博士論文

Bayesian Shrinkage Approaches to Parametric Inference
（パラメトリック推測への
ベイズ的な縮小化法によるアプローチ）

羽村靖之

Bayesian Shrinkage Approaches to Parametric Inference

by

Yasuyuki Hamura[1]

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Graduate School of Economics
of University of Tokyo
2020

[1]Graduate School of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN. JSPS Research Fellow.
E-Mail: yasu.stat@gmail.com

Abstract

In this thesis, we use shrinkage priors to obtain good Bayesian procedures for various statistical problems. In the first half of the thesis, we mainly use hierarchical priors constructed by assuming hyperpriors for global hyperparameters in order to prove domination results. In the second half of the thesis, properties of hyperpriors for local hyperparameters are analytically investigated in terms of shrinkage and robustness, improved numerical performance of global-local shrinkage priors is shown in simulation and empirical studies, and some results for Bayesian robust regression are also obtained.

In Part II of the thesis, we first consider in Chapter 2 the problems of estimating unknown parameters and predictive densities on the basis of observations of Poisson variables. Then, in Chapters 3 and 4, similar problems are treated in the negative multinomial case. Finally, in Chapter 5, we consider the prediction problem on the basis of Chi-squared and normal samples where the predictive density to be estimated is independent of the location parameter.

In Chapters 6 and 7 of Part III, we introduce classes of heavy-tailed distributions and investigate shrinkage and tail-robustness properties of corresponding Bayesian methods both analytically and numerically in the Poisson and normal cases. In Chapter 8 of Part III, the usefulness of our heavy-tailed distributions is further illustrated in the context of robust regression.

Acknowledgments

I would like to thank my advisor Tatsuya Kubokawa. His guidance, encouragement, and many valuable comments and helpful suggestions led to this thesis. I would also like to thank Kaoru Irie and Shonosuke Sugasawa and many other people for their support of my research. I thank my fellow graduate students. My research was supported in part by JSPS KAKENHI Grant Number JP20J10427 from Japan Society for the Promotion of Science. Finally, I thank my parents and other relatives.

Contents

Part I

Introduction

Chapter 1

Introduction

In this thesis, we take Bayesian shrinkage approaches to statistical inference for various parametric models and consider combining observed data with prior information or beliefs to obtain good estimators, predictive distributions, or, more generally, decisions which are superior to those based on the direct use of data from some theoretical and/or practical points of view. Theoretical aspects of classical shrinkage techniques are discussed for several models in Part II of the thesis, where Bayesian shrinkage estimators and predictive density estimators are derived and shown to have improved frequentist risk performance. In particular, we study Stein's phenomenon and prove that usual procedures are dominated by Bayesian shrinkage procedures under suitable conditions. Although Stein's phenomenon has been extensively investigated since Stein (1956) and many related problems have been considered for nonnormal and predictive models since Clevenson and Zidek (1975) and Komaki (2001), respectively, important new results are included in each chapter. Part III of the thesis treats more practical aspects as well. We propose classes of useful global-local shrinkage priors (Polson and Scott (2012a)) which have desirable properties in combining observed and prior information in possibly high-dimensional settings. The local priors are heavy-tailed distributions and the resulting estimators behave in such a way that small signals are shrunk toward prior means while large signals are kept unshrunk (tail robustness, Carvalho et al. (2010)). The usefulness of our heavy-tailed distributions is further illustrated in the context of robust regression.

The contents of Parts II and III can be considered as complementary to each other in terms of investigating the effects of global and local shrinkage. Let $f(x|\theta)$, $x \in \mathcal{X}$, be a likelihood function and $\pi(\theta|\lambda)$, $\theta \in \Theta$, be a conjugate prior with hyperparameter $\lambda \in \Lambda$, where $\mathcal{X}$ is the sample space and Θ is the parameter space. Suppose that $X_1, \ldots, X_m$ are independent observations from $f(x_1|\theta_1), \ldots, f(x_m|\theta_m)$, $x_1, \ldots, x_m \in \mathcal{X}$, $\theta_1, \ldots, \theta_m \in \Theta$, and consider using a joint prior of the form

$$
\begin{aligned}
&(\theta_1, \ldots, \theta_m) \sim \pi(\theta_1|\lambda_1\gamma) \cdots \pi(\theta_m|\lambda_m\gamma), \\
&(\lambda_1, \ldots, \lambda_m) \sim \psi^{\text{local}}(\lambda_1) \cdots \psi^{\text{local}}(\lambda_m), \quad \gamma \sim \psi^{\text{global}}(\gamma),
\end{aligned}
$$

where $\lambda_1, \ldots, \lambda_m$ and γ are local and global hyperparameters with hyperpriors $\psi^{\text{local}}(\lambda_1), \ldots, \psi^{\text{local}}(\lambda_m)$ and $\psi^{\text{global}}(\gamma)$ and satisfy $\lambda_1\gamma, \ldots, \lambda_m\gamma \in \Lambda$. In Part II of the thesis, we fix $\lambda_1, \ldots, \lambda_m$ and use the marginal joint prior of $\theta_1, \ldots, \theta_m$ based on $\psi^{\text{global}}(\gamma)$. On the other hand, in Chapters 6 and 7 of Part III, we first investigate the properties of $\psi^{\text{local}}(\lambda_1), \ldots, \psi^{\text{local}}(\lambda_m)$ analytically when γ is fixed and then investigate the numerical performance of the global-local shrinkage prior when

$\lambda_1, \ldots, \lambda_m$ and γ are not fixed.

Part II of the thesis is organized as follows.

- Chapter 2: Bayesian Point Estimators and Predictive Density Estimators Based on Poisson Observations

  In this chapter, we consider the problem of simultaneously estimating parameters of independent Poisson distributions in the presence of possibly unbalanced sample sizes under weighted standardized squared error loss. A class of heterogeneous Bayesian shrinkage estimators that utilize the unbalanced nature of sample sizes is proposed. To provide a theoretical justification, we first derive a necessary and sufficient condition for an estimator in the class to be proper Bayes and hence admissible and then obtain sufficient conditions for minimaxity that are compatible with the admissibility condition. Heterogeneous and homogeneous shrinkage estimators are compared by simulation. Several estimation methods are applied to data relating to the standardized mortality ratio. Finally, some extensions are considered. This chapter is based on Hamura and Kubokawa (2019b, 2020c).

- Chapter 3: Bayesian Shrinkage Estimation of Negative Multinomial Parameter Vectors

  The negative multinomial distribution is a multivariate generalization of the negative binomial distribution. In this chapter, we consider the problem of estimating an unknown matrix of probabilities on the basis of observations of negative multinomial variables under the standardized squared error loss. First, a general sufficient condition for a shrinkage estimator to dominate the UMVU estimator is derived and an empirical Bayes estimator satisfying the condition is constructed. Next, a hierarchical shrinkage prior is introduced, an associated Bayes estimator is shown to dominate the UMVU estimator under some conditions, and some remarks about posterior computation are presented. Finally, shrinkage estimators and the UMVU estimator are compared by simulation. This chapter is based on Hamura and Kubokawa (2020b).

- Chapter 4: Bayesian Shrinkage Approaches to Unbalanced Problems of Estimation and Prediction on the Basis of Negative Multinomial Samples

  In this chapter, we treat estimation and prediction problems where negative multinomial variables are observed and in particular consider unbalanced settings. First, the problem of estimating multiple negative multinomial parameter vectors under the standardized squared error loss is treated and a new empirical Bayes estimator which dominates the UMVU estimator under suitable conditions is derived. Second, we consider estimation of the joint predictive density of several multinomial tables under the Kullback-Leibler divergence and obtain a sufficient condition under which the Bayesian predictive density with respect to a hierarchical shrinkage prior dominates the Bayesian predictive density with respect to the Jeffreys prior. Third, our proposed Bayesian estimator and predictive density give risk improvements in simulations. Finally, the problem of estimating the joint predictive density of negative multinomial variables is discussed. This chapter is based on Hamura (2020).

- Chapter 5: Bayesian Predictive Density Estimation for a Chi-Squared Model Using Information from a Normal Observation with Unknown Mean and Variance

  In this chapter, we consider the problem of estimating the density function of a Chi-squared variable on the basis of observations of another Chi-squared variable and a normal variable

under the Kullback-Leibler divergence. We assume that these variables have a common unknown scale parameter and that the mean of the normal variable is also unknown. We compare the risk functions of two Bayesian predictive densities: one with respect to a hierarchical shrinkage prior and the other based on a noninformative prior. The hierarchical Bayesian predictive density depends on the normal variable while the Bayesian predictive density based on the noninformative prior does not. Sufficient conditions for the former to dominate the latter are obtained. These predictive densities are compared by simulation. This chapter is based on Hamura and Kubokawa (2020d).

Part III of the thesis is organized as follows.

- Chapter 6: On Global-Local Shrinkage Priors for Count Data

  Global-local shrinkage prior has been recognized as useful class of priors which can strongly shrink small signals towards prior means while keeping large signals unshrunk. Although such priors have been extensively discussed under Gaussian responses, we intensively encounter count responses in practice in which the previous knowledge of global-local shrinkage priors cannot be directly imported. In this chapter, we discuss global-local shrinkage priors for analyzing sequence of counts. We provide sufficient conditions under which the posterior mean keeps the observation as it is for very large signals, known as tail robustness property. Then, we propose tractable priors to meet the derived conditions approximately or exactly and develop an efficient posterior computation algorithm for Bayesian inference. The proposed methods are free from tuning parameters, that is, all the hyperparameters are automatically estimated based on the data. We demonstrate the proposed methods through simulation and an application to a real dataset. This chapter is based on Hamura, Irie and Sugasawa (2020a).

- Chapter 7: Shrinkage with Robustness: Log-Adjusted Priors for Sparse Signals

  We introduce a new class of distributions named log-adjusted shrinkage priors for the analysis of sparse signals, which extends the three parameter beta priors by multiplying an additional log-term to their densities. The proposed prior has density tails that are heavier than even those of the Cauchy distribution and realizes the tail-robustness of the Bayes estimator, while keeping the strong shrinkage effect on noises. We verify this property via the improved posterior mean squared errors in the tail. An integral representation with latent variables for the new density is available and enables fast and simple Gibbs samplers for the full posterior analysis. Our log-adjusted prior is significantly different from existing shrinkage priors with logarithms for allowing its further generalization by multiple log-terms in the density. The performance of the proposed priors is investigated through simulation studies and data analysis. This chapter is based on Hamura, Irie and Sugasawa (2020b).

- Chapter 8: Log-Regularly Varying Scale Mixture of Normals for Robust Regression

  Linear regression with the classical normality assumption for the error distribution may lead to an undesirable posterior inference of regression coefficients due to the potential outliers. This chapter considers the finite mixture of two components with thin and heavy tails as the error distribution, which has been routinely employed in applied statistics. For the heavily-tailed component, we introduce the novel class of distributions; their densities are log-regularly varying and have heavier tails than those of Cauchy distribution, yet

they are expressed as a scale mixture of normal distributions and enable the efficient posterior inference by Gibbs sampler. We prove the robustness to outliers of the posterior distributions under the proposed models with a minimal set of assumptions, which justifies the use of shrinkage priors with unbounded densities for the high-dimensional coefficient vector in the presence of outliers. The extensive comparison with the existing methods via simulation study shows the improved performance of our model in point and interval estimation, as well as its computational efficiency. Further, we confirm the posterior robustness of our method in the empirical study with the shrinkage priors for regression coefficients. This chapter is based on Hamura, Irie and Sugasawa (2020c).

Some concluding remarks are given in Chapter 9. In particular, the results and contribution of the thesis are summarized.

Part II

Decision-Theoretic Estimation and Prediction Problems

Chapter 2

Bayesian Point Estimators and Predictive Density Estimators Based on Poisson Observations

2.1 Introduction

Since the work of Clevenson and Zidek (1975), simultaneous estimation of parameters of independent Poisson distributions has been studied by many authors including Tsui (1979a), Tsui and Press (1982), Hwang (1982), and Chang and Shinozaki (2019). However, most of the existing work either concerns with the case of balanced sample sizes or deals with estimators in the unbalanced case which do not utilize the fact that the sample sizes are unbalanced. In this chapter, we consider the estimation problem in the case of unbalanced sample sizes and construct shrinkage estimators whose shrinkage factors reflect the fact that the sample sizes are unbalanced.

Suppose that $X_1, \dots, X_m$ are mutually independent Poisson random variables with means $n_1\lambda_1, \dots, n_m\lambda_m$, respectively, and that $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m) \in (0, \infty)^m$ is the unknown parameter while $n_1, \dots, n_m$ are positive known constants. This situation arises, for example, when for each $i = 1, \dots, m$, the observation X_i is the sum of $n_i (\in \mathbb{N})$ random sample from the Poisson distribution with mean λ_i. An example where $n_1, \dots, n_m$ are positive (possibly noninteger) real numbers is given in Komaki (2015). We treat the problem of estimating $\boldsymbol{\lambda}$ on the basis of $\boldsymbol{X} = (X_1, \dots, X_m)$.

In the balanced case with $n_1 = \cdots = n_m = 1$, the model becomes equivalent to that considered by Clevenson and Zidek (1975). When $n_1 = \cdots = n_m = 1$, for the avoidance of confusion, we use the different notation $\mathring{X}_1 = X_1, \dots, \mathring{X}_m = X_m$. Then, they showed that the estimator

$$\Big(1 - \frac{\beta_0 + m - 1}{\sum_{i=1}^m \mathring{X}_i + \beta_0 + m - 1}\Big)(\mathring{X}_1, \dots, \mathring{X}_m) \tag{2.1.1}$$

is admissible for $1 < \beta_0$ and minimax for $m \geq 2$ and $0 \leq \beta_0 \leq m - 1$ relative to the loss function $\sum_{i=1}^m (d_i - \lambda_i)^2/\lambda_i$.

Using their result, we can readily verify that the estimator

$$\Big(1-\frac{\beta_0+m-1}{\sum_{i=1}^m X_i+\beta_0+m-1}\Big)\Big(\frac{X_1}{n_1},\dots,\frac{X_m}{n_m}\Big) \tag{2.1.2}$$

dominates the ML estimator $(X_1/n_1,\dots,X_m/n_m)$ if $m\geq 2$ and $0\leq\beta_0\leq m-1$ under the loss

$$\sum_{i=1}^m \frac{n_i}{\lambda_i}(d_i-\lambda_i)^2. \tag{2.1.3}$$

However, the estimator given by (2.1.2) is not necessarily a natural shrinkage estimator from a practical point of view because the shrinkage factor $1-(\beta_0+m-1)/(\sum_{i=1}^m X_i+\beta_0+m-1)$ is common to all the samples irrespective of $\boldsymbol{n}=(n_1,\dots,n_m)$. In many applications, one of the purposes of using shrinkage estimators is to reduce the instability of ML estimators. In the present setting, for all $i,j=1,\dots,m$ such that $n_i<n_j$, the ML estimator X_i/n_i tends to be more unstable than X_j/n_j since the variance of X_i/n_i is approximately n_j/n_i times the variance of X_j/n_j if $\lambda_i\approx\lambda_j$. In addition, for each $i=1,\dots,m$, the sample size n_i can be interpreted as representing the amount of information the observation X_i contains about the unknown parameter λ_i. Thus, it seems reasonable to use a shrinkage estimator such that it shrinks the ML estimator X_i/n_i more toward the origin than X_j/n_j for all $i,j=1,\dots,m$ such that $n_i<n_j$. Furthermore, it turns out in Section 2.2 that the estimator given by (2.1.2) with $m\geq 2$ and $\beta_0\geq 0$ is the Bayes estimator with respect to a perhaps unnatural shrinkage prior which depends on $\boldsymbol{n}$ and puts less weight on the smaller values of λ_i than on the smaller values of λ_j for all $i,j=1,\dots,m$ such that $n_i<n_j$.

In this chapter, we consider the class of heterogeneous shrinkage estimators

$$\Big(\{1-\phi_1(\boldsymbol{X})\}\frac{X_1}{n_1},\dots,\{1-\phi_m(\boldsymbol{X})\}\frac{X_m}{n_m}\Big), \tag{2.1.4}$$

where the functions $\phi_1,\dots,\phi_m\colon\{0,1,2,\dots\}^m\to[0,1]$ satisfy that $\phi_i(\boldsymbol{x})>\phi_j(\boldsymbol{x})$ for all $\boldsymbol{x}=(x_1,\dots,x_m)\in\{0,1,2,\dots\}^m$ and $i,j=1,\dots,m$ such that $x_i,x_j\geq 1$ and $n_i<n_j$. We evaluate estimators under the weighted standardized squared loss function given by

$$L_{\boldsymbol{c}}(\boldsymbol{d},\boldsymbol{\lambda})=\sum_{i=1}^m\frac{c_i}{\lambda_i}(d_i-\lambda_i)^2, \tag{2.1.5}$$

where $\boldsymbol{c}=(c_1,\dots,c_m)\in(0,\infty)^m$ is a vector of weights possibly different form $\boldsymbol{n}$ and where $\boldsymbol{d}=(d_1,\dots,d_m)$ denotes a m-dimensional vector. For a discussion of the estimation of normal means in the presence of unequal weights as well as unequal variances, see Morris (1983).

Hamura and Kubokawa (2019b) constructed shrinkage estimators of the form (2.1.4) by using a class of improper priors introduced by Komaki (2015). However, they did not prove the admissibility of the estimators. In this chapter, we introduce a class of priors which includes both the proper priors of Clevenson and Zidek (1975) and the improper priors of Komaki (2015), construct proper Bayes estimators of the form (2.1.4), and derive sufficient conditions for the estimators to be minimax. The results for proper prior distributions are not straightforward generalizations of those for improper prior distributions. The main contribution of this chapter is to construct Bayes estimators of the form (2.1.4) which are both admissible and minimax.

In Section 2.2, we introduce the class of priors mentioned above, derive a necessary and sufficient condition for a prior in the class to be proper, and express the corresponding Bayes estimators explicitly. In Section 2.3, we derive sufficient conditions for minimaxity. In Section 2.4, some Monte Carlo evidence is presented. In Section 2.5, we treat real data. In Section 2.6, we consider some extensions. All the proofs of the lemmas in Sections 2.2 and 2.3 are given in the Appendix.

2.2 A Class of Bayes Estimators

We begin by providing a class of priors which includes the priors of both Clevenson and Zidek (1975) and Komaki (2015). Let

$$\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}(\boldsymbol{\lambda}) = \frac{\prod_{i=1}^m {\lambda_i}^{\beta_i-1}}{(\sum_{i=1}^m \lambda_i/\gamma_i)^\alpha} \int_0^\infty \frac{u^{\alpha-1+\beta_0}}{(u/\gamma_0+\sum_{i=1}^m \lambda_i/\gamma_i)^{\beta_0}} e^{-u} du \tag{2.2.1}$$

for $\alpha > 0$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m) \in (0,\infty)^m$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m) \in (0,\infty)^m$, $\beta_0 \geq 0$, and $\gamma_0 > 0$. By making the change of variables $u' = u/(\sum_{i=1}^m \lambda_i/\gamma_i)$, we can write (2.2.1) as

$$\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}(\boldsymbol{\lambda}) = \Big(\prod_{i=1}^m {\lambda_i}^{\beta_i-1}\Big) \int_0^\infty \frac{u^{\alpha-1+\beta_0}}{(1+u/\gamma_0)^{\beta_0}} e^{-u\sum_{i=1}^m \lambda_i/\gamma_i} du. \tag{2.2.2}$$

The class of priors of Clevenson and Zidek (1975) is expressed as

$$\pi_{m-1,\boldsymbol{j},\boldsymbol{j};\beta_0,1}(\boldsymbol{\lambda}) = \frac{1}{(\sum_{i=1}^m \lambda_i)^{m-1}} \int_0^\infty \frac{u^{m-2+\beta_0}}{(u+\sum_{i=1}^m \lambda_i)^{\beta_0}} e^{-u} du, \tag{2.2.3}$$

where $\boldsymbol{j} = (1, \ldots, 1) \in \mathbb{R}^m$, when $m \geq 2$ or $\beta_0 > 0$. The prior (2.2.3) is proper if $\beta_0 > 1$, as shown by Clevenson and Zidek (1975). On the other hand, the class of priors of Komaki (2015) is described by

$$\frac{\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};0,1}(\boldsymbol{\lambda})}{\Gamma(\alpha)} = \frac{\prod_{i=1}^m {\lambda_i}^{\beta_i-1}}{(\sum_{i=1}^m \lambda_i/\gamma_i)^\alpha}.$$

This prior is improper for all values of α, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$, which can be verified by, for example, Lemma 2.2.1 below.

The following lemma gives a necessary and sufficient condition for the prior $\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}$ to be proper. Let $\beta_\cdot = \sum_{i=1}^m \beta_i$.

Lemma 2.2.1 *The prior $\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}$ satisfies*

$$\int \cdots \int_{(0,\infty)^m} \pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}(\boldsymbol{\lambda}) d\boldsymbol{\lambda} < \infty$$

if and only if $\alpha < \beta_\cdot < \alpha + \beta_0$.

Next we derive an explicit form of the Bayes estimator against the prior $\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}$. To this end, we define

$$K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha; \gamma_0, \beta_0) = \int_0^\infty \frac{u^{\alpha-1}}{(1+u/\gamma_0)^{\beta_0}} \prod_{i=1}^m \frac{1}{(1+u/\gamma_i)^{\xi_i}} du$$

for $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m) \in [0, \infty)^m$ such that $\beta_0 + \sum_{i=1}^m \xi_i > \alpha$. This function is a generalization of the function given by Komaki (2015) which generalizes the beta function. Indeed, when $\gamma_0 = \gamma_1 = \cdots = \gamma_m$, we have

$$K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha; \gamma_0, \beta_0) = {\gamma_0}^{\alpha} B(\alpha, \beta_0 + \xi_\cdot - \alpha) \tag{2.2.4}$$

for $\xi_\cdot = \sum_{i=1}^m \xi_i$. The function K satisfies the following properties. Let $\boldsymbol{e}_i$ denote the i-th unit vector in $\mathbb{R}^m$, namely the i-th row of the $m \times m$ identity matrix, for $i = 1, \ldots, m$.

Lemma 2.2.2 *The following relations hold.*

(i)

$$\begin{aligned} \alpha K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha; \gamma_0, \beta_0) = \frac{\beta_0}{\gamma_0} K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha + 1; \gamma_0, \beta_0 + 1) \\ + \sum_{i=1}^m \frac{\xi_i}{\gamma_i} K(\boldsymbol{\gamma}, \boldsymbol{\xi} + \boldsymbol{e}_i, \alpha + 1; \gamma_0, \beta_0). \end{aligned} \tag{2.2.5}$$

(ii) *For $i = 1, \ldots, m$,*

$$K(\boldsymbol{\gamma}, \boldsymbol{\xi} + \boldsymbol{e}_i, \alpha + 1; \gamma_0, \beta_0) = \gamma_i \{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha; \gamma_0, \beta_0) - K(\boldsymbol{\gamma}, \boldsymbol{\xi} + \boldsymbol{e}_i, \alpha; \gamma_0, \beta_0)\}. \tag{2.2.6}$$

For the case of $\beta_0 = 0$, the relations (2.2.5) and (2.2.6) are given in Lemma 5 of Komaki (2015).

The following lemma gives some more properties of the function K and is crucial in Section 2.3 when we prove the existence of a heterogeneous shrinkage estimator which is both admissible and minimax.

Lemma 2.2.3 *Suppose that $\alpha < \beta_0 + \sum_{i=1}^m \xi_i - 1$. Then the following inequalities hold.*

(i) *For $i = 1, \ldots, m$,*

$$\frac{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha + 1; \gamma_0, \beta_0)}{K(\boldsymbol{\gamma}, \boldsymbol{\xi} - \boldsymbol{e}_i, \alpha; \gamma_0, \beta_0)} \geq \frac{K(\boldsymbol{\gamma}, \boldsymbol{\xi} + \boldsymbol{e}_i, \alpha + 1; \gamma_0, \beta_0)}{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha; \gamma_0, \beta_0)}. \tag{2.2.7}$$

Similarly,

$$\frac{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha + 1; \gamma_0, \beta_0)}{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha; \gamma_0, \beta_0 - 1)} \geq \frac{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha + 1; \gamma_0, \beta_0 + 1)}{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha; \gamma_0, \beta_0)}. \tag{2.2.8}$$

(ii) *For $i = 1, \ldots, m$,*

$$\frac{K(\boldsymbol{\gamma}, \boldsymbol{\xi} + \boldsymbol{e}_i, \alpha + 2; \gamma_0, \beta_0)}{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha; \gamma_0, \beta_0)} \geq \frac{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha + 1; \gamma_0, \beta_0)}{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha; \gamma_0, \beta_0)} \frac{K(\boldsymbol{\gamma}, \boldsymbol{\xi} + \boldsymbol{e}_i, \alpha + 1; \gamma_0, \beta_0)}{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha; \gamma_0, \beta_0)}. \tag{2.2.9}$$

Similarly,

$$\frac{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha + 2; \gamma_0, \beta_0 + 1)}{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha; \gamma_0, \beta_0)} \geq \frac{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha + 1; \gamma_0, \beta_0)}{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha; \gamma_0, \beta_0)} \frac{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha + 1; \gamma_0, \beta_0 + 1)}{K(\boldsymbol{\gamma}, \boldsymbol{\xi}, \alpha; \gamma_0, \beta_0)}. \tag{2.2.10}$$

For $\boldsymbol{v} = (v_1, \ldots, v_m) \in \mathbb{R}^m$ and $\tilde{\boldsymbol{v}} = (\tilde{v}_1, \ldots, \tilde{v}_m) \in \mathbb{R}^m$, we write $\boldsymbol{v} \circ \tilde{\boldsymbol{v}} = (v_1\tilde{v}_1, \ldots, v_m\tilde{v}_m)$. Let $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. For $\boldsymbol{x} = (x_1, \ldots, x_m) \in {\mathbb{N}_0}^m$ and $i = 1, \ldots, m$, we define

$$\phi_i^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{x}) = \begin{cases} \frac{1}{n_i\gamma_i}\frac{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{\beta},\alpha+\beta_0+1;\gamma_0,\beta_0)}{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{\beta}-\boldsymbol{e}_i,\alpha+\beta_0;\gamma_0,\beta_0)} & \text{if } x_i+\beta_i > 1 \text{ and } \sum_{j=1}^m (x_j+\beta_j) > \alpha+1 \\ 1 & \text{otherwise.} \end{cases}$$

The following lemma gives an explicit form of the Bayes estimator based on $\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}$.

Lemma 2.2.4 *Suppose $\alpha < \beta$.. Then the estimator $\widehat{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}$ defined by*

$$\Big(\{1-\phi_1^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{X})\}\frac{X_1+\beta_1-1}{n_1}, \ldots, \{1-\phi_m^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{X})\}\frac{X_m+\beta_m-1}{n_m}\Big) \tag{2.2.11}$$

is the unique Bayes estimator of $\boldsymbol{\lambda}$ on the basis of $\boldsymbol{X}$ against the prior $\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}$ under the loss function $L_{\boldsymbol{c}}$ given by (2.1.5).

It is worth noting that the Bayes estimator $\widehat{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}$ is robust in the sense that it does not depend on $\boldsymbol{c}$. We remark that the estimator with $\boldsymbol{\beta} = \boldsymbol{j}$ shrinks the ML estimator toward the origin.

Let $\boldsymbol{\nu} = (1/n_1, \ldots, 1/n_m)$ be the vector whose elements are the reciprocals of the sample sizes, so that $\boldsymbol{n}\circ\boldsymbol{\nu} = \boldsymbol{j}$. Suppose that $m \geq 2$. Then the Bayes estimator with $\alpha = m-1$, $\boldsymbol{\beta} = \boldsymbol{j}$, $\boldsymbol{\gamma} = \boldsymbol{\nu}$, and $\gamma_0 = 1$, namely $\widehat{\boldsymbol{\lambda}}^{(m-1,\boldsymbol{j},\boldsymbol{\nu};\beta_0,1)}$, reduces to (2.1.2) by (2.2.4). Thus, (2.1.2) is the Bayes estimator against the prior

$$\pi_{m-1,\boldsymbol{j},\boldsymbol{\nu};\beta_0,1}(\boldsymbol{\lambda}) = \frac{1}{(\sum_{i=1}^m n_i\lambda_i)^{m-1}}\int_0^\infty \frac{u^{m-2+\beta_0}}{(u+\sum_{i=1}^m n_i\lambda_i)^{\beta_0}}e^{-u}du.$$

In the context of shrinkage estimation, however, this choice of prior may be inappropriate since it depends on $\boldsymbol{n}$ and puts less weight on the smaller values of λ_i than on the smaller values of λ_j for all $i, j = 1, \ldots, m$ such that $n_i < n_j$. Indeed, the shrinkage factor of the resulting Bayes estimator (2.1.2) fails to reflect the fact that the sample size $\boldsymbol{n}$ is unbalanced.

Finally, we propose an estimator of the form (2.1.4) which shrinks the ML estimator X_i/n_i more toward the origin than X_j/n_j for all $i, j = 1, \ldots, m$ such that $n_i < n_j$. We consider the case where $\boldsymbol{\beta} = \boldsymbol{\gamma} = \boldsymbol{j}$ and $\alpha < m$ for $\boldsymbol{j} = (1, \ldots, 1) \in \mathbb{R}^m$. Then the prior is

$$\pi_{\alpha,\boldsymbol{j},\boldsymbol{j};\beta_0,\gamma_0}(\boldsymbol{\lambda}) = \frac{1}{(\sum_{i=1}^m \lambda_i)^{\alpha}}\int_0^\infty \frac{u^{\alpha-1+\beta_0}}{(u/\gamma_0+\sum_{i=1}^m \lambda_i)^{\beta_0}}e^{-u}du,$$

which is a shrinkage prior symmetric in $\lambda_1, \ldots, \lambda_m$. The resulting estimator can be expressed as

$$\widehat{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{j},\boldsymbol{j};\beta_0,\gamma_0)} = \Big(\{1-\phi_1^{(\alpha,\boldsymbol{j},\boldsymbol{j};\beta_0,\gamma_0)}(\boldsymbol{X})\}\frac{X_1}{n_1}, \ldots, \{1-\phi_m^{(\alpha,\boldsymbol{j},\boldsymbol{j};\beta_0,\gamma_0)}(\boldsymbol{X})\}\frac{X_m}{n_m}\Big), \tag{2.2.12}$$

where

$$\phi_i^{(\alpha,\boldsymbol{j},\boldsymbol{j};\beta_0,\gamma_0)}(\boldsymbol{x}) = \begin{cases} \frac{1}{n_i}\frac{K(\boldsymbol{n},\boldsymbol{x}+\boldsymbol{j},\alpha+\beta_0+1;\gamma_0,\beta_0)}{K(\boldsymbol{n},\boldsymbol{x}+\boldsymbol{j}-\boldsymbol{e}_i,\alpha+\beta_0;\gamma_0,\beta_0)} & \text{if } x_i \geq 1 \\ 1 & \text{if } x_i = 0 \end{cases} \tag{2.2.13}$$

for $\boldsymbol{x} = (x_1, \ldots, x_m) \in {\mathbb{N}_0}^m$ and $i = 1, \ldots, m$. This shrinkage estimator has the following heterogeneity properties.

Lemma 2.2.5 *Let* $\boldsymbol{x} = (x_1, \ldots, x_m) \in {\mathbb{N}_0}^m$ *and suppose* $\alpha < m$.

(i) *Let* $i \in \{1, \ldots, m\}$. *Then*

$$0 < \phi_i^{(\alpha, \boldsymbol{j}, \boldsymbol{j}; \beta_0, \gamma_0)}(\boldsymbol{x}) \leq 1. \tag{2.2.14}$$

Equality holds if and only if $x_i = 0$.

(ii) *Let* $i, j \in \{1, \ldots, m\}$ *and suppose* $x_i, x_j \geq 1$. *Then*

$$\phi_i^{(\alpha, \boldsymbol{j}, \boldsymbol{j}; \beta_0, \gamma_0)}(\boldsymbol{x}) > \phi_j^{(\alpha, \boldsymbol{j}, \boldsymbol{j}; \beta_0, \gamma_0)}(\boldsymbol{x}) \quad \text{if and only if} \quad n_i < n_j. \tag{2.2.15}$$

(iii) *Let* $i \in \{1, \ldots, m\}$ *and suppose* $x_i \geq 1$. *Suppose further that* $\alpha < m - 2$. *Then*

$$\lim_{n_i \to \infty} \phi_i^{(\alpha, \boldsymbol{j}, \boldsymbol{j}; \beta_0, \gamma_0)}(\boldsymbol{x}) = 0. \tag{2.2.16}$$

In the case where $\boldsymbol{n} = \boldsymbol{j}$ and $\alpha + 1 = m \geq 2$ and $\gamma_0 = 1$, both of the estimators (2.1.2) and (2.2.12) coincide with the estimator (2.1.1) given by Clevenson and Zidek (1975). However, we propose the latter as an important generalization of (2.1.1) which satisfies the heterogeneity properties (2.2.14), (2.2.15), and (2.2.16).

2.3 Sufficient Conditions for Minimaxity

In this section, we derive sufficient conditions for the estimator $\widehat{\boldsymbol{\lambda}}^{(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}; \beta_0, \gamma_0)}$ given by (2.2.11) to be minimax under the loss function $L_{\boldsymbol{c}}$ given by (2.1.5). Since it can be shown that the ML estimator $\widehat{\boldsymbol{\lambda}}^{\mathrm{ML}} = (\hat{\lambda}_1^{\mathrm{ML}}, \ldots, \hat{\lambda}_m^{\mathrm{ML}}) = (X_1/n_1, \ldots, X_m/n_m)$ is the constant risk minimax estimator, it suffices to find conditions under which $\widehat{\boldsymbol{\lambda}}^{(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}; \beta_0, \gamma_0)}$ dominates $\widehat{\boldsymbol{\lambda}}^{\mathrm{ML}}$. Hereafter, we restrict our attention to the case of $\alpha < m$ and $\boldsymbol{\beta} = \boldsymbol{j}$ and consider the shrinkage estimator

$$\widehat{\boldsymbol{\lambda}}^{(\alpha, \boldsymbol{j}, \boldsymbol{\gamma}; \beta_0, \gamma_0)} = \Big(\{1 - \phi_1^{(\alpha, \boldsymbol{j}, \boldsymbol{\gamma}; \beta_0, \gamma_0)}(\boldsymbol{X})\} \frac{X_1}{n_1}, \ldots, \{1 - \phi_m^{(\alpha, \boldsymbol{j}, \boldsymbol{\gamma}; \beta_0, \gamma_0)}(\boldsymbol{X})\} \frac{X_m}{n_m} \Big), \tag{2.3.1}$$

where

$$\phi_i^{(\alpha, \boldsymbol{j}, \boldsymbol{\gamma}; \beta_0, \gamma_0)}(\boldsymbol{x}) = \begin{cases} \frac{1}{n_i \gamma_i} \frac{K(\boldsymbol{n} \circ \boldsymbol{\gamma}, \boldsymbol{x} + \boldsymbol{j}, \alpha + \beta_0 + 1; \gamma_0, \beta_0)}{K(\boldsymbol{n} \circ \boldsymbol{\gamma}, \boldsymbol{x} + \boldsymbol{j} - \boldsymbol{e}_i, \alpha + \beta_0; \gamma_0, \beta_0)} & \text{if } x_i \geq 1 \\ 1 & \text{if } x_i = 0 \end{cases}$$

for $\boldsymbol{x} = (x_1, \ldots, x_m) \in {\mathbb{N}_0}^m$ and $i = 1, \ldots, m$.

The following result, due to Hudson (1978), is used in the proof of Theorem 2.3.1 below.

Lemma 2.3.1 *Let* $h \colon {\mathbb{N}_0}^m \to \mathbb{R}$ *and suppose that* $E_{\boldsymbol{\lambda}}[|h(\boldsymbol{X})|] < \infty$. *Then for all* $i = 1, \ldots, m$, *if* $h(\boldsymbol{x}) = 0$ *for all* $\boldsymbol{x} = (x_1, \ldots, x_m) \in {\mathbb{N}_0}^m$ *such that* $x_i = 0$, *we have*

$$E_{\boldsymbol{\lambda}}\Big[\frac{h(\boldsymbol{X})}{n_i \lambda_i}\Big] = E_{\boldsymbol{\lambda}}\Big[\frac{h(\boldsymbol{X} + \boldsymbol{e}_i)}{X_i + 1}\Big].$$

For simplicity of notation, we let $a_i = n_i \gamma_i$ and $C_i = (c_i/n_i)(1/a_i) = c_i/({n_i}^2 \gamma_i)$ for $i = 1, \ldots, m$ and let $\underline{a} = \min_{1 \leq i \leq m} a_i$, $\overline{a} = \max_{1 \leq i \leq m} a_i$, $\underline{C} = \min_{1 \leq i \leq m} C_i$, $\overline{C} = \max_{1 \leq i \leq m} C_i$, and $C_\cdot = \sum_{i=1}^m C_i$. The following theorem, which will be proved later in this section, gives two sufficient conditions for the minimaxity of $\widehat{\boldsymbol{\lambda}}^{(\alpha, \boldsymbol{j}, \boldsymbol{\gamma}; \beta_0, \gamma_0)}$.

Theorem 2.3.1 *Assume that $\alpha < m$ and that $\gamma_0 \le \underline{a}$.*

(i) *Suppose that*

$$\alpha + \beta_0 \le \frac{2}{3}\Big(\frac{C_\cdot}{\overline{C}} - 1\Big)\Big(\frac{\beta_0}{m} + 1\Big). \tag{2.3.2}$$

Then the estimator $\widehat{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{j},\boldsymbol{\gamma};\beta_0,\gamma_0)}$ is minimax under the loss $L_{\boldsymbol{c}}$.

(ii) *Let $\rho = \{\overline{C}(\alpha+\beta_0+1) - C_\cdot\}/\{\underline{C}(\alpha+\beta_0)\}$. Suppose that $0 \le \rho \le 1 - (1/2)(\overline{a}/\underline{a})$ and that*

$$2\rho\Big(\beta_0 + m + \frac{\underline{a}}{\overline{a}}\Big) \le \frac{\overline{C}}{\underline{C}}(\alpha + 2\beta_0 + 1) - \frac{C_\cdot}{\underline{C}} + 2\frac{\underline{a}}{\overline{a}} - 1. \tag{2.3.3}$$

Then the estimator $\widehat{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{j},\boldsymbol{\gamma};\beta_0,\gamma_0)}$ is minimax under the loss $L_{\boldsymbol{c}}$.

Part (i) of Theorem 2.3.1 is a generalization of Theorem 3 of Hamura and Kubokawa (2019b). They consider the case of $\beta_0 = 0$. In this case, the prior is improper by Lemma 2.2.1 but whenever $m \ge 2$, there is always a value of $\alpha > 0$ that satisfies the sufficient condition (2.3.2) for the minimaxity of the estimator $\widehat{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{j},\boldsymbol{\gamma};0,\underline{a})}$. On the other hand, when the prior is proper, assumption (2.3.2) implies $m < (2/3)(C_\cdot/\overline{C} - 1)(\beta_0/m + 1) \le 2(C_\cdot/\overline{C} - 1)$. Therefore, there exist $C_1, \ldots, C_m$ such that the condition (2.3.2) is violated for any choice of a proper prior. We can also generalize Theorem 4 of Hamura and Kubokawa (2019b) to obtain another sufficient condition for the case that $\overline{C}(\alpha+\beta_0+1) \le C_\cdot$: if $\underline{a}/\overline{a} \ge 1/2$ and $\alpha + \beta_0 \le C_\cdot/\overline{C} - 1$, then $\widehat{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{j},\boldsymbol{\gamma};\beta_0,\gamma_0)}$ is minimax under the loss $L_{\boldsymbol{c}}$.

Let $\underline{n} = \min_{1\le i\le m} n_i$ and $\overline{n} = \max_{1\le i\le m} n_i$. Let $C_i^* = c_i/{n_i}^2$ for $i = 1, \ldots, m$ and define $\underline{C}^*$, $\overline{C}^*$, and $C_\cdot^*$ analogously. Combining Lemmas 2.2.1, 2.2.4, and 2.2.5 and Theorem 2.3.1, we obtain the following theorem.

Theorem 2.3.2 *Suppose that $\alpha < m < \alpha + \beta_0$ and that $\gamma_0 \le \underline{n}$. Suppose further that one of the following two conditions holds:*

(i)

$$\alpha + \beta_0 \le \frac{2}{3}\Big(\frac{C_\cdot^*}{\overline{C}^*} - 1\Big)\Big(\frac{\beta_0}{m} + 1\Big).$$

(ii)

$$\frac{\overline{C}^*(\alpha+\beta_0+1) - C_\cdot^*}{\underline{C}^*(\alpha+\beta_0)} \le \min\Big\{1 - \frac{1}{2}\frac{\overline{n}}{\underline{n}}, \frac{(\overline{C}^*/\underline{C}^*)(\alpha+2\beta_0+1) - C_\cdot^*/\underline{C}^* + 2\underline{n}/\overline{n} - 1}{2(\beta_0 + m + \underline{n}/\overline{n})}\Big\}.$$

Then the estimator $\widehat{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{j},\boldsymbol{j};\beta_0,\gamma_0)}$ given by (2.2.12) is admissible and minimax under the loss $L_{\boldsymbol{c}}$. Furthermore, for all $\boldsymbol{x} = (x_1, \ldots, x_m) \in {\mathbb{N}_0}^m$ and $i, j \in \{1, \ldots, m\}$ such that $x_i, x_j \ge 1$, it satisfies (2.2.14), (2.2.15), and, if $\alpha < m - 2$, (2.2.16).

It can be seen that there exists an admissible minimax shrinkage estimator that satisfies (2.2.14), (2.2.15), and (2.2.16) by, for example, applying part (i) of Theorem 2.3.2 to the case where $(\alpha, \beta_0, \gamma_0) = (1, m, \underline{n})$ and m is sufficiently large and $\underline{C}^*/\overline{C}^*$ is sufficiently close to 1. Furthermore, though the details are omitted here, it can be shown from part (i) of Theorem 2.3.2 that there exists $\alpha > 0$, $\beta_0 \geq 0$, and $\gamma_0 > 0$ such that the conclusion of Theorem 2.3.2 holds if $2 \leq m < (4/3)(\underline{C}^*/\overline{C}^* - 1)$. This condition reduces to

$$2 \leq m < \frac{4}{3}\Big(\sum_{i=1}^{m} \frac{\underline{n}^k}{{n_i}^k} - 1\Big)$$

with $k = 1$ when $c_i = n_i$ and with $k = 2$ when $c_i = 1$.

In the particular case of $\boldsymbol{n} = \boldsymbol{c} = \boldsymbol{\gamma} = \boldsymbol{j}$ and $\gamma_0 = 1$, the condition for $\widehat{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{j},\boldsymbol{j};\beta_0,\gamma_0)}$ to be admissible and minimax given in part (i) of Theorem 2.3.2 is

$$\alpha < m < \alpha + \beta_0 \quad \text{and} \quad \alpha + \Big(1 - \frac{2}{3}\frac{m-1}{m}\Big)\beta_0 \leq \frac{2}{3}(m-1), \tag{2.3.4}$$

whereas that given in part (ii) of Theorem 2.3.2 is

$$\alpha < m < \alpha + \beta_0 \leq 2(m-1) \quad \text{and} \quad \frac{\alpha + \beta_0 + 1 - m}{\alpha + \beta_0} \leq \frac{\alpha + 2\beta_0 - m + 2}{2(\beta_0 + m + 1)}. \tag{2.3.5}$$

Conditions (2.3.4) and (2.3.5) correspond to (2.3.2) and (2.3.3), respectively. The condition given by Clevenson and Zidek (1975) is

$$\alpha = m - 1 \quad \text{and} \quad 1 < \beta_0 \leq m - 1.$$

When $m \geq 2$ and $\alpha = m - 1$, condition (2.3.4) is not satisfied for any values of β_0 but condition (2.3.5) becomes

$$1 < \beta_0 \leq (m-1)/3.$$

Thus, although the result of Clevenson and Zidek (1975) is not completely included, Theorem 2.3.1 or 2.3.2, which was derived for estimating $\boldsymbol{\lambda}$ when $\boldsymbol{n}$ is unbalanced, gives the sufficient condition which is close to that of Clevenson and Zidek (1975) even in the case of balanced sample sizes.

Proof of Theorem 2.3.1. Let $\Delta = E_{\boldsymbol{\lambda}}[L_{\boldsymbol{c}}(\widehat{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{j},\boldsymbol{\gamma};\beta_0,\gamma_0)}, \boldsymbol{\lambda})] - E_{\boldsymbol{\lambda}}[L_{\boldsymbol{c}}(\widehat{\boldsymbol{\lambda}}^{\mathrm{ML}}, \boldsymbol{\lambda})]$. From (2.3.1),

$$\begin{aligned}
\Delta &= E_{\boldsymbol{\lambda}}\Big[\sum_{i=1}^{m}\Big[\frac{c_i}{\lambda_i}\Big\{\frac{X_i}{n_i} - \lambda_i - \frac{X_i}{n_i}\phi_i^{(\alpha,\boldsymbol{j},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{X})\Big\}^2 - \frac{c_i}{\lambda_i}\Big(\frac{X_i}{n_i} - \lambda_i\Big)^2\Big]\Big] \\
&= E_{\boldsymbol{\lambda}}\Big[\sum_{i=1}^{m}\Big(\frac{n_i c_i}{n_i \lambda_i}\Big[\Big\{\frac{X_i}{n_i}\phi_i^{(\alpha,\boldsymbol{j},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{X})\Big\}^2 - 2\Big(\frac{X_i}{n_i}\Big)^2\phi_i^{(\alpha,\boldsymbol{j},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{X})\Big] \\
&\quad + 2c_i\frac{X_i}{n_i}\phi_i^{(\alpha,\boldsymbol{j},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{X})\Big)\Big],
\end{aligned}$$

which is, by application of Lemma 2.3.1,

$$\Delta = E_{\boldsymbol{\lambda}}\Big[\sum_{i=1}^{m}\Big(\frac{n_i c_i}{X_i+1}\Big[\Big\{\frac{X_i+1}{n_i}\phi_i^{(\alpha,\boldsymbol{j},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{X}+\boldsymbol{e}_i)\Big\}^2 - 2\Big(\frac{X_i+1}{n_i}\Big)^2\phi_i^{(\alpha,\boldsymbol{j},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{X}+\boldsymbol{e}_i)\Big] + 2c_i\frac{X_i}{n_i}\phi_i^{(\alpha,\boldsymbol{j},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{X})\Big)\Big].$$

Therefore, we can write the risk difference as $\Delta = E_{\boldsymbol{\lambda}}[I_1(\boldsymbol{X}) - 2I_2(\boldsymbol{X}) + 2I_3(\boldsymbol{X})]$, where

$$I_1(\boldsymbol{x}) = \sum_{i=1}^{m}\frac{c_i}{n_i}(x_i+1)\Big\{\frac{1}{n_i\gamma_i}\frac{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{j}+\boldsymbol{e}_i,\alpha+\beta_0+1;\gamma_0,\beta_0)}{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{j},\alpha+\beta_0;\gamma_0,\beta_0)}\Big\}^2,$$

$$I_2(\boldsymbol{x}) = \sum_{i=1}^{m}\frac{c_i}{n_i}\frac{x_i+1}{n_i\gamma_i}\frac{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{j}+\boldsymbol{e}_i,\alpha+\beta_0+1;\gamma_0,\beta_0)}{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{j},\alpha+\beta_0;\gamma_0,\beta_0)},$$

$$I_3(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \boldsymbol{x}=\boldsymbol{0} \\ \displaystyle\sum_{i=1}^{m}\frac{c_i}{n_i}\frac{x_i}{n_i\gamma_i}\frac{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{j},\alpha+\beta_0+1;\gamma_0,\beta_0)}{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{j}-\boldsymbol{e}_i,\alpha+\beta_0;\gamma_0,\beta_0)} & \text{otherwise,}\end{cases}$$

for $\boldsymbol{x}=(x_1,\ldots,x_m)\in{\mathbb{N}_0}^m$. We have $I_1(\boldsymbol{0}) - 2I_2(\boldsymbol{0}) + 2I_3(\boldsymbol{0}) < 0$ since

$$\frac{1}{n_i\gamma_i}\frac{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{j}+\boldsymbol{e}_i,\alpha+\beta_0+1;\gamma_0,\beta_0)}{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{j},\alpha+\beta_0;\gamma_0,\beta_0)} \in [0,1].$$

Thus, it is sufficient to show that $I_1(\boldsymbol{x}) - 2I_2(\boldsymbol{x}) + 2I_3(\boldsymbol{x}) \le 0$ for all $\boldsymbol{x}\in{\mathbb{N}_0}^m\setminus\{\boldsymbol{0}\}$.

Fix $\boldsymbol{x}=(x_1,\ldots,x_m)\in{\mathbb{N}_0}^m\setminus\{\boldsymbol{0}\}$. Hereafter, for simplicity, we use the abbreviated notation

$$I_1 = I_1(\boldsymbol{x}),\quad I_2 = I_2(\boldsymbol{x}),\quad I_3 = I_3(\boldsymbol{x}),$$

$$I = I_1 - 2I_2 + 2I_3,$$

$$H(c) = \frac{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{j},\alpha+\beta_0+c;\gamma_0,\beta_0)}{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{j},\alpha+\beta_0;\gamma_0,\beta_0)},$$

$$H(0,c) = \frac{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{j},\alpha+\beta_0+c;\gamma_0,\beta_0+1)}{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{j},\alpha+\beta_0;\gamma_0,\beta_0)},$$

and

$$H(\pm i,c) = \frac{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{j}\pm\boldsymbol{e}_i,\alpha+\beta_0+c;\gamma_0,\beta_0)}{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{j},\alpha+\beta_0;\gamma_0,\beta_0)}$$

for $c=0,1,2$ and $i=1,\ldots,m$ when well defined.

For part (i), we have

$$I_1 \le \overline{C}H(1)\sum_{i=1}^{m}\frac{x_i+1}{a_i}H(i,1).$$

By part (ii) of Lemma 2.2.2, we obtain

$$I_2 = \sum_{i=1}^{m}\frac{c_i}{n_i}\frac{x_i+1}{a_i}H(1) - \sum_{i=1}^{m}C_i\frac{x_i+1}{a_i}H(i,2)$$

and

$$I_3 = \sum_{i=1}^{m} \frac{c_i}{n_i}\frac{x_i}{a_i}\Big[H(1) - \Big\{H(1) - \frac{H(1)}{H(-i,0)}\Big\}\Big]$$
$$= \sum_{i=1}^{m} \frac{c_i}{n_i}\frac{x_i}{a_i}H(1) - \sum_{i=1}^{m} \frac{c_i}{n_i}\frac{x_i}{a_i}\frac{1}{a_i}\frac{H(1)}{H(-i,0)}H(1).$$

Then, from part (i) of Lemma 2.2.3,

$$I_3 \le \sum_{i=1}^{m} \frac{c_i}{n_i}\frac{x_i}{a_i}H(1) - \sum_{i=1}^{m} \frac{c_i}{n_i}\frac{x_i}{a_i}\frac{1}{a_i}H(i,1)H(1) \tag{2.3.6}$$
$$\le \sum_{i=1}^{m} \frac{c_i}{n_i}\frac{x_i}{a_i}H(1) - \sum_{i=1}^{m} \frac{c_i}{n_i}\frac{x_i+1}{a_i}\frac{1}{a_i}H(i,1)H(1) + \overline{C}\sum_{i=1}^{m}\frac{1}{a_i}H(i,1)H(1).$$

Therefore,

$$I \le \overline{C}H(1)\sum_{i=1}^{m}\frac{x_i+1}{a_i}H(i,1) - 2C.H(1)$$
$$+ 2\sum_{i=1}^{m} C_i\frac{x_i+1}{a_i}\{H(i,2) - H(i,1)H(1)\} + 2\overline{C}\sum_{i=1}^{m}\frac{1}{a_i}H(i,1)H(1)$$
$$\le \overline{C}H(1)\sum_{i=1}^{m}\frac{x_i+1}{a_i}H(i,1) - 2C.H(1)$$
$$+ 2\overline{C}\sum_{i=1}^{m}\frac{x_i+1}{a_i}\{H(i,2) - H(i,1)H(1)\} + 2\overline{C}\sum_{i=1}^{m}\frac{1}{a_i}H(i,1)H(1),$$

where the second inequality follows since $H(i,2) - H(i,1)H(1) \ge 0$ for all $i = 1, \ldots, m$ by part (ii) of Lemma 2.2.3. By applying part (i) of Lemma 2.2.2, we obtain

$$I \le \overline{C}H(1)\Big\{\alpha + \beta_0 - \frac{\beta_0}{\gamma_0}H(0,1)\Big\} - 2C.H(1) + 2\overline{C}\sum_{i=1}^{m}\frac{1}{a_i}H(i,1)H(1)$$
$$+ 2\overline{C}\Big\{(\alpha+\beta_0+1)H(1) - \frac{\beta_0}{\gamma_0}H(0,2) - (\alpha+\beta_0)H(1) + \frac{\beta_0}{\gamma_0}H(0,1)H(1)\Big\}$$
$$\le \overline{C}H(1)\Big\{\alpha + \beta_0 - \sum_{i=1}^{m}\frac{\beta_0/m}{a_i}H(i,1)\Big\} - 2C.H(1)$$
$$+ 2\overline{C}\sum_{i=1}^{m}\frac{1}{a_i}H(i,1)H(1) + 2\overline{C}H(1)$$
$$= \{\overline{C}(\alpha+\beta_0+2) - 2C.\}H(1) + \overline{C}\sum_{i=1}^{m}\frac{2-\beta_0/m}{a_i}H(i,1)H(1),$$

where the second inequality follows from the assumption that $\gamma_0 \le \underline{a}$ and part (ii) of Lemma 2.2.3. Since

$$\alpha + \beta_0 = \sum_{i=1}^{m}\Big\{\frac{x_i+1}{a_i}H(i,1) + \frac{\beta_0/m}{\gamma_0}H(0,1)\Big\} \ge \sum_{i=1}^{m}\frac{1+\beta_0/m}{a_i}H(i,1)$$

by part (i) of Lemma 2.2.2 and the assumption that $\gamma_0 \le \underline{a}$ and since assumption (2.3.2) implies $2 - \beta_0/m \ge 0$, we conclude that

$$I/H(1) \le \overline{C}(\alpha + \beta_0 + 2) - 2C. + \overline{C}\frac{2 - \beta_0/m}{1 + \beta_0/m}(\alpha + \beta_0) \le 0,$$

where the second inequality follows from assumption (2.3.2). This completes the proof of part (i).

For part (ii), let $\underline{i} \in \{1, \ldots, m\}$ be an index such that $a_{\underline{i}} = \underline{a}$. Then we have

$$I_1 - 2I_2 \le \Big\{\frac{1}{\underline{a}}H(\underline{i}, 1) - 2\Big\}I_2. \tag{2.3.7}$$

Note that, by part (ii) and part (i) of Lemma 2.2.2,

$$\begin{aligned} I_2 &= \sum_{i=1}^{m} C_i(x_i + 1)H(1) - \sum_{i=1}^{m} C_i(x_i + 1)\frac{1}{a_i}H(i, 2) \\ &\ge \sum_{i=1}^{m} C_i(x_i + 1)H(1) - \overline{C}\sum_{i=1}^{m}(x_i + 1)\frac{1}{a_i}H(i, 2) \\ &= \Big\{\sum_{i=1}^{m} C_i(x_i + 1) - \overline{C}(\alpha + \beta_0 + 1)\Big\}H(1) + \overline{C}\frac{\beta_0}{\gamma_0}H(0, 2). \end{aligned} \tag{2.3.8}$$

Since $(1/\underline{a})H(\underline{i}, 1) \le 1$, it follows from (2.3.7) and (2.3.8) that

$$\begin{aligned} I_1 - 2I_2 &\le \Big\{\frac{1}{\underline{a}}H(\underline{i}, 1) - 2\Big\} \\ &\quad \times \Big[\Big\{\sum_{i=1}^{m} C_i(x_i + 1) - \overline{C}(\alpha + \beta_0 + 1)\Big\}H(1) + \overline{C}\frac{\beta_0}{\gamma_0}H(0, 2)\Big] \\ &= -\Big\{2 - \frac{1}{\underline{a}}H(\underline{i}, 1)\Big\}\sum_{i=1}^{m} C_i x_i H(1) - \Big\{2 - \frac{1}{\underline{a}}H(\underline{i}, 1)\Big\}\overline{C}\frac{\beta_0}{\gamma_0}H(0, 2) \\ &\quad + \Big\{2 - \frac{1}{\underline{a}}H(\underline{i}, 1)\Big\}\{\overline{C}(\alpha + \beta_0 + 1) - C.\}H(1). \end{aligned} \tag{2.3.9}$$

Combining (2.3.6) and (2.3.9) gives

$$\begin{aligned} I &\le \frac{1}{\underline{a}}H(\underline{i}, 1)\sum_{i=1}^{m} C_i x_i H(1) - \Big\{2 - \frac{1}{\underline{a}}H(\underline{i}, 1)\Big\}\overline{C}\frac{\beta_0}{\gamma_0}H(0, 2) \\ &\quad + \Big\{2 - \frac{1}{\underline{a}}H(\underline{i}, 1)\Big\}\{\overline{C}(\alpha + \beta_0 + 1) - C.\}H(1) - 2\sum_{i=1}^{m} C_i\frac{x_i}{a_i}H(i, 1)H(1). \end{aligned} \tag{2.3.10}$$

Note that

$$\sum_{i=1}^{m} C_i\frac{x_i}{a_i}H(i, 1) \ge \frac{\underline{a}}{\overline{a}}\frac{1}{\underline{a}}H(\underline{i}, 1)\sum_{i=1}^{m} C_i x_i$$

and that

$$\sum_{i=1}^{m} C_i \frac{x_i}{a_i} H(i,1) \geq \underline{C} \sum_{i=1}^{m} \frac{x_i+1}{a_i} H(i,1) - \underline{C} \sum_{i=1}^{m} \frac{1}{a_i} H(i,1)$$
$$= \underline{C}(\alpha+\beta_0) - \underline{C}\frac{\beta_0}{\gamma_0} H(0,1) - \underline{C} \sum_{i=1}^{m} \frac{1}{a_i} H(i,1) \tag{2.3.11}$$

by part (i) of Lemma 2.2.2. Then we have

$$\begin{aligned}
I \leq{}& \frac{1}{\underline{a}} H(\underline{i},1) \sum_{i=1}^{m} C_i x_i H(1) - \Big\{2 - \frac{1}{\underline{a}} H(\underline{i},1)\Big\} \overline{C} \frac{\beta_0}{\gamma_0} H(0,2) \\
&+ \Big\{2 - \frac{1}{\underline{a}} H(\underline{i},1)\Big\} \{\overline{C}(\alpha+\beta_0+1) - C.\} H(1) \\
&- 2\Big(1 - \frac{1}{2}\frac{\overline{a}}{\underline{a}} - \rho\Big) \sum_{i=1}^{m} C_i \frac{x_i}{a_i} H(i,1) H(1) - 2\frac{1}{2}\frac{\overline{a}}{\underline{a}}\frac{\underline{a}}{\overline{a}}\frac{1}{\underline{a}} H(\underline{i},1) \sum_{i=1}^{m} C_i x_i H(1) \\
&- 2\rho\Big\{\underline{C}(\alpha+\beta_0) - \underline{C}\frac{\beta_0}{\gamma_0} H(0,1) - \underline{C} \sum_{i=1}^{m} \frac{1}{a_i} H(i,1)\Big\} H(1) \\
={}& -\Big\{2 - \frac{1}{\underline{a}} H(\underline{i},1)\Big\} \overline{C} \frac{\beta_0}{\gamma_0} H(0,2) - \{\overline{C}(\alpha+\beta_0+1) - C.\} \frac{1}{\underline{a}} H(\underline{i},1) H(1) \\
&+ 2\rho\underline{C}\frac{\beta_0}{\gamma_0} H(0,1) H(1) + 2\rho\underline{C} \sum_{i=1}^{m} \frac{1}{a_i} H(i,1) H(1) \\
&- 2\Big(1 - \frac{1}{2}\frac{\overline{a}}{\underline{a}} - \rho\Big) \sum_{i=1}^{m} C_i \frac{x_i}{a_i} H(i,1) H(1) \\
\leq{}& -\Big\{2 - \frac{1}{\underline{a}} H(\underline{i},1)\Big\} \overline{C} \frac{\beta_0}{\gamma_0} H(0,2) - \{\overline{C}(\alpha+\beta_0+1) - C.\} \frac{1}{\underline{a}} H(\underline{i},1) H(1) \\
&+ 2\rho\underline{C}\frac{\beta_0}{\gamma_0} H(0,1) H(1) + 2\rho\underline{C} m \frac{1}{\underline{a}} H(\underline{i},1) H(1) \\
&- 2\Big(1 - \frac{1}{2}\frac{\overline{a}}{\underline{a}} - \rho\Big) \underline{C}\frac{\underline{a}}{\overline{a}}\frac{1}{\underline{a}} H(\underline{i},1) H(1)
\end{aligned}$$

since $0 \leq \rho \leq 1 - (1/2)(\overline{a}/\underline{a})$ by assumption and since $\boldsymbol{x} \neq \mathbf{0}$. Now since $(1/\underline{a})H(\underline{i},1) \leq 1$ and since

$$\frac{1}{\gamma_0} H(0,2) \geq \frac{1}{\gamma_0} H(0,1) H(1) \geq \frac{1}{\underline{a}} H(\underline{i},1) H(1)$$

by part (ii) of Lemma 2.2.3, it follows that

$$\begin{aligned}
&- \Big\{2 - \frac{1}{\underline{a}} H(\underline{i},1)\Big\} \overline{C} \frac{\beta_0}{\gamma_0} H(0,2) + 2\rho\underline{C}\frac{\beta_0}{\gamma_0} H(0,1) H(1) \\
&\leq -\overline{C}\frac{\beta_0}{\gamma_0} H(0,1) H(1) + 2\rho\underline{C}\frac{\beta_0}{\gamma_0} H(0,1) H(1) \\
&\leq -(\overline{C} - 2\rho\underline{C})\frac{\beta_0}{\underline{a}} H(\underline{i},1) H(1),
\end{aligned} \tag{2.3.12}$$

where we have used the fact that $\overline{C} - 2\rho\underline{C} \geq \underline{C}(1-2\rho) \geq 0$ by assumption. Thus,

$$I/\Big\{\frac{1}{\underline{a}}H(\underline{i},1)H(1)\Big\} \leq -(\overline{C}-2\rho\underline{C})\beta_0 - \{\overline{C}(\alpha+\beta_0+1) - C.\} + 2\rho\underline{C}m \\ - 2\Big(1 - \frac{1}{2}\frac{\overline{a}}{\underline{a}} - \rho\Big)\underline{C}\frac{\underline{a}}{\overline{a}}.$$

The right-hand side of the above inequality is not positive by assumption (2.3.3). This completes the proof of part (ii). □

Remark 2.3.1 The major difference of the setting considered above from that considered by Hamura and Kubokawa (2019b) is that now the parameter β_0 may take on positive values in (2.3.8), yielding the additional terms in (2.3.9). In the present setting, we need to evaluate the factor $(1/\gamma_0)H(0,2)$ appropriately. Indeed, if (2.3.9) is replaced by

$$I_1 - 2I_2 \leq -\Big\{2 - \frac{1}{\underline{a}}H(\underline{i},1)\Big\}\sum_{i=1}^{m} C_i x_i H(1) \\ + \Big\{2 - \frac{1}{\underline{a}}H(\underline{i},1)\Big\}\{\overline{C}(\alpha+\beta_0+1) - C.\}H(1),$$

then it leads to a sufficient condition that is incompatible with the condition for propriety given in Lemma 2.2.1. Note also that the term $\overline{C}(\alpha+\beta_0+1) - C.$ is positive if the prior is proper satisfying $\alpha < m < \alpha+\beta_0$, while it is nonpositive in the case they consider. Since $(1/\gamma_0)H(0,2)$ can be very small compared to $H(1)$ in general, it is not straightforward to extend their results to the case of proper priors. We evaluate the third term on the right side of (2.3.10) by using (2.3.11), and then apply part (ii) of Lemma 2.2.3 to the second term in (2.3.10) in order to evaluate the secondary terms deriving from the last two terms in (2.3.11). Thus, Lemma 2.2.3 is important for the above proof of the existence of a heterogeneous shrinkage estimator that is both admissible and minimax.

Remark 2.3.2 In theory, we can obtain a sufficient condition that generalizes part 4 of Theorem 2.5 of Clevenson and Zidek (1975). By part (i) of Lemma 2.2.2, we have

$$\alpha+\beta_0 = \sum_{i=1}^{m}\frac{x_i+1}{a_i}H(i,1) + \frac{\beta_0}{\gamma_0}H(0,1) \geq (x.+m+\beta_0)\frac{1}{\overline{a}}H(\bar{i},1),$$
$$(\alpha+\beta_0+1)H(1) = \sum_{i=1}^{m}\frac{x_i+1}{a_i}H(i,2) + \frac{\beta_0}{\gamma_0}H(0,2) \leq (x.+m+\beta_0)\frac{1}{\gamma_0}H(0,2),$$

where $x. = \sum_{i=1}^{m} x_i$ and $\bar{i} \in \{1,\ldots,m\}$ is an index such that $a_{\bar{i}} = \overline{a}$. Therefore, it follows that

$$\frac{1}{\underline{a}}H(\underline{i},1) \leq \frac{\overline{a}}{\underline{a}}\frac{\alpha+\beta_0}{x.+m+\beta_0} \leq \frac{\overline{a}}{\underline{a}}\frac{\alpha+\beta_0}{1+m+\beta_0}$$

and that

$$\frac{1}{\gamma_0}H(0,2) \geq \frac{\alpha+\beta_0+1}{\alpha+\beta_0}\frac{1}{\overline{a}}H(\bar{i},1)H(1) \geq \frac{\alpha+\beta_0+1}{\alpha+\beta_0}\frac{\gamma_0}{\overline{a}}\frac{1}{\gamma_0}H(0,1)H(1).$$

Hence, (2.3.12) can be replaced by

$$
\begin{aligned}
&-\Big\{2-\frac{1}{\underline{a}}H(\underline{i},1)\Big\}\overline{C}\frac{\beta_0}{\gamma_0}H(0,2)+2\rho\underline{C}\frac{\beta_0}{\gamma_0}H(0,1)H(1)\\
&\le-\Big[2-\min\Big\{1,\frac{\overline{a}}{\underline{a}}\frac{\alpha+\beta_0}{1+m+\beta_0}\Big\}\Big]\overline{C}\frac{\beta_0}{\gamma_0}H(0,1)H(1)\max\Big\{1,\frac{\alpha+\beta_0+1}{\alpha+\beta_0}\frac{\gamma_0}{\overline{a}}\Big\}\\
&\quad+2\rho\underline{C}\frac{\beta_0}{\gamma_0}H(0,1)H(1)\\
&=-\Big(\Big[2\overline{C}-\overline{C}\min\Big\{1,\frac{\overline{a}}{\underline{a}}\frac{\alpha+\beta_0}{1+m+\beta_0}\Big\}\Big]\max\Big\{1,\frac{\alpha+\beta_0+1}{\alpha+\beta_0}\frac{\gamma_0}{\overline{a}}\Big\}-2\rho\underline{C}\Big)\\
&\quad\times\frac{\beta_0}{\gamma_0}H(0,1)H(1)\\
&\le-\Big(\Big[2\overline{C}-\overline{C}\min\Big\{1,\frac{\overline{a}}{\underline{a}}\frac{\alpha+\beta_0}{1+m+\beta_0}\Big\}\Big]\max\Big\{1,\frac{\alpha+\beta_0+1}{\alpha+\beta_0}\frac{\gamma_0}{\overline{a}}\Big\}-2\rho\underline{C}\Big)\\
&\quad\times\frac{\beta_0}{\underline{a}}H(\underline{i},1)H(1),
\end{aligned}
$$

which leads to a condition generalizing the sufficient condition of Clevenson and Zidek (1975) for the balanced case.

Remark 2.3.3 The class of proper Bayes minimax estimators will be broadened by replacing the factor $u^{\alpha-1+\beta_0}/(1+u/\gamma_0)^{\beta_0}$ in (2.2.2) with $u^{\beta_\cdot}\psi(u)$, where ψ is a proper density on $(0,\infty)$. This class of priors is considered by Ghosh and Parsian (1981) for the balanced case with $\boldsymbol{\beta}=\boldsymbol{\gamma}=\boldsymbol{j}$. One choice for ψ is the exponential density $\psi(u)=e^{-u/\gamma_0}$ for $u>0$. The details are omitted.

2.4 Simulation Study

In this section, we investigate through simulation the numerical performance of the risk functions of the Bayes estimators given in Section 2.2 under the loss function $L_{\boldsymbol{c}}$ given by (2.1.5) with $\boldsymbol{c}=\boldsymbol{n}$ or $\boldsymbol{c}=\boldsymbol{j}$. For the case of $\boldsymbol{c}=\boldsymbol{n}$, the estimators which we compare are the following five:

ML: the ML estimator $\widehat{\boldsymbol{\lambda}}^{\mathrm{ML}}=(X_1/n_1,\ldots,X_m/n_m)$,

PB1: the proper Bayes estimator $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB1}}=\widehat{\boldsymbol{\lambda}}^{(m-1,\boldsymbol{j},\boldsymbol{\nu};2,1)}$ given by (2.1.2) with $\beta_0=2$,

GB1: the generalized Bayes estimator $\widehat{\boldsymbol{\lambda}}^{\mathrm{GB1}}=\widehat{\boldsymbol{\lambda}}^{(m-1,\boldsymbol{j},\boldsymbol{\nu};0,1)}$ given by (2.1.2) with $\beta_0=0$,

PB2: the proper Bayes estimator $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB2}}=\widehat{\boldsymbol{\lambda}}^{(m-1,\boldsymbol{j},\boldsymbol{j};2,\underline{n})}$ given by (2.2.12) with $(\alpha,\beta_0,\gamma_0)=(m-1,2,\underline{n})$,

GB2: the generalized Bayes estimator $\widehat{\boldsymbol{\lambda}}^{\mathrm{GB2}}=\widehat{\boldsymbol{\lambda}}^{(m-1,\boldsymbol{j},\boldsymbol{j};0,\underline{n})}$ given by (2.2.12) with $(\alpha,\beta_0,\gamma_0)=(m-1,0,\underline{n})$.

For the case of $\boldsymbol{c}=\boldsymbol{j}$, the estimators which we compare are the above five estimators and the following two:

PB3: the proper Bayes estimator $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB3}}=\widehat{\boldsymbol{\lambda}}^{(m-1,\boldsymbol{j},\boldsymbol{\nu}\circ\boldsymbol{\nu};2,1/\overline{n})}$ given by (2.2.11) with $\boldsymbol{\beta}=\boldsymbol{j}$, $\boldsymbol{\gamma}=\boldsymbol{\nu}\circ\boldsymbol{\nu}$, and $(\alpha,\beta_0,\gamma_0)=(m-1,2,1/\overline{n})$,

GB3: the generalized Bayes estimator $\widehat{\boldsymbol{\lambda}}^{\mathrm{GB3}} = \widehat{\boldsymbol{\lambda}}^{(m-1,\boldsymbol{j},\boldsymbol{\nu}\circ\boldsymbol{\nu};0,1/\overline{n})}$ given by (2.2.11) with $\boldsymbol{\beta} = \boldsymbol{j}$, $\boldsymbol{\gamma} = \boldsymbol{\nu} \circ \boldsymbol{\nu}$, and $(\alpha, \beta_0, \gamma_0) = (m-1, 0, 1/\overline{n})$.

The imbalanced cases in $\boldsymbol{a} = (n_1\gamma_1, \ldots, n_m\gamma_m)$ and $\boldsymbol{C} = (c_1/({n_1}^2\gamma_1), \ldots, c_m/({n_m}^2\gamma_m))$ are summarized in Table 2.1. We consider the two estimators $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB3}}$ and $\widehat{\boldsymbol{\lambda}}^{\mathrm{GB3}}$ for $\boldsymbol{c} = \boldsymbol{j}$ in order to include the case where $\boldsymbol{C} = \boldsymbol{j}$.

Table 2.1: Imbalanced cases in $\boldsymbol{a}$ and $\boldsymbol{C}$.

$\boldsymbol{c}$	$\boldsymbol{\gamma}$	$\boldsymbol{a}$	$\boldsymbol{C}$
$\boldsymbol{n}$	$\boldsymbol{\nu}$	$\boldsymbol{j}$	$\boldsymbol{j}$
$\boldsymbol{n}$	$\boldsymbol{j}_m$	$\boldsymbol{n}$	$\boldsymbol{\nu}$
$\boldsymbol{j}_m$	$\boldsymbol{\nu}$	$\boldsymbol{j}$	$\boldsymbol{\nu}$
$\boldsymbol{j}_m$	$\boldsymbol{j}_m$	$\boldsymbol{n}$	$\boldsymbol{\nu} \circ \boldsymbol{\nu}$
$\boldsymbol{j}_m$	$\boldsymbol{\nu} \circ \boldsymbol{\nu}$	$\boldsymbol{\nu}$	$\boldsymbol{j}$

When $\boldsymbol{c} = \boldsymbol{n}$, the homogeneous proper Bayes estimator $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB1}}$ is always admissible and, by part (ii) of Theorem 2.3.1, minimax. On the other hand, the heterogeneous proper Bayes estimator $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB2}}$ is admissible, but the minimaxity is not clear, because Theorem 2.3.1 cannot always be applied when $\boldsymbol{n}$ is unbalanced. However, the conditions for the minimaxity of $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB2}}$ given in Theorem 2.3.1 are somewhat restrictive especially when the sample sizes are unbalanced, and it is worth investigating the performance of $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB2}}$. The generalized Bayes estimators $\widehat{\boldsymbol{\lambda}}^{\mathrm{GB1}}$, $\widehat{\boldsymbol{\lambda}}^{\mathrm{GB2}}$, and $\widehat{\boldsymbol{\lambda}}^{\mathrm{GB3}}$ are similar to the corresponding proper Bayes estimators $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB1}}$, $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB2}}$, and $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB3}}$ but whether or not the generalized Bayes estimators are admissible is not clear.

We set $m = 30$ and $(n_i, \lambda_i) = (\underline{n}, \lambda^{(1)})$ for $i = 1, \ldots, 15$ and $(n_i, \lambda_i) = (\overline{n}, \lambda^{(2)})$ for $i = 16, \ldots, 30$ and we generate random numbers of $\boldsymbol{X}$ for $(\underline{n}, \overline{n}) = (1,1), (0.5,2), (0.1,10)$ and $(\lambda^{(1)}, \lambda^{(2)}) = (1,1), (3,3), (1,3), (3,1)$. For each estimator $\widehat{\boldsymbol{\lambda}}$, we obtain approximated values of the risk function $E_{\boldsymbol{\lambda}}[L_{\boldsymbol{n}}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda})]$ by simulation with $100,000$ replications. The integrals are calculated via the Monte Carlo simulation with $100,000$ replications. The percentage relative improvement in average loss (PRIAL) of an estimator $\widehat{\boldsymbol{\lambda}}$ over $\widehat{\boldsymbol{\lambda}}^{\mathrm{ML}}$ is defined by

$$\mathrm{PRIAL} = 100\{E_{\boldsymbol{\lambda}}[L_{\boldsymbol{n}}(\widehat{\boldsymbol{\lambda}}^{\mathrm{ML}}, \boldsymbol{\lambda})] - E_{\boldsymbol{\lambda}}[L_{\boldsymbol{n}}(\widehat{\boldsymbol{\lambda}}, \boldsymbol{\lambda})]\}/E_{\boldsymbol{\lambda}}[L_{\boldsymbol{n}}(\widehat{\boldsymbol{\lambda}}^{\mathrm{ML}}, \boldsymbol{\lambda})].$$

For the case of $\boldsymbol{c} = \boldsymbol{n}$, Table 2.2 reports values of the risks of the estimators with values of PRIAL given in parentheses. When $(\underline{n}, \overline{n}) = (1,1)$, the risk values of $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB1}}$ and $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB2}}$ are the same because $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB1}} = \widehat{\boldsymbol{\lambda}}^{\mathrm{PB2}}$. When $(\underline{n}, \overline{n}) = (0.5,2)$, the risk values of $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB2}}$ are smaller than those of $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB1}}$ except when $(\lambda^{(1)}, \lambda^{(2)}) = (3,1)$. When $(\underline{n}, \overline{n}) = (0.1,10)$, all risk the values of $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB2}}$ are much smaller than those of $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB1}}$, and the improvement of $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB2}}$ is significant. In addition, when $(\underline{n}, \overline{n}) = (0.1,10)$, $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB2}}$ has the largest values of PRIAL while $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB1}}$ has the smallest values of PRIAL. These results suggest that the heterogeneous shrinkage estimators can enjoy substantial improvement over the homogeneous shrinkage estimators in the more unbalanced cases. The risk values of the proper Bayes estimators are almost the same as the corresponding risk values of their generalized Bayes counterparts.

Table 2.2: Risks of the estimators ML, PB1, GB1, PB2, and GB2 for $\boldsymbol{c} = \boldsymbol{n}$. (Values of PRIAL of PB1, GB1, PG2, and GB2 are given in parentheses.)

$(\underline{n}, \overline{n})$	$(\lambda^{(1)}, \lambda^{(2)})$	ML	PB1	GB1	PB2	GB2
$(1, 1)$	$(1, 1)$	30.01	15.35 (48.83)	15.37 (48.77)	15.35 (48.83)	15.37 (48.77)
	$(3, 3)$	29.99	22.83 (23.89)	22.82 (23.91)	22.83 (23.89)	22.82 (23.92)
	$(1, 3)$	30.00	20.38 (32.09)	20.38 (32.09)	20.38 (32.08)	20.38 (32.08)
	$(3, 1)$	30.03	20.37 (32.16)	20.37 (32.15)	20.37 (32.16)	20.37 (32.15)
$(0.5, 2)$	$(1, 1)$	30.00	17.03 (43.21)	17.05 (43.17)	15.34 (48.88)	15.35 (48.83)
	$(3, 3)$	30.00	23.99 (20.03)	23.98 (20.06)	22.16 (26.14)	22.16 (26.13)
	$(1, 3)$	29.98	23.24 (22.47)	23.24 (22.49)	19.16 (36.09)	19.21 (35.93)
	$(3, 1)$	30.00	19.47 (35.10)	19.47 (35.09)	20.53 (31.57)	20.50 (31.66)
$(0.1, 10)$	$(1, 1)$	30.11	25.37 (15.73)	25.36 (15.75)	15.18 (49.56)	15.19 (49.56)
	$(3, 3)$	30.00	28.25 (5.85)	28.24 (5.87)	18.05 (39.85)	18.05 (39.84)
	$(1, 3)$	29.96	28.22 (5.80)	28.21 (5.83)	16.23 (45.82)	16.25 (45.77)
	$(3, 1)$	29.99	25.37 (15.42)	25.36 (15.45)	17.54 (41.53)	17.53 (41.54)

For the case of $\boldsymbol{c} = \boldsymbol{j}$, Table 2.3 reports values of the risks of the estimators with values of PRIAL given in parentheses. The performance of the five estimators $\widehat{\boldsymbol{\lambda}}^{\mathrm{ML}}$, $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB1}}$, $\widehat{\boldsymbol{\lambda}}^{\mathrm{GB1}}$, $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB2}}$, and $\widehat{\boldsymbol{\lambda}}^{\mathrm{GB2}}$ is almost the same as in the previous case. The estimators $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB3}}$ and $\widehat{\boldsymbol{\lambda}}^{\mathrm{GB3}}$, which satisfy the condition $C_1 = \cdots = C_m$, have the largest risk values for $(\underline{n}, \overline{n}) = (0.5, 2), (0.1, 10)$. In particular, when $(\underline{n}, \overline{n}) = (0.1, 10)$, these estimators have the values of PRIAL almost equal to zero.

Table 2.3: Risks of the estimators ML, PB1, GB1, PB2, GB2, PB3, and GB3 for $\boldsymbol{c} = \boldsymbol{j}$. (Values of PRIAL of PB1, GB1, PG2, GB2, PB3, and GB3 are given in parentheses.)

$(\underline{n}, \overline{n})$	$(\lambda^{(1)}, \lambda^{(2)})$	ML	PB1	GB1	PB2	GB2	PB3	GB3
$(1, 1)$	$(1, 1)$	30.01	15.37 (48.79)	15.39 (48.73)	15.37 (48.79)	15.38 (48.74)	15.37 (48.79)	15.38 (48.74)
	$(3, 3)$	30.03	22.86 (23.89)	22.85 (23.91)	22.86 (23.88)	22.85 (23.90)	22.86 (23.88)	22.85 (23.90)
	$(1, 3)$	30.04	20.38 (32.13)	20.38 (32.13)	20.38 (32.14)	20.38 (32.14)	20.38 (32.14)	20.38 (32.14)
	$(3, 1)$	30.04	20.38 (32.16)	20.38 (32.16)	20.38 (32.17)	20.38 (32.16)	20.38 (32.17)	20.38 (32.16)
$(0.5, 2)$	$(1, 1)$	37.52	17.76 (52.66)	18.01 (51.98)	15.30 (59.21)	15.32 (59.16)	24.55 (34.57)	24.82 (33.84)
	$(3, 3)$	37.57	27.54 (26.70)	27.77 (26.08)	24.79 (34.00)	24.81 (33.96)	32.99 (12.18)	33.07 (11.98)
	$(1, 3)$	37.61	25.35 (32.60)	25.69 (31.68)	18.44 (50.98)	18.63 (50.46)	32.73 (12.98)	32.82 (12.73)
	$(3, 1)$	37.54	23.55 (37.25)	23.62 (37.07)	25.10 (33.13)	24.91 (33.65)	26.86 (28.44)	27.03 (27.98)
$(0.1, 10)$	$(1, 1)$	151.49	105.39 (30.43)	107.62 (28.96)	15.06 (90.06)	15.07 (90.05)	150.85 (0.43)	150.86 (0.42)
	$(3, 3)$	151.62	133.27 (12.10)	134.32 (11.41)	36.35 (76.02)	36.40 (75.99)	151.42 (0.13)	151.42 (0.13)
	$(1, 3)$	152.12	133.46 (12.27)	134.54 (11.56)	17.90 (88.23)	18.09 (88.11)	151.93 (0.13)	151.93 (0.13)
	$(3, 1)$	151.37	106.84 (29.42)	108.95 (28.02)	38.55 (74.53)	38.47 (74.59)	150.72 (0.43)	150.74 (0.42)

2.5 Application

In this section, several estimation methods considered in the previous sections are applied to data relating to the standardized mortality ratio (SMR). (For the SMR, see, for example, Clayton and

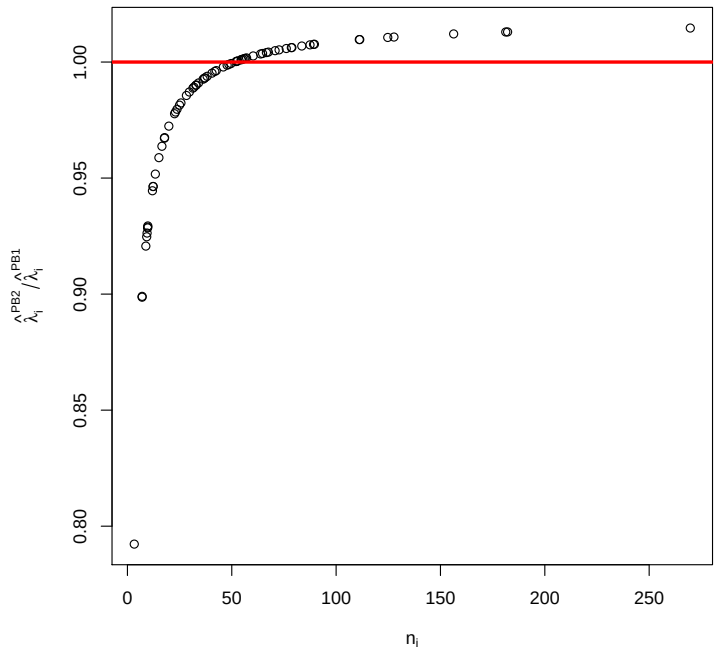

Figure 2.1: The ratio $\hat{\lambda}_i^{\mathrm{PB2}}/\hat{\lambda}_i^{\mathrm{PB1}}$.

Kaldor (1987).) More specifically, the data consist of actual and expected numbers of deaths of females from a specific cause in $m = 72$ districts in a prefecture in Japan during the 5 years from 2008 to 2012. For $i = 1, \ldots, m$, the actual and expected numbers of deaths in the i-th district are denoted by x_i and n_i, respectively. Each component of an estimator $\boldsymbol{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_m)$ is a measure of relative risk in a district calculated from the data.

We here consider only the three estimators $\widehat{\boldsymbol{\lambda}}^{\mathrm{ML}} = (\hat{\lambda}_1^{\mathrm{ML}}, \ldots, \hat{\lambda}_m^{\mathrm{ML}})$, $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB1}} = (\hat{\lambda}_1^{\mathrm{PB1}}, \ldots, \hat{\lambda}_m^{\mathrm{PB1}})$, and $\widehat{\boldsymbol{\lambda}}^{\mathrm{PB2}} = (\hat{\lambda}_1^{\mathrm{PB2}}, \ldots, \hat{\lambda}_m^{\mathrm{PB2}})$ given in Section 2.4. Integrals are calculated via the Monte Carlo simulation with $100,000$ replications. The data and the estimates for all the $m = 72$ districts are given in Tables 2.4 and 2.5.

The values of the ratio $\hat{\lambda}_i^{\mathrm{PB2}}/\hat{\lambda}_i^{\mathrm{PB1}}$ for all $i = 1, \ldots, m$ are plotted in Figure 2.1. For $i = 1, \ldots, m$, the heterogeneous estimator $\hat{\lambda}_i^{\mathrm{PB2}}$ shrinks the ML estimator $\hat{\lambda}_i^{\mathrm{ML}}$ toward the origin more than the homogeneous estimator $\hat{\lambda}_i^{\mathrm{PB1}}$ if $n_i \lessapprox 50$ and less than $\hat{\lambda}_i^{\mathrm{PB1}}$ if $n_i \gtrapprox 50$.

Table 2.4: The data and the estimates of relative risk for $i = 1, \ldots, 36$.

i	x_i	n_i	$\hat{\lambda}_i^{\mathrm{ML}}$	$\hat{\lambda}_i^{\mathrm{PB1}}$	$\hat{\lambda}_i^{\mathrm{PB2}}$
1	49	49.65	0.99	0.97	0.97
2	64	66.60	0.96	0.94	0.95
3	69	63.83	1.08	1.06	1.07
4	79	87.49	0.90	0.89	0.89
5	47	48.60	0.97	0.95	0.95
6	35	42.58	0.82	0.81	0.80
7	66	76.12	0.87	0.85	0.86
8	75	72.67	1.03	1.01	1.02
9	49	55.24	0.89	0.87	0.87
10	54	64.75	0.83	0.82	0.82
11	192	182.16	1.05	1.04	1.05
12	349	269.71	1.29	1.27	1.29
13	48	40.54	1.18	1.16	1.16
14	47	45.94	1.02	1.00	1.00
15	62	54.53	1.14	1.12	1.12
16	38	32.79	1.16	1.14	1.13
17	31	31.41	0.99	0.97	0.96
18	81	78.79	1.03	1.01	1.02
19	57	52.49	1.09	1.07	1.07
20	62	57.09	1.09	1.07	1.07
21	21	23.03	0.91	0.90	0.88
22	83	67.53	1.23	1.21	1.21
23	116	111.32	1.04	1.02	1.03
24	51	41.87	1.22	1.20	1.19
25	41	36.28	1.13	1.11	1.10
26	21	17.72	1.19	1.16	1.13
27	59	47.77	1.24	1.21	1.21
28	13	9.42	1.38	1.36	1.26
29	20	11.98	1.67	1.64	1.55
30	22	23.76	0.93	0.91	0.89
31	14	15.09	0.93	0.91	0.87
32	23	13.38	1.72	1.69	1.61
33	14	9.72	1.44	1.42	1.31
34	5	3.28	1.53	1.50	1.19
35	52	52.79	0.99	0.97	0.97
36	6	7.03	0.85	0.84	0.75

Table 2.5: The data and the estimates of relative risk for $i = 37, \ldots, 72$.

i	x_i	n_i	$\hat{\lambda}_i^{\mathrm{ML}}$	$\hat{\lambda}_i^{\mathrm{PB1}}$	$\hat{\lambda}_i^{\mathrm{PB2}}$
37	7	9.78	0.72	0.70	0.65
38	7	7.05	0.99	0.98	0.88
39	10	12.32	0.81	0.80	0.75
40	41	54.59	0.75	0.74	0.74
41	15	9.24	1.62	1.59	1.47
42	11	9.67	1.14	1.12	1.04
43	15	17.81	0.84	0.83	0.80
44	119	124.74	0.95	0.94	0.95
45	103	89.18	1.16	1.13	1.14
46	25	24.95	1.00	0.98	0.97
47	61	55.91	1.09	1.07	1.07
48	83	70.76	1.17	1.15	1.16
49	45	36.92	1.22	1.20	1.19
50	141	127.72	1.10	1.08	1.10
51	151	156.31	0.97	0.95	0.96
52	22	16.53	1.33	1.31	1.26
53	98	83.55	1.17	1.15	1.16
54	37	38.18	0.97	0.95	0.95
55	39	32.97	1.18	1.16	1.15
56	29	28.27	1.03	1.01	0.99
57	20	19.84	1.01	0.99	0.96
58	21	25.64	0.82	0.80	0.79
59	72	52.02	1.38	1.36	1.36
60	19	31.88	0.60	0.59	0.58
61	29	22.59	1.28	1.26	1.23
62	15	8.82	1.70	1.67	1.54
63	9	12.31	0.73	0.72	0.68
64	118	111.11	1.06	1.04	1.05
65	52	37.12	1.40	1.38	1.37
66	59	60.27	0.98	0.96	0.96
67	30	29.67	1.01	0.99	0.98
68	155	181.29	0.86	0.84	0.85
69	51	56.42	0.90	0.89	0.89
70	75	89.61	0.84	0.82	0.83
71	75	78.53	0.96	0.94	0.94
72	43	34.05	1.26	1.24	1.23

2.6 Extensions

2.6.1 Empirical Bayes estimators

In addition to the hierarchical Bayes estimators, we can also derive empirical Bayes estimators satisfying minimaxity. If we set $\boldsymbol{\beta} = \boldsymbol{j}$ in (2.2.2), we have

$$\pi_{\alpha,\boldsymbol{j},\boldsymbol{\gamma};\beta_0,\gamma_0}(\boldsymbol{\lambda}) \propto \int_0^\infty \frac{u^{\alpha+\beta_0-m-1}}{(1+u/\gamma_0)^{\beta_0}} \Big\{ \prod_{i=1}^m \frac{u}{\gamma_i} e^{-(u/\gamma_i)\lambda_i} \Big\} du,$$

which is regarded as a mixture distribution. In this subsection, we consider the variable u in the integration as an unknown hyper-parameter. Then the prior of λ_i has a gamma density proportional to $e^{-(u/\gamma_i)\lambda_i}$, and the resulting subjective Bayes estimator is $\widehat{\boldsymbol{\lambda}}^{\mathrm{B}}(u) = (\hat{\lambda}_1^{\mathrm{B}}(u), \ldots, \hat{\lambda}_m^{\mathrm{B}}(u))$ for

$$\hat{\lambda}_i^{\mathrm{B}}(u) = \frac{X_i}{n_i}\Big(1 - \frac{1}{n_i\gamma_i u^{-1} + 1}\Big).$$

Since

$$\begin{aligned}
&\int \cdots \int_{(0,\infty)^m} E_{\boldsymbol{\lambda}}\Big[\sum_{i=1}^m \frac{\tilde{c}_i}{n_i} X_i\Big] \Big\{ \prod_{i=1}^m \frac{u}{\gamma_i} e^{-(u/\gamma_i)\lambda_i} \Big\} d\boldsymbol{\lambda} \\
&= \sum_{i=1}^m \Big\{ \tilde{c}_i \int_0^\infty \lambda_i \frac{u}{\gamma_i} e^{-(u/\gamma_i)\lambda_i} d\lambda_i \Big\} = u^{-1} \sum_{i=1}^m \tilde{c}_i \gamma_i
\end{aligned}$$

for $\tilde{\boldsymbol{c}} = (\tilde{c}_1, \ldots, \tilde{c}_m) \in (0,\infty)^m$, we can estimate u^{-1} by

$$\widehat{u^{-1}} = \frac{\sum_{i=1}^m (\tilde{c}_i/n_i) X_i}{\sum_{i=1}^m \tilde{c}_i \gamma_i},$$

which is substituted into $\widehat{\boldsymbol{\lambda}}^{\mathrm{B}}(u)$ to get the empirical Bayes estimator $\widehat{\boldsymbol{\lambda}}^{\mathrm{EB}} = (\hat{\lambda}_1^{\mathrm{EB}}, \ldots, \hat{\lambda}_m^{\mathrm{EB}}) = \widehat{\boldsymbol{\lambda}}^{\mathrm{B}}\big|_{u^{-1}=\widehat{u^{-1}}}$, where for $\widetilde{X} = \sum_{j=1}^m (\tilde{c}_j/n_j) X_j$,

$$\hat{\lambda}_i^{\mathrm{EB}} = \frac{X_i}{n_i} \frac{\widetilde{X}}{\widetilde{X} + \sum_{j=1}^m \tilde{c}_j\gamma_j/(n_i\gamma_i)} = \frac{X_i}{n_i}\Big(1 - \frac{\sum_{j=1}^m \tilde{c}_j\gamma_j}{n_i\gamma_i\widetilde{X} + \sum_{j=1}^m \tilde{c}_j\gamma_j}\Big). \tag{2.6.1}$$

In the case of $\tilde{c}_i = n_i$ and $\gamma_i = 1/n_i$, the empirical Bayes estimator is

$$\hat{\lambda}_i^{\mathrm{EB1}} = \frac{X_i}{n_i}\Big(1 - \frac{m}{X_\cdot + m}\Big), \tag{2.6.2}$$

which was given by Clevenson and Zidek (1975) when $n_1 = \cdots = n_m = 1$. When $\tilde{c}_i = n_i$ and $\gamma_i = 1$, the empirical Bayes estimator is

$$\hat{\lambda}_i^{\mathrm{EB2}} = \frac{X_i}{n_i}\Big(1 - \frac{n_\cdot}{n_i X_\cdot + n_\cdot}\Big), \tag{2.6.3}$$

where $n_\cdot = \sum_{j=1}^m n_j$. On the other hand, when $\tilde{c}_i = 1$, it becomes

$$\hat{\lambda}_i^{\mathrm{EB3}} = \frac{X_i}{n_i}\Big(1 - \frac{\sum_{j=1}^m 1/n_j}{\sum_{j=1}^m X_j/n_j + \sum_{j=1}^m 1/n_j}\Big) \tag{2.6.4}$$

if $\gamma_i = 1/n_i$ and

$$\hat{\lambda}_i^{\mathrm{EB4}} = \frac{X_i}{n_i}\Big(1 - \frac{m}{n_i \sum_{j=1}^m X_j/n_j + m}\Big) \tag{2.6.5}$$

if $\gamma_i = 1$. It is seen that $\hat{\lambda}_i^{\mathrm{EB2}}$ and $\hat{\lambda}_i^{\mathrm{EB4}}$ are heterogeneous estimators in the sense that they shrink $\hat{\lambda}_i^{\mathrm{ML}}$ more toward the origin when n_i is smaller.

In the following, we use the notation $\max(v_i) = \max_{1\le i\le m} v_i$ and $\min(v_i) = \min_{1\le i\le m} v_i$ for $(v_1, \ldots, v_m) \in \mathbb{R}^m$.

Theorem 2.6.1 *Suppose that*

$$2\sum_{i=1}^m \frac{c_i}{n_i} \ge \Big[\max\Big\{\frac{\max(n_i\gamma_i)}{\min(n_i\gamma_i)}, \frac{\max(\tilde{c}_i\gamma_i) + \sum_{i=1}^m \tilde{c}_i\gamma_i}{\min(\tilde{c}_i\gamma_i) + \sum_{i=1}^m \tilde{c}_i\gamma_i}\Big\}\Big]\Big\{\frac{\sum_{i=1}^m \tilde{c}_i\gamma_i}{\min(n_i\gamma_i)} \max\Big(\frac{c_i}{\tilde{c}_i}\Big) + 2\max\Big(\frac{c_i}{n_i}\Big)\Big\}. \tag{2.6.6}$$

Then the empirical estimator $\widehat{\boldsymbol{\lambda}}^{\mathrm{EB}}$ is minimax under the loss $L_{\boldsymbol{c}}$ given by (2.1.5).

In the case of $\tilde{c}_i = c_i = n_i$ for $i = 1, \ldots, m$, condition (2.6.6) is

$$2m \ge \frac{\max(n_i\gamma_i)}{\min(n_i\gamma_i)}\Big\{\frac{\sum_{i=1}^m n_i\gamma_i}{\min(n_i\gamma_i)} + 2\Big\}. \tag{2.6.7}$$

If in addition $\gamma_i = 1$ for $i = 1, \ldots, m$, this becomes

$$2m(\underline{n}/\overline{n}) \ge n./\underline{n} + 2, \tag{2.6.8}$$

where $\overline{n} = \max(n_i)$ and $\underline{n} = \min(n_i)$.

Proof of Theorem 2.6.1. Let $\overline{a} = \max(n_i\gamma_i)$ and $\underline{a} = \min(n_i\gamma_i)$, let $\overline{\tilde{b}} = \max(\tilde{c}_i\gamma_i)$, $\underline{\tilde{b}} = \min(\tilde{c}_i\gamma_i)$, and $\tilde{b}. = \sum_{i=1}^m \tilde{c}_i\gamma_i$, and let $\overline{e} = \max(c_i/n_i)$ and $e. = \sum_{i=1}^m c_i/n_i$. The risk difference $\Delta^{\mathrm{EB}} = E_{\boldsymbol{\lambda}}[L_{\boldsymbol{c}}(\widehat{\boldsymbol{\lambda}}^{\mathrm{EB}}, \boldsymbol{\lambda})] - E_{\boldsymbol{\lambda}}[L_{\boldsymbol{c}}(\widehat{\boldsymbol{\lambda}}^{\mathrm{ML}}, \boldsymbol{\lambda})]$ is

$$\begin{aligned}
\Delta^{\mathrm{EB}} &= E_{\boldsymbol{\lambda}}\Big[\sum_{i=1}^m \Big\{\frac{c_i}{\lambda_i}\Big(\frac{X_i}{n_i} - \lambda_i - \frac{X_i}{n_i}\frac{\tilde{b}.}{n_i\gamma_i\widetilde{X} + \tilde{b}.}\Big)^2 - \frac{c_i}{\lambda_i}\Big(\frac{X_i}{n_i} - \lambda_i\Big)^2\Big\}\Big] \\
&= E_{\boldsymbol{\lambda}}\Big[\sum_{i=1}^m \Big[\frac{n_ic_i}{n_i\lambda_i}\Big\{\Big(\frac{X_i}{n_i}\Big)^2\Big(\frac{\tilde{b}.}{n_i\gamma_i\widetilde{X} + \tilde{b}.}\Big)^2 - 2\Big(\frac{X_i}{n_i}\Big)^2\frac{\tilde{b}.}{n_i\gamma_i\widetilde{X} + \tilde{b}.}\Big\} + 2\frac{c_i}{n_i}X_i\frac{\tilde{b}.}{n_i\gamma_i\widetilde{X} + \tilde{b}.}\Big]\Big].
\end{aligned} \tag{2.6.9}$$

From (2.6.9) and Lemma 2.3.1, it follows that

$$\begin{aligned}
\Delta^{\mathrm{EB}} &= E_{\boldsymbol{\lambda}}\Big[\sum_{i=1}^m \Big[\frac{n_ic_i}{X_i+1}\Big\{\Big(\frac{X_i+1}{n_i}\Big)^2\Big(\frac{\tilde{b}.}{n_i\gamma_i\widetilde{X} + \tilde{c}_i\gamma_i + \tilde{b}.}\Big)^2 - 2\Big(\frac{X_i+1}{n_i}\Big)^2\frac{\tilde{b}.}{n_i\gamma_i\widetilde{X} + \tilde{c}_i\gamma_i + \tilde{b}.}\Big\} \\
&\quad + 2\frac{c_i}{n_i}X_i\frac{\tilde{b}.}{n_i\gamma_i\widetilde{X} + \tilde{c}_i\gamma_i + \tilde{b}.} + 2\frac{c_i}{n_i}X_i\Big(\frac{\tilde{b}.}{n_i\gamma_i\widetilde{X} + \tilde{b}.} - \frac{\tilde{b}.}{n_i\gamma_i\widetilde{X} + \tilde{c}_i\gamma_i + \tilde{b}.}\Big)\Big]\Big] \\
&= E_{\boldsymbol{\lambda}}\Big[\sum_{i=1}^m \frac{c_i}{n_i}\frac{(X_i+1)\tilde{b}.^2}{(n_i\gamma_i\widetilde{X} + \tilde{c}_i\gamma_i + \tilde{b}.)^2} - 2\sum_{i=1}^m \frac{c_i}{n_i}\frac{\tilde{b}.}{n_i\gamma_i\widetilde{X} + \tilde{c}_i\gamma_i + \tilde{b}.} \\
&\quad + 2\sum_{i=1}^m \frac{c_i}{n_i}X_i\frac{\tilde{b}.\tilde{c}_i\gamma_i}{(n_i\gamma_i\widetilde{X} + \tilde{b}.)(n_i\gamma_i\widetilde{X} + \tilde{c}_i\gamma_i + \tilde{b}.)}\Big].
\end{aligned}$$

Hence, letting $\overline{\rho} = \max(c_i/\tilde{c}_i)$, the risk difference is evaluated as

$$\begin{aligned}
\Delta^{\mathrm{EB}} &\le E_{\boldsymbol{\lambda}}\Big[\frac{(\overline{\rho}\widetilde{X}+e.)\tilde{b}_\cdot^2}{(\underline{a}\widetilde{X}+\underline{\tilde{b}}+\tilde{b}.)^2} - 2\frac{e.\tilde{b}.}{\overline{a}\widetilde{X}+\overline{\tilde{b}}+\tilde{b}.} + 2\frac{\tilde{b}.}{\underline{a}\widetilde{X}+\underline{\tilde{b}}+\tilde{b}.}\sum_{i=1}^m \frac{c_i}{n_i}X_i\frac{\tilde{c}_i\gamma_i}{n_i\gamma_i\widetilde{X}+\tilde{b}.}\Big] \\
&\le E_{\boldsymbol{\lambda}}\Big[\frac{\tilde{b}_\cdot^2}{\underline{a}^2}\frac{\overline{\rho}\widetilde{X}+e.}{(\widetilde{X}+\underline{\tilde{b}}/\underline{a}+\tilde{b}./\underline{a})^2} - 2\frac{e.\tilde{b}.}{\overline{a}\widetilde{X}+\overline{\tilde{b}}+\tilde{b}.} + 2\frac{\tilde{b}.}{\underline{a}\widetilde{X}+\underline{\tilde{b}}+\tilde{b}.}\sum_{i=1}^m \overline{e}\frac{X_i(\tilde{c}_i/n_i)}{\widetilde{X}+\tilde{b}./\overline{a}}\Big] \\
&\le E_{\boldsymbol{\lambda}}\Big[\frac{\tilde{b}_\cdot^2}{\underline{a}^2}\frac{\overline{\rho}}{\widetilde{X}+\underline{\tilde{b}}/\underline{a}+\tilde{b}./\underline{a}} - 2\frac{e.\tilde{b}.}{\overline{a}\widetilde{X}+\overline{\tilde{b}}+\tilde{b}.} + 2\frac{\tilde{b}.\overline{e}}{\underline{a}\widetilde{X}+\underline{\tilde{b}}+\tilde{b}.}\Big], \qquad (2.6.10)
\end{aligned}$$

where the last inequality follows since

$$\frac{\overline{\rho}\widetilde{X}+e.}{\widetilde{X}+\underline{\tilde{b}}/\underline{a}+\tilde{b}./\underline{a}} \le \overline{\rho}\frac{\widetilde{X}+e./\overline{\rho}}{\widetilde{X}+\tilde{b}./\underline{a}} = \overline{\rho}\frac{\widetilde{X}+\sum_{i=1}^m(c_i/n_i)/\overline{\rho}}{\widetilde{X}+\sum_{i=1}^m \tilde{c}_i\gamma_i/\underline{a}} \le \overline{\rho}\frac{\widetilde{X}+\sum_{i=1}^m(c_i/n_i)/(c_i/\tilde{c}_i)}{\widetilde{X}+\sum_{i=1}^m \tilde{c}_i\gamma_i/(n_i\gamma_i)} = \overline{\rho}.$$

From (2.6.10), it is concluded that

$$\begin{aligned}
\Delta^{\mathrm{EB}} &\le \tilde{b}.E_{\boldsymbol{\lambda}}\Big[\Big(\frac{\tilde{b}.}{\underline{a}}\overline{\rho}+2\overline{e}\Big)\frac{1}{\underline{a}\widetilde{X}+\underline{\tilde{b}}+\tilde{b}.} - 2\frac{e.}{\overline{a}\widetilde{X}+\overline{\tilde{b}}+\tilde{b}.}\Big] \\
&= \tilde{b}.E_{\boldsymbol{\lambda}}\Big[\frac{1}{\underline{a}\widetilde{X}+\underline{\tilde{b}}+\tilde{b}.}\frac{1}{\overline{a}\widetilde{X}+\overline{\tilde{b}}+\tilde{b}.} \\
&\qquad \times \Big[\widetilde{X}\underline{a}\Big\{\frac{\overline{a}}{\underline{a}}\Big(\frac{\tilde{b}.}{\underline{a}}\overline{\rho}+2\overline{e}\Big)-2e.\Big\} + (\underline{\tilde{b}}+\tilde{b}.)\Big\{\frac{\overline{\tilde{b}}+\tilde{b}.}{\underline{\tilde{b}}+\tilde{b}.}\Big(\frac{\tilde{b}.}{\underline{a}}\overline{\rho}+2\overline{e}\Big)-2e.\Big\}\Big]\Big],
\end{aligned}$$

which is less than or equal to 0, because $2e. \ge [\max\{\overline{a}/\underline{a}, (\overline{\tilde{b}}+\tilde{b}.)/(\underline{\tilde{b}}+\tilde{b}.)\}]\{(\tilde{b}./\underline{a})\overline{\rho}+2\overline{e}\}$ by assumption. □

2.6.2 Estimation under the Kullback-Leibler loss

We can also evaluate the risk of the Bayes estimator with respect to the prior $\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}$ given by (2.2.2) under the loss function

$$\tilde{L}_{\boldsymbol{c}}(\boldsymbol{d},\boldsymbol{\lambda}) = \sum_{i=1}^m c_i\lambda_i\Big(\frac{d_i}{\lambda_i} - 1 - \log\frac{d_i}{\lambda_i}\Big) = \sum_{i=1}^m c_i\Big(d_i - \lambda_i - \lambda_i\log\frac{d_i}{\lambda_i}\Big), \qquad (2.6.11)$$

which is the loss function considered by Ghosh and Yang (1988) for the balanced case and by Hamura and Kubokawa (2019b) for the unbalanced case. The Bayes estimator is given by

$$\widetilde{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)} = \widetilde{\boldsymbol{\lambda}}^{(\boldsymbol{\beta})} \circ \Big(1-\tilde{\phi}_1^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{X}), \dots, 1-\tilde{\phi}_m^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{X})\Big)$$

for $\alpha < \beta.$, where $\widetilde{\boldsymbol{\lambda}}^{(\boldsymbol{\beta})} = ((X_1+\beta_1)/n_1, \dots, (X_m+\beta_m)/n_m)$ is the Bayes estimator against the improper prior $\pi_{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \prod_{i=1}^m {\lambda_i}^{\beta_i-1}$ and where

$$\tilde{\phi}_i^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{X}) = \frac{1}{n_i\gamma_i}\frac{K(\boldsymbol{n}\circ\boldsymbol{\gamma}, \boldsymbol{X}+\boldsymbol{\beta}+\boldsymbol{e}_i, \alpha+\beta_0+1; \gamma_0, \beta_0)}{K(\boldsymbol{n}\circ\boldsymbol{\gamma}, \boldsymbol{X}+\boldsymbol{\beta}, \alpha+\beta_0; \gamma_0, \beta_0)}$$

determines the amount of shrinkage for $i = 1, \ldots, m$. The prior $\pi_{\boldsymbol{\beta}}$ coincides with the Jeffreys prior when $\boldsymbol{\beta} = \boldsymbol{j}/2$. A calculation similar to that in the proof of Theorem 2.3.1 shows that the risk difference between the two estimators is

$$E_{\boldsymbol{\lambda}}[\tilde{L}_{\boldsymbol{c}}(\widetilde{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}, \boldsymbol{\lambda})] - E_{\boldsymbol{\lambda}}[L_{\boldsymbol{c}}(\widetilde{\boldsymbol{\lambda}}^{(\boldsymbol{\beta})}, \boldsymbol{\lambda})] = E_{\boldsymbol{\lambda}}[\tilde{D}^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{X})],$$

where $\tilde{D}^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{0}) = -\sum_{i=1}^m C_i\beta_i K(\boldsymbol{n}\circ\boldsymbol{\gamma}, \boldsymbol{\beta}+\boldsymbol{e}_i, \alpha+\beta_0+1; \gamma_0, \beta_0)/K(\boldsymbol{n}\circ\boldsymbol{\gamma}, \boldsymbol{\beta}, \alpha+\beta_0; \gamma_0, \beta_0)$ and

$$\begin{aligned}\tilde{D}^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}(\boldsymbol{x}) = &-\sum_{i=1}^m \frac{c_i}{n_i}\frac{x_i+\beta_i}{n_i\gamma_i}\frac{K(\boldsymbol{n}\circ\boldsymbol{\gamma}, \boldsymbol{x}+\boldsymbol{\beta}+\boldsymbol{e}_i, \alpha+\beta_0+1; \gamma_0, \beta_0)}{K(\boldsymbol{n}\circ\boldsymbol{\gamma}, \boldsymbol{x}+\boldsymbol{\beta}, \alpha+\beta_0; \gamma_0, \beta_0)} \\ &+\sum_{i=1}^m \frac{c_i}{n_i} x_i \log\Big\{1+\frac{1}{n_i\gamma_i}\frac{K(\boldsymbol{n}\circ\boldsymbol{\gamma}, \boldsymbol{x}+\boldsymbol{\beta}, \alpha+\beta_0+1; \gamma_0, \beta_0)}{K(\boldsymbol{n}\circ\boldsymbol{\gamma}, \boldsymbol{x}+\boldsymbol{\beta}, \alpha+\beta_0; \gamma_0, \beta_0)}\Big\}\end{aligned}$$

for $\boldsymbol{x} = (x_1, \ldots, x_m) \in {\mathbb{N}_0}^m \setminus \{\boldsymbol{0}\}$. Using Lemma 2.2.2 to evaluate the first term on the right and applying the inequality $\log(1+\xi) \le \xi$ for $\xi \ge 0$ to the second term will lead to a sufficient condition for $\widetilde{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}$ to improve on $\widetilde{\boldsymbol{\lambda}}^{(\boldsymbol{\beta})}$ that is similar to the condition of Theorem 1 of Hamura and Kubokawa (2019b) and incompatible with the condition for propriety. In contrast, applying the sharper inequality

$$\log(1+\xi) \le \xi - \frac{\xi^2}{2(1+\xi)}$$

for $\xi \ge 0$ leads to a result applicable to proper Bayes estimators. This sharper inequality is similar to the inequality of Lemma 3.1 of Dey, Ghosh, and Srinivasan (1987), which is used by Ghosh and Yang (1988).

Theorem 2.6.2 *Let* $\tilde{\rho} = 2\{\overline{C}(\alpha+\beta_0+1) - \sum_{i=1}^m C_i\beta_i\}/\{\underline{C}(\alpha+\beta_0)\}$. *Suppose that one of the following two conditions holds:*

(i)

$$\alpha + \beta_0 \le \sum_{i=1}^m C_i\beta_i/\overline{C} - 1.$$

(ii) $\alpha < \sum_{i=1}^m \beta_i$, $\gamma_0 \le \underline{a}$, $0 \le \tilde{\rho} \le 1$, *and*

$$\tilde{\rho}\Big(\beta_0 + \sum_{i=1}^m \beta_i + \frac{\underline{a}}{\overline{a}}\Big) \le 2\frac{\overline{C}}{\underline{C}}\beta_0 + \frac{\underline{a}}{\overline{a}}.$$

Then $E_{\boldsymbol{\lambda}}[\tilde{L}_{\boldsymbol{c}}(\widetilde{\boldsymbol{\lambda}}^{(\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0)}, \boldsymbol{\lambda})] < E_{\boldsymbol{\lambda}}[L_{\boldsymbol{c}}(\widetilde{\boldsymbol{\lambda}}^{(\boldsymbol{\beta})}, \boldsymbol{\lambda})]$ *for all* $\boldsymbol{\lambda} \in (0,\infty)^m$.

2.6.3 Prediction under the Kullback-Leibler divergence

This subsection extends the result of the point estimation in Theorem 2.6.2 to a corresponding prediction problem. Suppose that $Y_1, \ldots, Y_m$ and $Z_1, \ldots, Z_m$ are independent Poisson random variables with means $r_1\lambda_1, \ldots, r_m\lambda_m$ and $s_1\lambda_1, \ldots, s_m\lambda_m$, respectively, and suppose that $\lambda_1, \ldots, \lambda_m > 0$ are unknown while $r_1, \ldots, r_m > 0$ and $s_1, \ldots, s_m > 0$ are known. Let $p_{\boldsymbol{Y}}(\cdot|\boldsymbol{\lambda})$ and

$p_{\boldsymbol{Z}}(\cdot|\boldsymbol{\lambda})$ be the densities of $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_m)$, respectively. We consider the problem of predicting the density $p_{\boldsymbol{Z}}(\cdot|\boldsymbol{\lambda})$ of $\boldsymbol{Z}$, where each predictor $\hat{p}(\cdot; \boldsymbol{Y})$ based on $\boldsymbol{Y}$ is evaluated in terms of the risk function relative to the Kullback-Leibler divergence, given by

$$R(\boldsymbol{\lambda}, \hat{p}) = E_{\boldsymbol{\lambda}}\Big[\sum_{\boldsymbol{z} \in \mathbb{N}_0{}^m} p_{\boldsymbol{Z}}(\boldsymbol{z}|\boldsymbol{\lambda}) \log \frac{p_{\boldsymbol{Z}}(\boldsymbol{z}|\boldsymbol{\lambda})}{\hat{p}(\boldsymbol{z}; \boldsymbol{Y})} \Big].$$

The Bayesian predictive density, denoted by $\hat{p}_\pi$, against prior π is

$$\hat{p}_\pi(\boldsymbol{z}; \boldsymbol{Y}) = \int p_{\boldsymbol{Z}}(\boldsymbol{z}|\boldsymbol{\xi}) p_{\boldsymbol{Y}}(\boldsymbol{Y}|\boldsymbol{\xi}) \pi(\boldsymbol{\xi}) d\boldsymbol{\xi} \Big/ \int p_{\boldsymbol{Y}}(\boldsymbol{Y}|\boldsymbol{\xi}) \pi(\boldsymbol{\xi}) d\boldsymbol{\xi}.$$

Let $\hat{p}_{\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}}$ and $\hat{p}_{\pi_{\boldsymbol{\beta}}}$ be the Bayesian predictive densities against the priors $\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}$ and $\pi_{\boldsymbol{\beta}}$, respectively. We extend Theorem 2.6.2 to the prediction problem by using the result of Lemma 2.6.1 below, which is a special case of Lemma 1 of Komaki (2015).

Lemma 2.6.1 *Let $W_i(\tau)$ be a Poisson random variable with mean $t_i(\tau)\lambda_i$ for $\tau \in [0,1]$ and $i = 1, \ldots, m$, where*

$$t_i(\tau) = r_i \frac{1 + s_i/r_i}{1 + (s_i/r_i)(1-\tau)}.$$

Let $\widetilde{\boldsymbol{\lambda}}^{(\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0})}(\tau) = (\tilde{\lambda}_1^{(\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0})}(\tau), \ldots, \tilde{\lambda}_m^{(\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0})}(\tau))$ and $\widetilde{\boldsymbol{\lambda}}^{(\pi_{\boldsymbol{\beta}})}(\tau) = (\tilde{\lambda}_1^{(\pi_{\boldsymbol{\beta}})}(\tau), \ldots, \tilde{\lambda}_m^{(\pi_{\boldsymbol{\beta}})}(\tau))$ for

$$\tilde{\lambda}_i^{(\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0})}(\tau) = \frac{W_i(\tau) + \beta_i}{t_i(\tau)} \frac{K(\boldsymbol{t}(\tau) \circ \boldsymbol{\gamma}, \boldsymbol{W}(\tau) + \boldsymbol{\beta} + \boldsymbol{e}_i, \alpha + \beta_0; \gamma_0, \beta_0)}{K(\boldsymbol{t}(\tau) \circ \boldsymbol{\gamma}, \boldsymbol{W}(\tau) + \boldsymbol{\beta}, \alpha + \beta_0; \gamma_0, \beta_0)}$$

and $\tilde{\lambda}_i^{(\pi_{\boldsymbol{\beta}})}(\tau) = \{W_i(\tau) + \beta_i\}/t_i(\tau)$, where $\boldsymbol{t}(\tau) = (t_1(\tau), \ldots, t_m(\tau))$ and $\boldsymbol{W}(\tau) = (W_1(\tau), \ldots, W_m(\tau))$. Then the risk difference between $\hat{p}_{\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}}$ and $\hat{p}_{\pi_{\boldsymbol{\beta}}}$ is expressed as

$$R(\boldsymbol{\lambda}, \hat{p}_{\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}}) - R(\boldsymbol{\lambda}, \hat{p}_{\pi_{\boldsymbol{\beta}}}) = \int_0^1 \{E_{\boldsymbol{\lambda}}[\tilde{L}_{\boldsymbol{t}'(\tau)}(\widetilde{\boldsymbol{\lambda}}^{(\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0})}(\tau), \boldsymbol{\lambda})] - E_{\boldsymbol{\lambda}}[\tilde{L}_{\boldsymbol{t}'(\tau)}(\widetilde{\boldsymbol{\lambda}}^{(\pi_{\boldsymbol{\beta}})}(\tau), \boldsymbol{\lambda})]\} d\tau,$$

where $\boldsymbol{t}'(\tau) = (t_1{}'(\tau), \ldots, t_m{}'(\tau)) = ((dt_1/d\tau)(\tau), \ldots, (dt_m/d\tau)(\tau))$.

Combining Theorem 2.6.2 and Lemma 2.6.1 and noting that $\{t_i{}'(\tau)/t_i(\tau)\}/\{t_i(\tau)\gamma_i\} = \{1/r_i - 1/(r_i + s_i)\}/\gamma_i$ for all $i = 1, \ldots, m$ for all $\tau \in [0,1]$, we have the following result, which gives a sufficient condition under which $\hat{p}_{\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}}$ dominates $\hat{p}_{\pi_{\boldsymbol{\beta}}}$.

Theorem 2.6.3 *Let $A_i = \{1/r_i - 1/(r_i + s_i)\}/\gamma_i$ for $i = 1, \ldots, m$ and let $\underline{A} = \min_{1 \le i \le m} A_i$ and $\overline{A} = \max_{1 \le i \le m} A_i$. Let $\sigma = 2\{\overline{A}(\alpha + \beta_0 + 1) - \sum_{i=1}^m A_i\beta_i\}/\{\underline{A}(\alpha + \beta_0)\}$. Suppose that either*

- $$\alpha + \beta_0 \le \sum_{i=1}^m \frac{A_i\beta_i}{\overline{A}} - 1$$

or

- *$\alpha < \beta.$, $\gamma_0 \le \min_{1 \le i \le m} t_i(\tau)\gamma_i$, $0 \le \sigma \le 1$, and*

$$\sigma\Big\{\beta_0 + \sum_{i=1}^m \beta_i + \frac{\min_{1 \le i \le m} t_i(\tau)\gamma_i}{\max_{1 \le i \le m} t_i(\tau)\gamma_i}\Big\} \le 2\frac{\overline{A}}{\underline{A}}\beta_0 + \frac{\min_{1 \le i \le m} t_i(\tau)\gamma_i}{\max_{1 \le i \le m} t_i(\tau)\gamma_i}$$

  for all $\tau \in [0,1]$.

Then we have $R(\boldsymbol{\lambda}, \hat{p}_{\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}}) < R(\boldsymbol{\lambda}, \hat{p}_{\pi_{\boldsymbol{\beta}}})$ for all $\boldsymbol{\lambda} \in (0,\infty)^m$.

2.7 Appendix

All the proofs of the lemmas in Sections 2.2 and 2.3 are given here. For $\boldsymbol{v} = (v_1, \ldots, v_m) \in \mathbb{R}^m$ and $\tilde{\boldsymbol{v}} = (\tilde{v}_1, \ldots, \tilde{v}_m) \in \mathbb{R}^m$, we write the inner product $v_1\tilde{v}_1 + \cdots + v_m\tilde{v}_m$ as $\boldsymbol{v} \cdot \tilde{\boldsymbol{v}}$.

Proof of Lemma 2.2.1. Let $J = \int \cdots \int_{(0,\infty)^m} \pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}(\boldsymbol{\lambda}) d\boldsymbol{\lambda}$. From (2.2.2), it follows that

$$J = \int_0^\infty \Big\{ \frac{u^{\alpha-1+\beta_0}}{(1+u/\gamma_0)^{\beta_0}} \int \cdots \int_{(0,\infty)^m} \Big(\prod_{i=1}^m {\lambda_i}^{\beta_i - 1} \Big) e^{-u\sum_{i=1}^m \lambda_i/\gamma_i} d\boldsymbol{\lambda} \Big\} du.$$

By making the change of variables

$$(\theta_1, \ldots, \theta_{m-1}, \Lambda) = \Big(\lambda_1 \Big(\sum_{i=1}^m \lambda_i \Big)^{-1}, \ldots, \lambda_{m-1} \Big(\sum_{i=1}^m \lambda_j \Big)^{-1}, \sum_{i=1}^m \lambda_i \Big),$$

we obtain

$$\begin{aligned} J &= \int_0^\infty \Big\{ \frac{u^{\alpha-1+\beta_0}}{(1+u/\gamma_0)^{\beta_0}} \\ &\quad \times \int \cdots \int_{D\times(0,\infty)} \Big(\Lambda^{\beta_\cdot - m} \prod_{i=1}^m {\theta_i}^{\beta_i - 1} \Big) e^{-\Lambda u \sum_{i=1}^m \theta_i/\gamma_i} \Lambda^{m-1} d\theta_1 \cdots d\theta_{m-1} d\Lambda \Big\} du \\ &= \int_0^\infty \Big\{ \frac{u^{\alpha-1+\beta_0}}{(1+u/\gamma_0)^{\beta_0}} \\ &\quad \times \int \cdots \int_D \Big(\prod_{i=1}^m {\theta_i}^{\beta_i - 1} \Big) \Gamma(\beta_\cdot) \Big(u \sum_{i=1}^m \theta_i/\gamma_i \Big)^{-\beta_\cdot} d\theta_1 \cdots d\theta_{m-1} \Big\} du, \end{aligned} \tag{2.7.1}$$

where θ_m denotes $1-(\theta_1+\cdots+\theta_{m-1})$ and $D = \{(\zeta_1, \ldots, \zeta_{m-1}) \in (0,1)^{m-1} : \zeta_1+\cdots+\zeta_{m-1} < 1\}$. Let $\overline{\gamma} = \max_{1\le i\le m} \gamma_i$ and $\underline{\gamma} = \min_{1\le i\le m} \gamma_i$. Then, from (2.7.1),

$$\begin{aligned} J &\le \int_0^\infty \Big\{ \frac{u^{\alpha-1+\beta_0}}{(1+u/\gamma_0)^{\beta_0}} \\ &\quad \times \int \cdots \int_D \Big(\prod_{i=1}^m {\theta_i}^{\beta_j - 1} \Big) \Gamma(\beta_\cdot) u^{-\beta_\cdot} \Big(\overline{\gamma} / \sum_{i=1}^m \theta_i \Big)^{\beta_\cdot} d\theta_1 \cdots d\theta_{m-1} \Big\} du \\ &= \Gamma(\beta_\cdot) \overline{\gamma}^{\beta_\cdot} \Big\{ \int \cdots \int_D \Big(\prod_{i=1}^m {\theta_i}^{\beta_i - 1} \Big) d\theta_1 \cdots d\theta_{m-1} \Big\} \int_0^\infty \frac{u^{\alpha-1+\beta_0-\beta_\cdot}}{(1+u/\gamma_0)^{\beta_0}} du, \end{aligned}$$

the right-hand side of which is finite if $\alpha < \beta_\cdot < \alpha + \beta_0$. Similarly,

$$J \ge \Gamma(\beta_\cdot) \underline{\gamma}^{\beta_\cdot} \Big\{ \int \cdots \int_D \Big(\prod_{i=1}^m {\theta_i}^{\beta_i - 1} \Big) d\theta_1 \cdots d\theta_{m-1} \Big\} \int_0^\infty \frac{u^{\alpha-1+\beta_0-\beta_\cdot}}{(1+u/\gamma_0)^{\beta_0}} du = \infty$$

if the condition $\alpha < \beta_\cdot < \alpha + \beta_0$ does not hold, and the proof is complete. □

Proof of Lemma 2.2.2. For part (i), we have by integration by parts that

$$\begin{aligned}
&K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0)\\
&=\Big[\frac{u^\alpha}{\alpha}\frac{1}{(1+u/\gamma_0)^{\beta_0}}\prod_{i=1}^m\frac{1}{(1+u/\gamma_i)^{\xi_i}}\Big]_0^\infty\\
&\quad-\int_0^\infty\frac{u^\alpha}{\alpha}\Big(\frac{-\beta_0/\gamma_0}{1+u/\gamma_0}+\sum_{i=1}^m\frac{-\xi_i/\gamma_i}{1+u/\gamma_i}\Big)\frac{1}{(1+u/\gamma_0)^{\beta_0}}\prod_{i=1}^m\frac{1}{(1+u/\gamma_i)^{\xi_i}}du\\
&=0+\frac{1}{\alpha}\Big\{\frac{\beta_0}{\gamma_0}K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+1;\gamma_0,\beta_0+1)+\sum_{i=1}^m\frac{\xi_i}{\gamma_i}K(\boldsymbol{\gamma},\boldsymbol{\xi}+\boldsymbol{e}_i,\alpha+1;\gamma_0,\beta_0)\Big\}.
\end{aligned}$$

Part (ii) follows since

$$\begin{aligned}
&K(\boldsymbol{\gamma},\boldsymbol{\xi}+\boldsymbol{e}_i,\alpha+1;\gamma_0,\beta_0)\\
&=\int_0^\infty\frac{u^{\alpha-1}}{(1+u/\gamma_0)^{\beta_0}}\frac{u}{1+u/\gamma_i}\prod_{j=1}^m\frac{1}{(1+u/\gamma_j)^{\xi_j}}du\\
&=\int_0^\infty\frac{u^{\alpha-1}}{(1+u/\gamma_0)^{\beta_0}}\gamma_i\Big(1-\frac{1}{1+u/\gamma_i}\Big)\prod_{j=1}^m\frac{1}{(1+u/\gamma_j)^{\xi_j}}du\\
&=\gamma_i\{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0)-K(\boldsymbol{\gamma},\boldsymbol{\xi}+\boldsymbol{e}_i,\alpha;\gamma_0,\beta_0)\}
\end{aligned}$$

for $i=1,\ldots,m$. This completes the proof. □

Proof of Lemma 2.2.3. For part (i), let $f(u)=u^{\alpha-1}(1+u/\gamma_0)^{-\beta_0}\prod_{i=1}^m(1+u/\gamma_i)^{-\xi_i}$ for $u>0$ and let $\Delta_K=K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+1;\gamma_0,\beta_0)K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0)-K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+1;\gamma_0,\beta_0+1)K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0-1)$. Note that

$$\begin{aligned}
\Delta_K&=\int_0^\infty uf(u)du\int_0^\infty f(u)du-\int_0^\infty\frac{u}{1+u/\gamma_0}f(u)du\int_0^\infty\Big(1+\frac{u}{\gamma_0}\Big)f(u)du\\
&=\int_0^\infty uf(u)du\int_0^\infty f(u)du\\
&\quad-\gamma_0\int_0^\infty\Big(1-\frac{1}{1+u/\gamma_0}\Big)f(u)du\int_0^\infty\Big(1+\frac{u}{\gamma_0}\Big)f(u)du\\
&=-\gamma_0\Big\{\int_0^\infty f(u)du\Big\}^2+\gamma_0\int_0^\infty\frac{1}{1+u/\gamma_0}f(u)du\int_0^\infty\Big(1+\frac{u}{\gamma_0}\Big)f(u)du.
\end{aligned}$$

Then it follows from the Cauchy-Schwarz inequality that $\Delta_K\geq 0$, which can be rewritten as (2.2.8). The inequality (2.2.7) can be similarly shown. Next we prove part (ii). From (2.2.8), we have

$$0\leq\frac{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+1;\gamma_0,\beta_0)}{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0-1)}-\frac{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+1;\gamma_0,\beta_0+1)}{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0)}.$$

By adding and subtracting $K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+1;\gamma_0,\beta_0)/K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0)$, we obtain

$$
\begin{aligned}
0 &\le -\frac{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+1;\gamma_0,\beta_0)}{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0)}\Big(1-\frac{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0)}{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0-1)}\Big)\\
&\quad+\frac{1}{\gamma_0}\frac{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+2;\gamma_0,\beta_0+1)}{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0)}\\
&= -\frac{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+1;\gamma_0,\beta_0)}{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0)}\frac{1}{\gamma_0}\frac{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+1;\gamma_0,\beta_0)}{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0-1)}\\
&\quad+\frac{1}{\gamma_0}\frac{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+2;\gamma_0,\beta_0+1)}{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0)}\\
&\le -\frac{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+1;\gamma_0,\beta_0)}{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0)}\frac{1}{\gamma_0}\frac{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+1;\gamma_0,\beta_0+1)}{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0)}\\
&\quad+\frac{1}{\gamma_0}\frac{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha+2;\gamma_0,\beta_0+1)}{K(\boldsymbol{\gamma},\boldsymbol{\xi},\alpha;\gamma_0,\beta_0)},
\end{aligned}
$$

where the second inequality follows from (2.2.8), and thus (2.2.10) follows. The inequality (2.2.9) can be similarly shown. The proof of Lemma 2.2.3 is complete. □

Proof of Lemma 2.2.4. Let $\boldsymbol{x}=(x_1,\dots,x_m)\in{\mathbb{N}_0}^m$. For $i=1,\dots,m$, the posterior mean of $1/\lambda_i$ with respect to the observation $\boldsymbol{X}=\boldsymbol{x}$ and the prior $\pi_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma};\beta_0,\gamma_0}$, denoted $E^{\boldsymbol{\lambda}|\boldsymbol{X}}[1/\lambda_i|\boldsymbol{X}=\boldsymbol{x}]$, is given by

$$
\frac{\int_0^\infty\big[\frac{u^{\alpha-1+\beta_0}}{(1+u/\gamma_0)^{\beta_0}}\int\cdots\int_{(0,\infty)^m}\big\{\frac{1}{\lambda_i}(\prod_{j=1}^m{\lambda_j}^{x_j+\beta_j-1}e^{-n_j\lambda_j})e^{-u\sum_{j=1}^m\lambda_j/\gamma_j}\big\}d\boldsymbol{\lambda}\big]du}{\int_0^\infty\big[\frac{u^{\alpha-1+\beta_0}}{(1+u/\gamma_0)^{\beta_0}}\int\cdots\int_{(0,\infty)^m}\big\{(\prod_{j=1}^m{\lambda_j}^{x_j+\beta_j-1}e^{-n_j\lambda_j})e^{-u\sum_{j=1}^m\lambda_j/\gamma_j}\big\}d\boldsymbol{\lambda}\big]du},
$$

which can be rewritten as

$$
\begin{aligned}
&\frac{\int_0^\infty\frac{u^{\alpha-1+\beta_0}}{(1+u/\gamma_0)^{\beta_0}}\prod_{j=1}^m\int_0^\infty{\lambda_j}^{x_j+\beta_j-1-\delta_{ij}}e^{-\lambda_j(n_j+u/\gamma_j)}d\lambda_j du}{\int_0^\infty\frac{u^{\alpha-1+\beta_0}}{(1+u/\gamma_0)^{\beta_0}}\prod_{j=1}^m\int_0^\infty{\lambda_j}^{x_j+\beta_j-1}e^{-\lambda_j(n_j+u/\gamma_j)}d\lambda_j du}\\
&=\frac{\int_0^\infty\frac{u^{\alpha-1+\beta_0}}{(1+u/\gamma_0)^{\beta_0}}\prod_{j=1}^m\frac{\Gamma(x_j+\beta_j-\delta_{ij})}{(n_j+u/\gamma_j)^{x_j+\beta_j-\delta_{ij}}}du}{\int_0^\infty\frac{u^{\alpha-1+\beta_0}}{(1+u/\gamma_0)^{\beta_0}}\prod_{j=1}^m\frac{\Gamma(x_j+\beta_j)}{(n_j+u/\gamma_j)^{x_j+\beta_j}}du}\\
&=n_i\frac{\Gamma(x_i+\beta_i-1)}{\Gamma(x_i+\beta_i)}\frac{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{\beta}-\boldsymbol{e}_i,\alpha+\beta_0;\gamma_0,\beta_0)}{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{\beta},\alpha+\beta_0;\gamma_0,\beta_0)},
\end{aligned}
$$

where $\delta_{ij}=\boldsymbol{e}_i\cdot\boldsymbol{e}_j$ for $j=1,\dots,m$, $\Gamma(t)=\infty$ for $t\le 0$, and $K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{0}+\boldsymbol{\beta}-\boldsymbol{e}_i,\alpha+\beta_0;\gamma_0,\beta_0)=\infty$ for $\alpha\ge\beta_\cdot-1$. Similarly, we have

$$
E^{\boldsymbol{\lambda}|\boldsymbol{X}}[\lambda_i|\boldsymbol{X}=\boldsymbol{x}]=\frac{1}{n_i}\frac{\Gamma(x_i+\beta_i+1)}{\Gamma(x_i+\beta_i)}\frac{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{\beta}+\boldsymbol{e}_i,\alpha+\beta_0;\gamma_0,\beta_0)}{K(\boldsymbol{n}\circ\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{\beta},\alpha+\beta_0;\gamma_0,\beta_0)},
$$

which is finite. Hence, for all $\boldsymbol{d} = (d_1, \ldots, d_m) \in \mathbb{R}^m$, we have

$$\begin{aligned}
&E^{\boldsymbol{\lambda}|\boldsymbol{X}}[L_{\boldsymbol{c}}(\boldsymbol{d}, \boldsymbol{\lambda})|\boldsymbol{X} = \boldsymbol{x}] \\
&= \sum_{i=1}^{m} c_i E^{\boldsymbol{\lambda}|\boldsymbol{X}}\Big[\frac{{d_i}^2}{\lambda_i} - 2d_i + \lambda_i \Big| \boldsymbol{X} = \boldsymbol{x}\Big] \\
&= \sum_{i \in S} c_i \Big\{ E^{\boldsymbol{\lambda}|\boldsymbol{X}}\Big[\frac{1}{\lambda_i}\Big|\boldsymbol{X} = \boldsymbol{x}\Big]\Big(d_i - \frac{1}{E^{\boldsymbol{\lambda}|\boldsymbol{X}}\big[\frac{1}{\lambda_i}\big|\boldsymbol{X} = \boldsymbol{x}\big]}\Big)^2 + A_i \Big\} \\
&\quad + \sum_{i \in S^c} c_i({d_i}^2 \cdot \infty - 2d_i + E^{\boldsymbol{\lambda}|\boldsymbol{X}}[\lambda_i|\boldsymbol{X} = \boldsymbol{x}]),
\end{aligned}$$

where $S = \{i \in \{1, \ldots, m\} : E^{\boldsymbol{\lambda}|\boldsymbol{X}}[1/\lambda_i|\boldsymbol{X} = \boldsymbol{x}] < \infty\}$ and $A_i = -(E^{\boldsymbol{\lambda}|\boldsymbol{X}}[1/\lambda_i|\boldsymbol{X} = \boldsymbol{x}])^{-1} + E^{\boldsymbol{\lambda}|\boldsymbol{X}}[\lambda_i|\boldsymbol{X} = \boldsymbol{x}]$ for $i \in S$. Therefore, $E^{\boldsymbol{\lambda}|\boldsymbol{X}}[L_{\boldsymbol{c}}(\boldsymbol{d}, \boldsymbol{\lambda})|\boldsymbol{X} = \boldsymbol{x}]$ is finite if and only if $d_i = 0$ for all $i \in S^c$. Furthermore, in this case, it is minimized if and only if $d_i = (E^{\boldsymbol{\lambda}|\boldsymbol{X}}[1/\lambda_i|\boldsymbol{X} = \boldsymbol{x}])^{-1}$ for all $i \in S$. Thus, $E^{\boldsymbol{\lambda}|\boldsymbol{X}}[L_{\boldsymbol{c}}(\boldsymbol{d}, \boldsymbol{\lambda})|\boldsymbol{X} = \boldsymbol{x}]$ is uniquely minimized at

$$\boldsymbol{d} = \sum_{i \in S} \frac{x_i + \beta_i - 1}{n_i} \frac{K(\boldsymbol{n} \circ \boldsymbol{\gamma}, \boldsymbol{x} + \boldsymbol{\beta}, \alpha + \beta_0; \gamma_0, \beta_0)}{K(\boldsymbol{n} \circ \boldsymbol{\gamma}, \boldsymbol{x} + \boldsymbol{\beta} - \boldsymbol{e}_i, \alpha + \beta_0; \gamma_0, \beta_0)} \boldsymbol{e}_i,$$

which can be expressed as

$$\Big(\{1 - \phi_1^{(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}; \beta_0, \gamma_0)}(\boldsymbol{x})\}\frac{x_1 + \beta_1 - 1}{n_1}, \ldots, \{1 - \phi_m^{(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}; \beta_0, \gamma_0)}(\boldsymbol{x})\}\frac{x_m + \beta_m - 1}{n_m}\Big)$$

by part (ii) of Lemma 2.2.2. Thus, the desired result is obtained. □

Proof of Lemma 2.2.5. For part (i), suppose $x_i \geq 1$. Then

$$\begin{aligned}
&n_i K(\boldsymbol{n}, \boldsymbol{x} + \boldsymbol{j} - \boldsymbol{e}_i, \alpha + \beta_0; \gamma_0, \beta_0) \\
&= \int_0^\infty \frac{u^{\alpha + \beta_0 - 1}}{(1 + u/\gamma_0)^{\beta_0}} (n_i + u) \prod_{j=1}^{m} \frac{1}{(1 + u/n_j)^{x_j + 1}} du \qquad (2.7.2) \\
&> \int_0^\infty \frac{u^{\alpha + \beta_0}}{(1 + u/\gamma_0)^{\beta_0}} \prod_{j=1}^{m} \frac{1}{(1 + u/n_j)^{x_j + 1}} du \\
&= K(\boldsymbol{n}, \boldsymbol{x} + \boldsymbol{j}, \alpha + \beta_0 + 1; \gamma_0, \beta_0).
\end{aligned}$$

This shows the desired result. Part (ii) follows immediately from (2.7.2). For part (iii), let $f_i(u) = (1 + u/\gamma_0)^{-\beta_0} \prod_{j \neq i} (1 + u/n_j)^{-(x_j + 1)}$ for $u > 0$ and let $k = 0, 1$. Then we have that

$$0 \leq u^{\alpha + \beta_0 - k} \frac{1}{(1 + u/n_i)^{x_i + 1 - k}} f_i(u) \uparrow u^{\alpha + \beta_0 - k} f_i(u)$$

as $n_i \to \infty$ for every $u > 0$. Since $\alpha < m - 2$, it follows from the dominated convergence theorem that

$$\begin{aligned}
&K(\boldsymbol{n}, \boldsymbol{x} + \boldsymbol{j} - k\boldsymbol{e}_i, \alpha + \beta_0 + 1 - k; \gamma_0, \beta_0) \\
&= \int_0^\infty u^{\alpha + \beta_0 - k} \frac{1}{(1 + u/n_i)^{x_i + 1 - k}} f_i(u) du \\
&\to \int_0^\infty u^{\alpha + \beta_0 - k} f_i(u) du \in (0, \infty)
\end{aligned}$$

as $n_i \to \infty$, and this completes the proof. □

Proof of Lemma 2.3.1. It can be seen that

$$\begin{aligned}
&E_{\boldsymbol{\lambda}}\Big[\frac{h(\boldsymbol{X}+\boldsymbol{e}_i)}{X_i+1}\Big] \\
&= \sum_{\boldsymbol{x}\in{\mathbb{N}_0}^m} \frac{h(\boldsymbol{x}+\boldsymbol{e}_i)}{x_i+1} \prod_{j=1}^{m} \frac{(n_j\lambda_j)^{x_j}}{x_j!} e^{-n_j\lambda_j} \\
&= \sum_{\boldsymbol{x}\in{\mathbb{N}_0}^m} \frac{h(\boldsymbol{x})}{n_i\lambda_i} \prod_{j=1}^{m} \frac{(n_j\lambda_j)^{x_j}}{x_j!} e^{-n_j\lambda_j} - \sum_{\boldsymbol{x}\in{\mathbb{N}_0}^m,\, \boldsymbol{x}\cdot\boldsymbol{e}_i=0} \frac{h(\boldsymbol{x})}{n_i\lambda_i} \prod_{j=1}^{m} \frac{(n_j\lambda_j)^{x_j}}{x_j!} e^{-n_j\lambda_j} \\
&= E_{\boldsymbol{\lambda}}\Big[\frac{h(\boldsymbol{X})}{n_i\lambda_i}\Big],
\end{aligned}$$

which proves Lemma 2.3.1. □

Chapter 3

Bayesian Shrinkage Estimation of Negative Multinomial Parameter Vectors

3.1 Introduction

Stein's phenomenon for the estimation of parameters of discrete distributions has been extensively studied since Clevenson and Zidek (1975) showed that the usual estimator of the mean vector of independent Poisson distributions is dominated by a Bayesian shrinkage estimator under the standardized squared error loss. For example, Ghosh and Parsian (1981), Tsui (1979b), Tsui and Press (1982), and Ghosh and Yang (1988) considered different estimators of Poisson parameters under different loss functions. Estimation for discrete exponential families including the Poisson and the negative binomial distributions was treated by Tsui (1979a), Hwang (1982), and Ghosh, Hwang, and Tsui (1983). Tsui (1984), Tsui (1986a), and Tsui (1986b) explored the robustness of Clevenson–Zidek-type estimators in estimating means when the observations are not Poisson-distributed. In particular, Tsui (1986b) considered the case of dependent observations following the negative multinomial distribution, which is a multivariate generalization of the negative binomial distribution and arises as the joint distribution of the frequencies of multiple events in inverse sampling. The negative multinomial distribution is also included in the general classes of discrete distributions of Chou (1991) and Dey and Chung (1992).

However, little attention has been paid to the construction of Bayesian shrinkage estimators when the underlying distributions are not Poisson. This could be partly because tractable hierarchical models may not be so widely known in such cases; some difficulties with the beta-binomial hierarchy are discussed in Example 4.5.3 of Lehmann and Casella (1998). In this chapter, we consider the Bayesian estimation of multiple negative multinomial parameter vectors.

The m-dimensional negative multinomial distribution with parameters $r > 0$ and $\mathring{\boldsymbol{p}} = (\mathring{p}_1, \dots, \mathring{p}_m)^\top \in D_m = \{(\tilde{p}_1, \dots, \tilde{p}_m)^\top | \tilde{p}_1, \dots, \tilde{p}_m > 0, \sum_{i=1}^m \tilde{p}_i < 1\}$, denoted by $\mathrm{NM}_m(r, \mathring{\boldsymbol{p}})$, has probability mass function

$$\mathrm{NM}_m(\boldsymbol{x}|r, \mathring{\boldsymbol{p}}) = \frac{\Gamma\left(r + \sum_{i=1}^m x_i\right)}{\Gamma(r) \prod_{i=1}^m x_i!} \mathring{p}_0^{\,r} \prod_{i=1}^m \mathring{p}_i^{x_i} \tag{3.1.1}$$

for $\boldsymbol{x} = (x_1, \dots, x_m)^\top \in {\mathbb{N}_0}^m = \{0, 1, 2, \dots\}^m$, where $\mathring{p}_0 = 1 - \mathring{p}_\cdot = 1 - \sum_{i=1}^m \mathring{p}_i$ and where r

corresponds to the number of successes in inverse sampling. Even if r is not an integer, the probability function (3.1.1) is well defined and has the Poisson-gamma mixture representation

$$\mathrm{NM}_m(\boldsymbol{x}|r,\mathring{\boldsymbol{p}}) = \int_0^\infty \frac{v^{r-1}}{\Gamma(r)} e^{-v} \Big[\prod_{i=1}^m \frac{\{(\mathring{p}_i/\mathring{p}_0)v\}^{x_i}}{x_i!} e^{-(\mathring{p}_i/\mathring{p}_0)v} \Big] dv. \tag{3.1.2}$$

The mean and variance of the negative multinomial distribution $\mathrm{NM}_m(r,\mathring{\boldsymbol{p}})$ are $r\mathring{\boldsymbol{p}}/\mathring{p}_0$ and $r\mathrm{diag}\,(\mathring{\boldsymbol{p}})/\mathring{p}_0 + r\mathring{\boldsymbol{p}}\mathring{\boldsymbol{p}}'/\mathring{p}_0^2$. The marginals are negative binomial. If $\mathring{\boldsymbol{X}}^{(1)} \sim \mathrm{NM}_m(r^{(1)},\mathring{\boldsymbol{p}})$ and $\mathring{\boldsymbol{X}}^{(2)} \sim \mathrm{NM}_m(r^{(2)},\mathring{\boldsymbol{p}})$ for $r^{(1)}, r^{(2)} > 0$, then $\mathring{\boldsymbol{X}}^{(1)} + \mathring{\boldsymbol{X}}^{(2)} \sim \mathrm{NM}_m(r^{(1)} + r^{(2)},\mathring{\boldsymbol{p}})$; therefore, r can also be interpreted as a sample size. For further properties and applications of the negative multinomial distribution, see, for example, Sibuya, Yoshimura, and Shimizu (1964) and Tsui (1986b) and the references therein.

Suppose that $\boldsymbol{X}_1 = (X_{1,1},\dots,X_{m,1})^\top, \dots, \boldsymbol{X}_N = (X_{1,N},\dots,X_{m,N})^\top$ are independently distributed according to $\mathrm{NM}_m(r,\boldsymbol{p}_1),\dots,\mathrm{NM}_m(r,\boldsymbol{p}_N)$, respectively, for $m, N \in \mathbb{N} = \{1,2,\dots\}$, where all the elements of $\boldsymbol{p} = (\boldsymbol{p}_1,\dots,\boldsymbol{p}_N) = ((p_{1,1},\dots,p_{m,1})^\top,\dots,(p_{1,N},\dots,p_{m,N})^\top) \in {D_m}^N$ are assumed to be unknown. For $n \in \{1,\dots,N\}$, we consider the problem of estimating the matrix $(\boldsymbol{p}_1,\dots,\boldsymbol{p}_n)$ on the basis of $\boldsymbol{X} = (\boldsymbol{X}_1,\dots,\boldsymbol{X}_N)$ under the standardized squared error loss

$$L_n(\boldsymbol{d},\boldsymbol{p}) = \sum_{\nu=1}^n \sum_{i=1}^m \frac{1}{p_{i,\nu}} (d_{i,\nu} - p_{i,\nu})^2, \tag{3.1.3}$$

where $\boldsymbol{d} = (d_{i,\nu})_{1\le i\le m,\,1\le\nu\le N} \in \mathbb{R}^{m\times N}$. Here, $n = N$ corresponds to the simultaneous estimation of all the parameters while $n = 1$ corresponds to the estimation of $\boldsymbol{p}_1$ relating to the first observation $\boldsymbol{X}_1$ by using all the information $\boldsymbol{X}$. The case of $n = N$ (with $m = 1$ or $N = 1$) has been considered in the literature. One motivation for our general framework is to borrow information from the entire population even when there are nuisance parameters.

As prior distribution for $\boldsymbol{p}$, we first use the conjugate Dirichlet distribution with density

$$\prod_{\nu=1}^N \mathrm{Dir}_m(\boldsymbol{p}_\nu|a_0,\boldsymbol{a}) = \prod_{\nu=1}^N \Big\{ \frac{\Gamma(a_0 + a_\cdot)}{\Gamma(a_0)\prod_{i=1}^m \Gamma(a_i)} {p_{0,\nu}}^{a_0-1} \prod_{i=1}^m {p_{i,\nu}}^{a_i-1} \Big\}, \tag{3.1.4}$$

where $a_0 \in \mathbb{R}$, $\boldsymbol{a} = (a_1,\dots,a_m)^\top \in (0,\infty)^m$, $a_\cdot = \sum_{i=1}^m a_i$, and $p_{0,\nu} = 1 - p_{\cdot,\nu} = 1 - \sum_{i=1}^m p_{i,\nu}$ for $\nu \in \{1,\dots,N\}$. As will be shown later, the UMVU estimator of $\boldsymbol{p}$ is $\hat{\boldsymbol{p}}^{\mathrm{U}} = (X_{i,\nu}/(r + X_{\cdot,\nu} - 1))_{1\le i\le m,\,1\le\nu\le N}$, where $X_{\cdot,\nu} = \sum_{i=1}^m X_{i,\nu}$ for $\nu \in \{1,\dots,N\}$, and corresponds to the Bayes estimator with respect to the prior (3.1.4) with $a_0 = -m$ and $\boldsymbol{a} = \boldsymbol{j}^{(m)}$ and the loss (3.1.3) with $n = N$, where $\boldsymbol{j}^{(m)} = (1,\dots,1)^\top \in \mathbb{R}^m$. Also, it will be seen that the Jeffreys prior is (3.1.4) with $a_0 = -(m-1)/2$ and $\boldsymbol{a} = \boldsymbol{j}^{(m)}/2$.

In Section 3.2, we first consider the general class of estimators

$$\hat{\boldsymbol{p}}^{(\delta)} = \Big(\frac{X_{i,\nu}}{r + X_{\cdot,\nu} - 1 + \delta(X_{\cdot,\cdot})} \Big)_{1\le i\le m,\,1\le\nu\le N}, \tag{3.1.5}$$

where $\delta(X_{\cdot,\cdot})$ is a strictly positive function of $X_{\cdot,\cdot} = \sum_{\nu=1}^N X_{\cdot,\nu} = \sum_{\nu=1}^N \sum_{i=1}^m X_{i,\nu}$, and derive a sufficient condition for the shrinkage estimator $\hat{\boldsymbol{p}}^{(\delta)}$ to dominate the UMVU estimator $\hat{\boldsymbol{p}}^{\mathrm{U}}$. Next we construct an empirical Bayes estimator based on the prior (3.1.4) with $\boldsymbol{a} = \boldsymbol{j}^{(m)}$ and

show that it dominates the UMVU estimator when m is sufficiently large by using the derived condition.

In Section 3.3, we obtain a shrinkage estimator of the form $(X_{i,\nu}/\{r+X_{\cdot,\nu}-1+\delta(\boldsymbol{X}_\cdot)\})_{1\le i\le m,\,1\le\nu\le N}$, where $\delta(\boldsymbol{X}_\cdot)>0$ is some symmetric function of $\boldsymbol{X}_\cdot=(X_{\cdot,1},\dots,X_{\cdot,N})^\top$, by introducing a hierarchical prior for $\boldsymbol{p}$. In a simple case, this prior becomes

$$\boldsymbol{p}\sim\Big(\prod_{\nu=1}^N p_{0,\nu}\Big)^{-m-1}/\Big(\sum_{\nu=1}^N\log\frac{1}{p_{0,\nu}}\Big)^{\alpha},$$

where $\alpha>0$. The above expression shows that the prior puts more probability around $p_{0,1}=\cdots=p_{0,N}=1$ than the Dirichlet prior $\boldsymbol{p}\sim\prod_{\nu=1}^N {p_{0,\nu}}^{-m-1}$. Our hierarchical Bayes estimator is shown to dominate the UMVU estimator under some conditions. Also, for sufficiently large m, we obtain an estimator based on our hierarchical prior which dominates a Bayes estimator against the Jeffreys prior under the loss

$$\tilde L_n(\tilde{\boldsymbol{d}},\boldsymbol{p})=\sum_{\nu=1}^n\sum_{i=1}^m\Big(\tilde d_{i,\nu}-p_{i,\nu}-p_{i,\nu}\log\frac{\tilde d_{i,\nu}}{p_{i,\nu}}\Big),\tag{3.1.6}$$

where $\tilde{\boldsymbol{d}}=(\tilde d_{i,\nu})_{1\le i\le m,\,1\le\nu\le N}\in(0,\infty)^{m\times N}$. In addition, it turns out that posterior computation is quite simple under our hierarchical prior.

Recently, Stoltenberg and Hjort (2019) also considered Bayesian multivariate models for count variables based on the Poisson likelihood. Hamura and Kubokawa (2019b, 2020c) considered estimation of Poisson parameters when sample sizes are unbalanced by using and generalizing the shrinkage prior of Komaki (2015). Interestingly, it is the method for evaluating integrals in Bayesian predictive probabilities of Poisson variables in the presence of unbalanced sample sizes, developed by Komaki (2015) and utilized by Hamura and Kubokawa (2019b, 2020c), that plays a crucial role in obtaining the results in Section 3.3 for our hierarchical Bayes estimators of negative multinomial parameters in the balanced setting.

The remainder of this chapter is organized as follows. In Sections 3.2 and 3.3, we consider empirical Bayes and hierarchical Bayes estimators, respectively. In Section 3.4, through simulation, we compare our proposed estimators with the UMVU estimator as well as an alternative estimator which estimates $\boldsymbol{p}_1,\dots,\boldsymbol{p}_N$ independently based on $\boldsymbol{X}_1,\dots,\boldsymbol{X}_N$, respectively. Some concluding remarks are given in Section 3.5. Proofs are in the Appendix.

3.2 Empirical Bayes Estimation

We first derive a sufficient condition for the shrinkage estimator $\hat{\boldsymbol{p}}^{(\delta)}$ given in (3.1.5) to dominate the UMVU estimator. Let, for $i\in\{1,\dots,m\}$ and $\nu\in\{1,\dots,N\}$,

$$\hat p^{\mathrm{U}}_{i,\nu}=\frac{X_{i,\nu}}{r+X_{\cdot,\nu}-1}.\tag{3.2.1}$$

The right-hand side is defined to be 0 if the denominator is 0. The same remark applies to (3.2.2) below. Then $\hat{\boldsymbol{p}}^{\mathrm{U}}=(\hat p^{\mathrm{U}}_{i,\nu})_{1\le i\le m,\,1\le\nu\le N}$ is the UMVU estimator of $\boldsymbol{p}$ since it is unbiased by Lemma 3.6.1 in the Appendix and since $\boldsymbol{X}$ is a minimal and complete sufficient statistic. Let $\delta\colon\mathbb{N}_0\to(0,\infty)$ and let, for $i\in\{1,\dots,m\}$ and $\nu\in\{1,\dots,N\}$,

$$\hat{\boldsymbol{p}}^{(\delta)}=(\hat p^{(\delta)}_{i,\nu})_{1\le i\le m,\,1\le\nu\le N},\quad \hat p^{(\delta)}_{i,\nu}=\frac{X_{i,\nu}}{r+X_{\cdot,\nu}-1+\delta(X_{\cdot,\cdot})}.\tag{3.2.2}$$

The following theorem, together with the other theorems in this chapter, shows that borrowing information from the independent observations actually is useful in improving risk performance even if only a subset of the unknown parameters are of interest.

Theorem 3.2.1 *Let $n \in \{1, \ldots, N\}$ and assume $r \geq 5/2$. Suppose that the function δ satisfies the following conditions for all $z \in \mathbb{N}$:*

(i) $z\delta(z) \leq (z+1)\delta(z+1)$.

(ii) *If $z \geq 2$, then*

- $\delta(z) \leq 2(m-3)$ *implies* $(m-6)\delta(z) + 2(m-3)r \geq 0$ *and*
- $\delta(z) > 2(m-3)$ *implies* $n\{(m-6)\delta(z) + 2(m-3)r\} \geq (z-1)\{\delta(z) - 2(m-3)\}$.

Then the shrinkage estimator $\hat{\boldsymbol{p}}^{(\delta)}$ dominates the UMVU estimator $\hat{\boldsymbol{p}}^{\mathrm{U}}$ under the loss $L_n(\boldsymbol{d}, \boldsymbol{p})$ given by (3.1.3).

For example, if $\delta(X_{\cdot,\cdot}) = c_0$ for some constant $0 < c_0 \leq 2(m-3)$, conditions (i) and (ii) are satisfied provided that $m \geq 6(r + c_0)/(2r + c_0)$. Also, condition (i) is satisfied if $\delta(X_{\cdot,\cdot}) = c_1 + c_2/X_{\cdot,\cdot}$ for some constants $c_1, c_2 > 0$ when $X_{\cdot,\cdot} \geq 1$.

Next, we construct an empirical Bayes estimator. Lemma 3.2.1 below states that the shrinkage estimator $\hat{\boldsymbol{p}}^{(\delta)}$ coincides with a Bayes solution in a simple case. Let $\delta^{(a_0)}(X_{\cdot,\cdot}) = a_0 + m$.

Lemma 3.2.1 *Suppose $a_0 > \max\{-m, -r\}$. Then the shrinkage estimator $\hat{\boldsymbol{p}}^{(\delta^{(a_0)})}$ is a Bayes solution with respect to the prior (3.1.4) with $\boldsymbol{a} = \boldsymbol{j}^{(m)}$ under the loss (3.1.3) for every $n \in \{1, \ldots, N\}$.*

The conditions $a_0 > -m$ and $a_0 > -r$ ensure, respectively, that $\hat{\boldsymbol{p}}^{(\delta^{(a_0)})}$ shrinks toward the origin and that the posterior distribution is proper. If $r \geq 5/2$ and $0 < a_0 \leq m - 6$, the estimator $\hat{\boldsymbol{p}}^{(\delta^{(a_0)})} = (X_{i,\nu}/(r + a_0 + X_{\cdot,\nu} + m - 1))_{1 \leq i \leq m,\, 1 \leq \nu \leq N}$ is proper Bayes by Lemma 3.2.1 and dominates the UMVU estimator by Theorem 3.2.1.

An empirical Bayes estimator is obtained by first assuming $a_0 > 1$ and then substituting for a_0 in $\hat{\boldsymbol{p}}^{(\delta^{(a_0)})}$ an estimator based on the marginal likelihood of $\boldsymbol{p}$ under the prior corresponding to $\hat{\boldsymbol{p}}^{(\delta^{(a_0)})}$. More specifically, when $a_0 > 1$, the prior expectation of the mean $E[X_{\cdot,\cdot}] = \sum_{\nu=1}^{N} \sum_{i=1}^{m} r p_{i,\nu}/p_{0,\nu}$ with respect to the Dirichlet prior (3.1.4) with $\boldsymbol{a} = \boldsymbol{j}^{(m)}$ is given by

$$\begin{aligned}\int_{{D_m}^N} E[X_{\cdot,\cdot}] \Big\{ \prod_{\nu'=1}^{N} \mathrm{Dir}_m(\boldsymbol{p}_{\nu'} | a_0, \boldsymbol{j}^{(m)}) \Big\} d\boldsymbol{p} &= \sum_{\nu=1}^{N} \sum_{i=1}^{m} r \int_{{D_m}^N} \frac{p_{i,\nu}}{p_{0,\nu}} \Big\{ \prod_{\nu'=1}^{N} \mathrm{Dir}_m(\boldsymbol{p}_{\nu'} | a_0, \boldsymbol{j}^{(m)}) \Big\} d\boldsymbol{p} \\ &= \frac{Nmr}{a_0 - 1}.\end{aligned}$$

Thus, an estimator of a_0 is obtained as $\hat{a}_0 = 1 + Nmr/X_{\cdot,\cdot}$ and our empirical Bayes estimator is

$$\hat{\boldsymbol{p}}^{\mathrm{EB}} = (\hat{p}_{i,\nu}^{\mathrm{EB}})_{1 \leq i \leq m,\, 1 \leq \nu \leq N} = \hat{\boldsymbol{p}}^{(\delta^{(a_0)})}|_{a_0 = \hat{a}_0} = \Big(\frac{X_{i,\nu}}{r + X_{\cdot,\nu} - 1 + \delta^{\mathrm{EB}}(X_{\cdot,\cdot})} \Big)_{1 \leq i \leq m,\, 1 \leq \nu \leq N}, \quad (3.2.3)$$

where

$$\delta^{\mathrm{EB}}(X_{\cdot,\cdot}) = 1 + m + Nmr/X_{\cdot,\cdot}$$

when $X_{\cdot,\cdot} \geq 1$ and $\delta^{\mathrm{EB}}(0) \in (1 + m + Nmr, \infty)$.

The following corollary gives a sufficient condition for $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ to dominate the UMVU estimator.

Corollary 3.2.1 *Suppose that $m \geq 7$ and that $r \geq 5/2$. Then $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ is an empirical Bayes estimator dominating the UMVU estimator $\hat{\boldsymbol{p}}^{\mathrm{U}}$ under the loss $L_n(\boldsymbol{d}, \boldsymbol{p})$ given by (3.1.3) for every $n \in \{1, \ldots, N\}$.*

It is worth noting that the condition given in the above corollary is independent of n, which shows some robustness of the empirical Bayes estimator $\hat{\boldsymbol{p}}^{\mathrm{EB}}$. Additionally, recall that r, m, and N can vary independently in our setting. When applying Corollary 3.2.1, we do not have to set $N > m, r$, nor do we need to assume $r > m$.

The UMVU estimator corresponds to $a_0 = -m$ since $\lim_{a_0 \to -m} \hat{\boldsymbol{p}}^{(\delta^{(a_0)})} = \hat{\boldsymbol{p}}^{\mathrm{U}}$ (when $r > m$). However, the empirical Bayes estimator $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ was derived under the assumption that $a_0 > 1$. Indeed, we have $\hat{a}_0 > 1$ since all the elements of the observations $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ are nonnegative. Thus, there is a discrepancy in the support of a_0 between the usual Bayes estimator and the empirical Bayes estimator. On the other hand, in the case of hierarchical Bayes estimation, a mixture of the priors $\boldsymbol{p} \sim \prod_{\nu=1}^{N} \mathrm{Dir}_m(\boldsymbol{p}_\nu | s, \boldsymbol{j}^{(m)})$, $s > -m$, will be considered in the next section.

3.3 Hierarchical Bayes Estimation

In this section, we first introduce a shrinkage prior for $\boldsymbol{p}$ and investigate its properties (Section 3.3.1). Next, using the prior, we construct a hierarchical Bayes estimator that dominates the UMVU estimator under some conditions (Section 3.3.2). Finally, some remarks about posterior computation are presented (Section 3.3.3).

3.3.1 A hierarchical shrinkage prior

For $\boldsymbol{p} = ((p_{1,1}, \ldots, p_{m,1})^\top, \ldots, (p_{1,N}, \ldots, p_{m,N})^\top) \in {D_m}^N$ and $p_{0,\nu} = 1 - p_{\cdot,\nu} = 1 - \sum_{i=1}^{m} p_{i,\nu}$, $\nu \in \{1, \ldots, N\}$, let

$$\pi_{\alpha,\beta,g,a_0,\boldsymbol{a}}(\boldsymbol{p}) = \int_0^\infty t^{\alpha-1} e^{-\beta t} g(t) \Big\{ \prod_{\nu=1}^{N} \Big({p_{0,\nu}}^{t+a_0-1} \prod_{i=1}^{m} {p_{i,\nu}}^{a_i-1} \Big) \Big\} dt, \tag{3.3.1}$$

where $\alpha > 0$, $\beta \geq 0$, $g\colon (0,\infty) \to (0,\infty)$ is a bounded and smooth function, $a_0 \in \mathbb{R}$, and $\boldsymbol{a} = (a_1, \ldots, a_m)^\top \in (0,\infty)^m$. When $g = g_1$, where $g_1\colon (0,\infty) \to (0,\infty)$ is the function defined by $g_1(t) = 1$, $t \in (0,\infty)$, the prior (3.3.1) becomes

$$\pi_{\alpha,\beta,g_1,a_0,\boldsymbol{a}}(\boldsymbol{p}) = \Gamma(\alpha) \Big\{ \prod_{\nu=1}^{N} \Big({p_{0,\nu}}^{a_0-1} \prod_{i=1}^{m} {p_{i,\nu}}^{a_i-1} \Big) \Big\} \Big/ \Big(\beta + \sum_{\nu=1}^{N} \log \frac{1}{p_{0,\nu}} \Big)^{\alpha}. \tag{3.3.2}$$

It can be seen that

$$\lim_{\alpha \to 0} \frac{\pi_{\alpha,\beta,g_1,a_0,\boldsymbol{a}}(\boldsymbol{p})}{\Gamma(\alpha)} = \prod_{\nu=1}^{N} \Big({p_{0,\nu}}^{a_0-1} \prod_{i=1}^{m} {p_{i,\nu}}^{a_i-1} \Big) \propto \prod_{\nu=1}^{N} \mathrm{Dir}_m(\boldsymbol{p}_\nu | a_0, \boldsymbol{a})$$

and that the denominator of (3.3.2) tends to infinity as $\min\{p_{0,1},\ldots,p_{0,N}\}\to 0$. Thus, $\pi_{\alpha,\beta,g,a_0,\boldsymbol{a}}(\boldsymbol{p})$ is a shrinkage prior based on the Dirichlet distribution. Furthermore, if $m=1$, $N\geq 2$, and $(\Lambda,\boldsymbol{\theta})\sim e^{-(a_0-1)\Lambda}$, where $\Lambda=\sum_{\nu'=1}^{N}\log(1/p_{0,\nu'})$ and $\theta_\nu=\{\log(1/p_{0,\nu})\}/\sum_{\nu'=1}^{N}\log(1/p_{0,\nu'})$, $\nu\in\{1,\ldots,N-1\}$, then $\boldsymbol{p}\sim\big(\prod_{\nu=1}^{N}{p_{0,\nu}}^{a_0-1}\big)/\big\{\sum_{\nu=1}^{N}\log(1/p_{0,\nu})\big\}^{N-1}\propto\pi_{N-1,0,g_1,a_0,1}(\boldsymbol{p})$.

Let $a_\cdot=\sum_{i=1}^{m}a_i$. Necessary and sufficient conditions for propriety of the prior and posterior distributions are as follows:

Lemma 3.3.1

(i) *The prior (3.3.1) is proper if and only if either*

- $a_0>0$ *and* $\int_1^\infty t^{\alpha-Na_\cdot-1}e^{-\beta t}g(t)dt<\infty$ *or*
- $a_0=0$, $\int_0^1 t^{\alpha-N-1}e^{-\beta t}g(t)dt<\infty$, *and* $\int_1^\infty t^{\alpha-Na_\cdot-1}e^{-\beta t}g(t)dt<\infty$.

(ii) *Under the prior (3.3.1), the posterior distribution of* $\boldsymbol{p}$ *given the observations* $\boldsymbol{X}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)$ *is proper for all* $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\in{\mathbb{N}_0}^m$ *if and only if either*

- $r+a_0>0$, $\int_1^\infty t^{\alpha-Na_\cdot-1}e^{-\beta t}g(t)dt<\infty$ *or*
- $r+a_0=0$, $\int_0^1 t^{\alpha-N-1}e^{-\beta t}g(t)dt<\infty$, *and* $\int_1^\infty t^{\alpha-Na_\cdot-1}e^{-\beta t}g(t)dt<\infty$.

When the condition of part (ii) of Lemma 3.3.1 is satisfied, we will simply say that the posterior is proper. For example, when $g=g_1$ and either $\alpha<Na_\cdot$ or $\beta>0$, the prior (3.3.1) is proper if $a_0>0$, while the posterior is proper if $r+a_0>0$. It is also worth noting that even when $a_0<0$ and the prior is improper, the condition for posterior propriety may still be satisfied.

The prior (3.3.1) is related to shrinkage priors in the Poisson case. Specifically, if $m=1$ and $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_N)=(\log(1/p_{0,1}),\ldots,\log(1/p_{0,N}))\approx{\boldsymbol{0}^{(N)}}^\top$, where $\boldsymbol{0}^{(N)}=(0,\ldots,0)^\top\in\mathbb{R}^N$, then $\boldsymbol{\lambda}$ is approximately distributed as

$$\boldsymbol{\lambda}\sim\int_0^\infty t^{\alpha-1}e^{-\beta t}g(t)\Big\{\prod_{\nu=1}^{N}(e^{-\lambda_\nu})^{t+a_0}(1-e^{-\lambda_\nu})^{a_1-1}\Big\}dt\approx\Big(\prod_{\nu=1}^{N}{\lambda_\nu}^{a_1-1}\Big)\int_0^\infty t^{\alpha-1}e^{-t(\beta+\lambda_\cdot)}g(t)dt, \tag{3.3.3}$$

where $\lambda_\cdot=\sum_{\nu=1}^{N}\lambda_\nu$. The density (3.3.3) corresponds to the prior considered by Ghosh and Parsian (1981) when $a_1=1$, to that considered by Komaki (2004) when $\beta=0$ and $g=g_1$, and to that considered by Komaki (2006) when $\alpha=ma_1-1$, $\beta=0$, and $g(t)=\{t/(1+\kappa t)\}^{c+1}$ for all $t\in(0,\infty)$ for some $c>-ma_1$ and $\kappa>0$. However, in order to prove the results in the next subsection, we need to extend the technique of the proof of Theorem 1 of Komaki (2015), who considered an unbalanced problem.

3.3.2 Dominance results

In order to derive an explicit form of a Bayes solution with respect to the prior (3.3.1), we define

$$K(\alpha,\beta,g,\xi_0,\boldsymbol{\xi})=\int_0^\infty t^{\alpha-1}e^{-\beta t}g(t)\Big\{\prod_{\nu=1}^{N}\frac{\Gamma(t+\xi_0)}{\Gamma(t+\xi_0+\xi_\nu)}\Big\}dt \tag{3.3.4}$$

for $\xi_0 \geq 0$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)^\top \in [0, \infty)^N$ and we let $\boldsymbol{j}^{(N)} = (1, \ldots, 1)^\top \in \mathbb{R}^N$. For now, we consider the case of $a_0 = -m$ and $\boldsymbol{a} = \boldsymbol{j}^{(m)}$ and assume that either

$$r > m \quad \text{and} \quad \int_1^\infty t^{\alpha - Nm - 1} e^{-\beta t} g(t) dt < \infty \tag{3.3.5}$$

or

$$r = m, \quad \int_0^1 t^{\alpha - N - 1} e^{-\beta t} g(t) dt < \infty, \quad \int_1^\infty t^{\alpha - Nm - 1} e^{-\beta t} g(t) dt < \infty. \tag{3.3.6}$$

Then, by Lemma 3.3.1, the posterior under the prior $\boldsymbol{p} \sim \pi_{\alpha,\beta,g,-m,\boldsymbol{j}^{(m)}}(\boldsymbol{p})$ is proper, $K(\alpha, \beta, g, r - m, \boldsymbol{z} + m\boldsymbol{j}^{(N)}) < \infty$ for all $\boldsymbol{z} \in {\mathbb{N}_0}^N$, and $K(\alpha + 1, \beta, g, r - m, \boldsymbol{z} + m\boldsymbol{j}^{(N)}) < \infty$ for all $\boldsymbol{z} \in {\mathbb{N}_0}^N \setminus \{\boldsymbol{0}^{(N)}\}$.

Define the function $\delta^{(\alpha,\beta,g)} \colon {\mathbb{N}_0}^N \to (0, \infty]$ by

$$\delta^{(\alpha,\beta,g)}(\boldsymbol{z}) = \frac{K(\alpha + 1, \beta, g, r - m, \boldsymbol{z} + m\boldsymbol{j}^{(N)})}{K(\alpha, \beta, g, r - m, \boldsymbol{z} + m\boldsymbol{j}^{(N)})}, \quad \boldsymbol{z} \in {\mathbb{N}_0}^N.$$

Let, for $i \in \{1, \ldots, m\}$ and $\nu \in \{1, \ldots, N\}$,

$$\begin{aligned} \hat{p}_{i,\nu}^{(\alpha,\beta,g)} &= \begin{cases} \dfrac{X_{i,\nu}}{r + X_{\cdot,\nu} - 1 + \delta^{(\alpha,\beta,g)}(\boldsymbol{X}_\cdot)}, & \text{if } X_{i,\nu} \geq 1, \\ 0, & \text{if } X_{i,\nu} = 0, \end{cases} \\ &= \frac{X_{i,\nu}}{r + X_{\cdot,\nu} - 1 + \delta^{(\alpha,\beta,g)}(\boldsymbol{X}_\cdot)} \end{aligned} \tag{3.3.7}$$

and let $\hat{\boldsymbol{p}}^{(\alpha,\beta,g)} = (\hat{p}_{i,\nu}^{(\alpha,\beta,g)})_{1 \leq i \leq m,\, 1 \leq \nu \leq N}$. Then $\hat{\boldsymbol{p}}^{(\alpha,\beta,g)}$ is our hierarchical Bayes estimator.

Lemma 3.3.2 *Suppose that (3.3.5) or (3.3.6) holds. Then the shrinkage estimator $\hat{\boldsymbol{p}}^{(\alpha,\beta,g)}$ is a Bayes solution with respect to the prior (3.3.1) with $a_0 = -m$ and $\boldsymbol{a} = \boldsymbol{j}^{(m)}$ under the loss (3.1.3) for every $n \in \{1, \ldots, N\}$.*

The term $\delta^{(\alpha,\beta,g)}(\boldsymbol{X}_\cdot)$ is at once expressed in closed form and symmetric in $X_{\cdot,1}, \ldots, X_{\cdot,N}$. Deriving such terms will be less straightforward in the case of empirical Bayes estimation except for those that are dependent only on $X_{\cdot,\cdot}$.

Let $\boldsymbol{e}_\nu^{(N)}$ denote the νth unit vector in $\mathbb{R}^N$, namely the νth column of the $N \times N$ identity matrix, for $\nu \in \{1, \ldots, N\}$. The function $\delta^{(\alpha,\beta,g)}$ satisfies the following properties.

Proposition 3.3.1 *Let $\boldsymbol{z} = (z_1, \ldots, z_N)^\top \in {\mathbb{N}_0}^N$ and suppose that (3.3.5) or (3.3.6) holds.*

(i) *We have $0 < \delta^{(\alpha,\beta,g)}(\boldsymbol{z}) \leq \infty$. Furthermore, $\delta^{(\alpha,\beta,g)}(\boldsymbol{z}) = \infty$ only if $\boldsymbol{z} = \boldsymbol{0}^{(N)}$.*

(ii) *Let $\nu \in \{1, \ldots, N\}$. Then $\delta^{(\alpha,\beta,g)}(\boldsymbol{z}) \geq \delta^{(\alpha,\beta,g)}(\boldsymbol{z} + \boldsymbol{e}_\nu^{(N)})$.*

(iii) *Let $\nu \in \{1, \ldots, N\}$. Then $\lim_{\mathbb{N} \ni k \to \infty} \delta^{(\alpha,\beta,g)}(\boldsymbol{z} + k\boldsymbol{e}_\nu^{(N)}) = 0$.*

(iv) *Suppose that $r > m$, that $\lim_{t \to 0} g(t) = g(0) \in (0, \infty)$, and that $\alpha + 1 < N$. Then $\lim_{\mathbb{N} \setminus \{1\} \ni k \to \infty} [\delta^{(\alpha,\beta,g)}(\boldsymbol{z} + k\boldsymbol{j}^{(N)}) / \{(\alpha/N) / \log k\}] = 1$.*

Properties (iii) and (iv) above are in contrast to the fact that $\lim_{z\to\infty}\delta^{\mathrm{EB}}(z) = 1+m > 0$.

The following theorem provides a sufficient condition for $\hat{\boldsymbol{p}}^{(\alpha,\beta,g)}$ to dominate $\hat{\boldsymbol{p}}^{\mathrm{U}}$.

Theorem 3.3.1 *Let $n \in \{1,\dots,N\}$. Assume that (3.3.5) or (3.3.6) holds. Assume that g is nonincreasing. Suppose further that*

$$\alpha + 1 \le \min\{n(m-2), nm/2 + \beta r\}. \tag{3.3.8}$$

Then $\hat{\boldsymbol{p}}^{(\alpha,\beta,g)}$ is a hierarchical Bayes estimator dominating the UMVU estimator $\hat{\boldsymbol{p}}^{\mathrm{U}}$ under the loss $L_n(\boldsymbol{d},\boldsymbol{p})$ given by (3.1.3).

There exist $\alpha > 0$ and $\beta \ge 0$ satisfying assumption (3.3.8) if and only if $n(m-2) > 1$. When $r = m$ and g is nonincreasing, the condition $\int_0^1 t^{\alpha-N-1}e^{-\beta t}g(t)dt < \infty$ becomes $\alpha > N$. Even if $r = m$, the conditions of Theorem 3.3.1 can be satisfied when m is sufficiently large.

In the remainder of this subsection, we consider the problem of estimating $\boldsymbol{p}$ under the loss (3.1.6), a weighted version of Stein's loss, in order to show some robustness of our prior. Since the risk function of the UMVU estimator $\hat{\boldsymbol{p}}^{\mathrm{U}}$ is not defined under the loss (3.1.6) as well as under Stein's loss, we first derive the Jeffreys prior.

Lemma 3.3.3 *The Dirichlet prior (3.1.4) with $a_0 = (1-m)/2$ and $\boldsymbol{a} = \boldsymbol{j}^{(m)}/2$ is the Jeffreys prior.*

We note that if Stein's loss is used instead of the loss (3.1.6), the posterior risk with respect to the Jeffreys prior is identically infinite when at least one component of the matrix $(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$ is 0.

Next we show that under the loss (3.1.6), Bayes estimators are obtained as posterior means of $\boldsymbol{p}$.

Lemma 3.3.4 *Let $\boldsymbol{p} \sim \pi(\boldsymbol{p})$ be a strictly positive prior density and assume that the posterior is proper, that is, that $\int_{{D_m}^N}\{\prod_{\nu=1}^N \mathrm{NM}_m(\boldsymbol{x}_\nu|r,\boldsymbol{p}_\nu)\}\pi(\boldsymbol{p})d\boldsymbol{p} < \infty$ for all $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N \in {\mathbb{N}_0}^m$. Then the posterior mean of $\boldsymbol{p}$ is a Bayes solution under the loss (3.1.6) for every $n \in \{1,\dots,N\}$.*

The posterior under the Dirichlet prior (3.1.4) is proper if and only if $r + a_0 > 0$, in which case the posterior mean of $\boldsymbol{p}$ is

$$\hat{\boldsymbol{p}}^{(a_0,\boldsymbol{a})} = (\hat{p}_{i,\nu}^{(a_0,\boldsymbol{a})})_{1\le i\le m,\,1\le\nu\le N} = \Big(\frac{X_{i,\nu}+a_i}{r+a_0+X_{\cdot,\nu}+a_\cdot}\Big)_{1\le i\le m,\,1\le\nu\le N}. \tag{3.3.9}$$

The posterior under the hierarchical prior (3.3.1) is proper if and only if the condition of part (ii) of Lemma 3.3.1 is satisfied. In this case, the posterior mean of $\boldsymbol{p}$ is

$$\hat{\boldsymbol{p}}^{(\alpha,\beta,g,a_0,\boldsymbol{a})} = (\hat{p}_{i,\nu}^{(\alpha,\beta,g,a_0,\boldsymbol{a})})_{1\le i\le m,\,1\le\nu\le N} = \Big(\frac{X_{i,\nu}+a_i}{r+a_0+X_{\cdot,\nu}+a_\cdot+\delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{X}_\cdot)}\Big)_{1\le i\le m,\,1\le\nu\le N},$$

where $\delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}\colon {\mathbb{N}_0}^N \to (0,\infty)$ is the function defined by

$$\delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z}) = \frac{K(\alpha+1,\beta,g,r+a_0,\boldsymbol{z}+a_\cdot\boldsymbol{j}^{(N)}+\boldsymbol{e}_\nu^{(N)})}{K(\alpha,\beta,g,r+a_0,\boldsymbol{z}+a_\cdot\boldsymbol{j}^{(N)}+\boldsymbol{e}_\nu^{(N)})}, \quad \boldsymbol{z} \in {\mathbb{N}_0}^N,$$

for $\nu \in \{1,\dots,N\}$. Some properties of the functions $\delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}$, $\nu \in \{1,\dots,N\}$, are given in the following proposition, which corresponds to Proposition 3.3.1.

Proposition 3.3.2 *Let $\boldsymbol{z} = (z_1, \ldots, z_N)^\top \in {\mathbb{N}_0}^N$ and $\nu \in \{1, \ldots, N\}$. Suppose that the condition of part (ii) of Lemma 3.3.1 is satisfied.*

(i) *We have $0 < \delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z}) < \infty$.*

(ii) *Let $\nu' \in \{1, \ldots, N\}$. Then $\delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z}) \geq \delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z} + \boldsymbol{e}_{\nu'}^{(N)})$.*

(iii) *Let $\nu' \in \{1, \ldots, N\}$. Then $\lim_{\mathbb{N} \ni k \to \infty} \delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z} + k\boldsymbol{e}_{\nu'}^{(N)}) = 0$.*

(iv) *Suppose that $r + a_0 > 0$, that $\lim_{t \to 0} g(t) = g(0) \in (0, \infty)$, and that $\alpha + 1 < N$. Then $\lim_{\mathbb{N} \setminus \{1\} \ni k \to \infty} [\delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z} + k\boldsymbol{j}^{(N)}) / \{(\alpha/N)/\log k\}] = 1$.*

Theorem 3.3.2 provides a sufficient condition for $\hat{\boldsymbol{p}}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}$ to dominate $\hat{\boldsymbol{p}}^{(a_0,\boldsymbol{a})}$ under the loss (3.1.6).

Theorem 3.3.2 *Let $n \in \{1, \ldots, N\}$. Assume that the condition of part (ii) of Lemma 3.3.1 is satisfied. Assume that g is nonincreasing. Suppose further that $a_0 + a_\cdot + 1 \geq 0$ and that*

$$\alpha + 1 \leq n(-a_0 - 2). \tag{3.3.10}$$

Then $\hat{\boldsymbol{p}}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}$ dominates $\hat{\boldsymbol{p}}^{(a_0,\boldsymbol{a})}$ under the loss $\tilde{L}_n(\tilde{\boldsymbol{d}}, \boldsymbol{p})$ given by (3.1.6).

In particular, we have the following result for the case of the Jeffreys prior.

Corollary 3.3.1 *Let $n \in \{1, \ldots, N\}$. Assume that either*

- *$r > (m-1)/2$ and $\int_1^\infty t^{\alpha - Na_\cdot - 1} e^{-\beta t} g(t) dt < \infty$*

or

- *$r = (m-1)/2$, $\alpha > N$, and $\int_1^\infty t^{\alpha - Na_\cdot - 1} e^{-\beta t} g(t) dt < \infty$.*

Assume that g is nonincreasing. Suppose further that

$$\alpha + 1 \leq n(m-5)/2. \tag{3.3.11}$$

Then $\hat{\boldsymbol{p}}^{(\alpha,\beta,g,(1-m)/2,\boldsymbol{j}^{(m)}/2)}$ dominates $\hat{\boldsymbol{p}}^{((1-m)/2,\boldsymbol{j}^{(m)}/2)}$ under the loss $\tilde{L}_n(\tilde{\boldsymbol{d}}, \boldsymbol{p})$ given by (3.1.6).

3.3.3 Posterior computation

In order to approximate the integral

$$K(\alpha, \beta, g, \xi_0, \boldsymbol{\xi}) = \int_0^\infty t^{\alpha-1} e^{-\beta t} g(t) \Big\{ \prod_{\nu=1}^N \frac{\Gamma(t + \xi_0)}{\Gamma(t + \xi_0 + \xi_\nu)} \Big\} dt,$$

we could in principle use i.i.d. gamma variables (when $\beta > 0$) or rewrite the integral as

$$K(\alpha, \beta, g, \xi_0, \boldsymbol{\xi}) = \int_0^1 \Big(\frac{\omega}{1-\omega}\Big)^{\alpha-1} e^{-\beta\omega/(1-\omega)} g\Big(\frac{\omega}{1-\omega}\Big) \Big\{ \prod_{\nu=1}^N \frac{\Gamma(\omega/(1-\omega) + \xi_0)}{\Gamma(\omega/(1-\omega) + \xi_0 + \xi_\nu)} \Big\} \frac{1}{(1-\omega)^2} d\omega$$

and use i.i.d. uniform variables, for example. However, this can be numerically unstable because the gamma function appears in the integrand. If $\boldsymbol{\xi} \in {\mathbb{N}_0}^N$, the problem would be alleviated to some extent by using the relation

$$\prod_{\nu=1}^{N} \frac{\Gamma(t+\xi_0)}{\Gamma(t+\xi_0+\xi_\nu)} = \prod_{\nu=1}^{N} \frac{1}{(t+\xi_0)\cdots(t+\xi_0+\xi_\nu-1)}$$

for all $t \in (0,\infty)$.

When $g = g_1$, a more convenient way to compute the hierarchical Bayes estimators in the previous subsection is to use MCMC samples since they are functions of posterior expectations. In order to describe a Gibbs sampler, we introduce a fully conjugate prior. For $\alpha > 0$, $\beta \geq 0$, $a_0 \in \mathbb{R}$, and $(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N) = ((a_{1,1},\ldots,a_{m,1})^\top,\ldots,(a_{1,N},\ldots,a_{m,N})^\top) \in (0,\infty)^{m\times N}$, let

$$\pi(\boldsymbol{p},t|\alpha,\beta,a_0,\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N) = t^{\alpha-1}e^{-\beta t}\prod_{\nu=1}^{N}\Big({p_{0,\nu}}^{t+a_0-1}\prod_{i=1}^{m}{p_{i,\nu}}^{a_{i,\nu}-1}\Big) \tag{3.3.12}$$

and

$$\pi(\boldsymbol{p}|\alpha,\beta,a_0,\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N) = \Gamma(\alpha)\Big\{\prod_{\nu=1}^{N}\Big({p_{0,\nu}}^{a_0-1}\prod_{i=1}^{m}{p_{i,\nu}}^{a_{i,\nu}-1}\Big)\Big\}\Big/\Big(\beta+\sum_{\nu=1}^{N}\log\frac{1}{p_{0,\nu}}\Big)^{\alpha}, \tag{3.3.13}$$

where $t \in (0,\infty)$ and where $\boldsymbol{p} = ((p_{1,1},\ldots,p_{m,1})^\top,\ldots,(p_{1,N},\ldots,p_{m,N})^\top) \in {D_m}^N$ and $p_{0,\nu} = 1-\sum_{i=1}^{m}p_{i,\nu}$ for $\nu \in \{1,\ldots,N\}$. When $\boldsymbol{a}_1 = \cdots = \boldsymbol{a}_N = \boldsymbol{a}$, the prior (3.3.13) becomes the original prior (3.3.2).

Some basic properties of the priors (3.3.12) and (3.3.13) are summarized in the following proposition. Let $a_{\cdot,\nu} = \sum_{i=1}^{m}a_{i,\nu}$ for $\nu \in \{1,\ldots,N\}$ and let $a_{\cdot,\cdot} = \sum_{\nu=1}^{N}a_{\cdot,\nu}$.

Proposition 3.3.3 *The priors (3.3.12) and (3.3.13) satisfy the following properties:*

(i) *The following statements are equivalent:*

- $\int_{{D_m}^N\times(0,\infty)}\pi(\boldsymbol{p},t|\alpha,\beta,a_0,\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N)d(\boldsymbol{p},t) < \infty$.
- $\int_{{D_m}^N}\pi(\boldsymbol{p}|\alpha,\beta,a_0,\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N)d\boldsymbol{p} < \infty$.
- $\min\{\max\{a_0,\alpha-N\},\max\{a_{\cdot,\cdot}-\alpha,\beta\}\} > 0$.

(ii) *If* $\boldsymbol{p} \sim \pi(\boldsymbol{p}|\alpha,\beta,a_0,\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N)$ *and* $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)|\boldsymbol{p} \sim \prod_{\nu=1}^{N}\mathrm{NM}_m(\boldsymbol{x}_\nu|r,\boldsymbol{p}_\nu)$, *then*

$$\boldsymbol{p}|(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N) \sim \pi(\boldsymbol{p}|\alpha,\beta,r+a_0,\boldsymbol{x}_1+\boldsymbol{a}_1,\ldots,\boldsymbol{x}_N+\boldsymbol{a}_N).$$

(iii) *If* $(\boldsymbol{p},t) \sim \pi(\boldsymbol{p},t|\alpha,\beta,a_0,\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N)$, *then* $\boldsymbol{p} \sim \pi(\boldsymbol{p}|\alpha,\beta,a_0,\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N)$.

(iv) *If* $(\boldsymbol{p},t) \sim \pi(\boldsymbol{p},t|\alpha,\beta,a_0,\boldsymbol{a}_1,\ldots,\boldsymbol{a}_N)$, *then*

$$\boldsymbol{p}|t \sim \prod_{\nu=1}^{N}\mathrm{Dir}_m(\boldsymbol{p}_\nu|t+a_0,\boldsymbol{a}_\nu), \quad t|\boldsymbol{p} \sim \mathrm{Ga}\Big(t\Big|\alpha,\beta+\sum_{\nu=1}^{N}\log\frac{1}{p_{0,\nu}}\Big).$$

Part (ii) of Proposition 3.3.3 shows that the prior (3.3.13) is conjugate. Furthermore, part (iii) of the proposition shows that in order to generate samples of $\boldsymbol{p}$ from the prior (3.3.13), it is sufficient to sample from the joint prior (3.3.12). Therefore, we describe a Gibbs sampler for (3.3.12) based on part (iv) of the proposition. In order to generate MCMC samples corresponding to (3.3.12) when it is proper, given a current sample of $(\boldsymbol{p}, t)$, denoted by $(\tilde{\boldsymbol{p}}, \tilde{t}) = ((\tilde{p}_{i,\nu})_{1\le i\le m,\,1\le \nu\le N}, \tilde{t})$, we generate a new sample as follows:

- sample $t^* \sim \mathrm{Ga}\big(t\big|\alpha, \beta + \sum_{\nu=1}^N \log\big\{1/\big(1 - \sum_{i=1}^m \tilde{p}_{i,\nu}\big)\big\}\big)$;
- sample $\boldsymbol{p}^* \sim \prod_{\nu=1}^N \mathrm{Dir}_m(\boldsymbol{p}_\nu | t^* + a_0, \boldsymbol{a}_\nu)$.

Then samples of $\boldsymbol{p}$ can be used to approximate expectations of functions of $\boldsymbol{p} \sim \pi(\boldsymbol{p}|\alpha, \beta, a_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_N)$. Also, samples of t may be used to approximate $\delta^{(\alpha,\beta,g)}(\boldsymbol{z})$ and $\delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z})$, $\boldsymbol{z} \in {\mathbb{N}_0}^N$, $\nu \in \{1, \ldots, N\}$, even if $g \ne g_1$.

3.4 Simulation Study

In this section, we investigate through simulation the numerical performance of the risk functions of the Bayes estimators given in the previous two sections under the standardized squared error loss given by (3.1.3) with $n = N$. The estimators which we compare are the following four:

U: the UMVU estimator $\hat{\boldsymbol{p}}^{\mathrm{U}}$ given by (3.2.1),

EB0: the alternative empirical Bayes estimator which estimates $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N$ independently based on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$, respectively, namely $\hat{\boldsymbol{p}}^{\mathrm{EB0}} = (X_{i,\nu}/(r + X_{\cdot,\nu} + m + mr/X_{\cdot,\nu}))_{1\le i\le m,\,1\le\nu\le N}$,

EB: the empirical Bayes estimator $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ given by (3.2.3),

HB: the hierarchical Bayes estimator $\hat{\boldsymbol{p}}^{\mathrm{HB}} = \hat{\boldsymbol{p}}^{(\alpha,1,g_1)}$ given by (3.3.7) with $(\beta, g) = (1, g_1)$.

We consider the following cases:

(i) Let $(r, m, N) = (8, 7, 1)$, $\alpha = 4$, and $\boldsymbol{p} = \boldsymbol{p}^{(1)}(1), \boldsymbol{p}^{(1)}(2), \boldsymbol{p}^{(1)}(3)$, where

$$\boldsymbol{p}^{(1)}(1) = (1,1,1,1,1,1,1)^\top/8, \quad \boldsymbol{p}^{(1)}(2) = (1,1,1,1,2,2,2)^\top/12, \quad \boldsymbol{p}^{(1)}(3) = (2,2,2,2,1,1,1)^\top/12.$$

(ii) Let $(r, m, N) = (8, 7, 3)$, $\alpha = 14$, and $\boldsymbol{p} = \boldsymbol{p}^{(2)}(1), \boldsymbol{p}^{(2)}(2), \boldsymbol{p}^{(2)}(3)$, where

$$\begin{aligned}
\boldsymbol{p}^{(2)}(1) &= ((1,1,1,1,1,1,1)^\top/8, (1,1,1,1,1,1,1)^\top/8, (1,1,1,1,1,1,1)^\top/8),\\
\boldsymbol{p}^{(2)}(2) &= ((1,1,1,1,2,2,2)^\top/12, (1,1,1,1,1,1,1)^\top/8, (1,1,1,1,2,2,2)^\top/12),\\
\boldsymbol{p}^{(2)}(3) &= ((1,1,1,1,2,2,2)^\top/12, (1,1,1,1,1,1,1)^\top/8, (2,2,2,2,1,1,1)^\top/12).
\end{aligned}$$

(iii) Let $(r, m, N) = (4, 3, 7)$, $\alpha = 6$, and $\boldsymbol{p} = \boldsymbol{p}^{(3)}(1), \boldsymbol{p}^{(3)}(2), \boldsymbol{p}^{(3)}(3)$, where

$$\boldsymbol{p}^{(3)}(1) = \left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} /4 \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} /4 \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} /4 \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} /4 \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} /4 \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} /4 \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} /4 \right),$$

$$\boldsymbol{p}^{(3)}(2) = \left(\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} /6 \quad \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} /6 \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} /4 \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} /4 \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} /4 \quad \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} /6 \quad \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} /6 \right),$$

$$\boldsymbol{p}^{(3)}(3) = \left(\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} /6 \quad \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} /6 \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} /4 \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} /4 \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} /4 \quad \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} /6 \quad \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} /6 \right).$$

(iv) Let $(r, m, N) = (2, 1, 7)$, $\alpha = 6$, and $\boldsymbol{p} = \boldsymbol{p}^{(4)}(1), \boldsymbol{p}^{(4)}(2), \boldsymbol{p}^{(4)}(3)$, where

$$\boldsymbol{p}^{(4)}(1) = (1/2, 1/2, 1/2, 1/2, 1/2, 1/2, 1/2), \quad \boldsymbol{p}^{(4)}(2) = (1/3, 1/3, 1/2, 1/2, 1/2, 1/3, 1/3),$$
$$\boldsymbol{p}^{(4)}(3) = (1/3, 1/3, 1/2, 1/2, 1/2, 2/3, 2/3).$$

Case (ii) is a case where $m > N$ while Case (iii) is where $m < N$. Case (i) corresponds to a single negative multinomial observation while Case (iv) corresponds to multiple negative binomial observations. Table 3.1 summarizes whether the sufficient conditions for dominance in Sections 3.2 and 3.3 are applicable.

Table 3.1: Whether or not the conditions for dominance are satisfied by the alternative empirical Bayes estimator (EB0), the proposed empirical Bayes estimator (EB), and the hierarchical Bayes estimator with $(\beta, g) = (1, g_1)$ (HB). When one of the conditions is satisfied, $+$ is marked, and $-$ is marked otherwise.

Case	EB0	EB	HB
(i)	$+$	$+$	$+$
(ii)	$+$	$+$	$+$
(iii)	$-$	$-$	$+$
(iv)	$-$	$-$	$-$

For each estimator $\hat{\boldsymbol{p}}$, we obtain approximated values of the risk function $E[L_N(\hat{\boldsymbol{p}}, \boldsymbol{p})]$ by simulation with $1,000$ replications. The hierarchical Bayes estimator $\hat{\boldsymbol{p}}^{\mathrm{HB}}$ was computed based on the Gibbs sampler described in Section 3.3.3 by generating $50,000$ posterior samples after discarding the first $50,000$ samples. The percentage relative improvement in average loss (PRIAL) of an estimator $\hat{\boldsymbol{p}}$ over $\hat{\boldsymbol{p}}^{\mathrm{U}}$ is defined by

$$\mathrm{PRIAL} = 100\{E[L_N(\hat{\boldsymbol{p}}^{\mathrm{U}}, \boldsymbol{p})] - E[L_N(\hat{\boldsymbol{p}}, \boldsymbol{p})]\}/E[L_N(\hat{\boldsymbol{p}}^{\mathrm{U}}, \boldsymbol{p})].$$

For Case (i), Table 3.2 reports values of the risks of the estimators with values of PRIAL given in parentheses. Since $\hat{\boldsymbol{p}}^{\mathrm{EB0}}$ and $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ are identical, they have the same values of PRIAL. Although the dominance of $\hat{\boldsymbol{p}}^{\mathrm{HB}}$ over $\hat{\boldsymbol{p}}^{\mathrm{U}}$ is guaranteed, the difference in risks between the two estimators is small. It is clear from the values of PRIAL that $\hat{\boldsymbol{p}}^{\mathrm{EB0}}$ and $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ are superior to $\hat{\boldsymbol{p}}^{\mathrm{HB}}$ in this case.

Table 3.2: Risks of the UMVU estimator (U), the alternative empirical Bayes estimator (EB0), the proposed empirical Bayes estimator (EB), and the hierarchical Bayes estimator with $(\beta, g) = (1, g_1)$ (HB) for Case (i). Values of PRIAL of EB0, EB, and HB are given in parentheses.

$\boldsymbol{p}$	U	EB0	EB	HB
$\boldsymbol{p}^{(1)}(1)$	0.11	0.10 (7.58)	0.10 (7.58)	0.10 (2.86)
$\boldsymbol{p}^{(1)}(2)$	0.15	0.13 (11.15)	0.13 (11.15)	0.14 (4.59)
$\boldsymbol{p}^{(1)}(3)$	0.07	0.07 (4.92)	0.07 (4.92)	0.07 (1.75)

For Case (ii), Table 3.3 reports values of the risks and PRIAL. In all cases, the risk values of $\hat{\boldsymbol{p}}^{\mathrm{EB0}}$ are smaller than those of $\hat{\boldsymbol{p}}^{\mathrm{HB}}$, and the risk values of $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ are still smaller. These three estimators have the largest values of PRIAL when $\boldsymbol{p} = \boldsymbol{p}^{(2)}(2)$. Also, it can be seen that in the balanced case of $\boldsymbol{p} = \boldsymbol{p}^{(2)}(1)$, the risk values of the three estimators are smaller than those of $\hat{\boldsymbol{p}}^{\mathrm{U}}$ even when the loss is (3.1.3) with $n = 1$.

Table 3.3: Risks of the UMVU estimator (U), the alternative empirical Bayes estimator (EB0), the proposed empirical Bayes estimator (EB), and the hierarchical Bayes estimator with $(\beta, g) = (1, g_1)$ (HB) for Case (ii). Values of PRIAL of EB0, EB, and HB are given in parentheses.

$\boldsymbol{p}$	U	EB0	EB	HB
$\boldsymbol{p}^{(2)}(1)$	0.32	0.30 (7.18)	0.29 (7.91)	0.31 (3.70)
$\boldsymbol{p}^{(2)}(2)$	0.39	0.35 (9.65)	0.35 (10.61)	0.37 (5.17)
$\boldsymbol{p}^{(2)}(3)$	0.32	0.29 (9.10)	0.29 (9.92)	0.31 (4.15)

For Case (iii), Table 3.4 reports values of the risks and PRIAL. Although the empirical Bayes estimators do not satisfy the condition of Corollary 3.2.1, $\hat{\boldsymbol{p}}^{\mathrm{EB0}}$ is competitive with $\hat{\boldsymbol{p}}^{\mathrm{HB}}$ and $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ is superior to $\hat{\boldsymbol{p}}^{\mathrm{HB}}$. Importantly, even if the loss is (3.1.3) with $n = 1$, $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ has smaller values of risks than $\hat{\boldsymbol{p}}^{\mathrm{EB0}}$ when $\boldsymbol{p} = \boldsymbol{p}^{(3)}(1)$.

Table 3.4: Risks of the UMVU estimator (U), the alternative empirical Bayes estimator (EB0), the proposed empirical Bayes estimator (EB), and the hierarchical Bayes estimator with $(\beta, g) = (1, g_1)$ (HB) for Case (iii). Values of PRIAL of EB0, EB, and HB are given in parentheses.

$\boldsymbol{p}$	U	EB0	EB	HB
$\boldsymbol{p}^{(3)}(1)$	1.21	1.16 (4.25)	1.10 (9.27)	1.14 (6.16)
$\boldsymbol{p}^{(3)}(2)$	1.44	1.30 (9.90)	1.23 (14.82)	1.32 (8.55)
$\boldsymbol{p}^{(3)}(3)$	1.22	1.15 (6.21)	1.08 (11.58)	1.14 (6.87)

Finally, Table 3.5 reports values of the risks and PRIAL for Case (iv). The estimators $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ and $\hat{\boldsymbol{p}}^{\mathrm{HB}}$ do not satisfy the conditions for dominance but their risk values are smaller than those of $\hat{\boldsymbol{p}}^{\mathrm{U}}$. In particular, $\hat{\boldsymbol{p}}^{\mathrm{HB}}$ has large values of PRIAL. In contrast to Case (i), $\hat{\boldsymbol{p}}^{\mathrm{HB}}$ is superior to $\hat{\boldsymbol{p}}^{\mathrm{EB0}}$ and $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ in the present case where N is much larger than m.

Table 3.5: Risks of the UMVU estimator (U), the alternative empirical Bayes estimator (EB0), the proposed empirical Bayes estimator (EB), and the hierarchical Bayes estimator with $(\beta, g) = (1, g_1)$ (HB) for Case (iv). Values of PRIAL of EB0, EB, and HB are given in parentheses.

$\boldsymbol{p}$	U	EB0	EB	HB
$\boldsymbol{p}^{(4)}(1)$	1.34	1.38 (−3.35)	1.33 (0.75)	1.00 (24.99)
$\boldsymbol{p}^{(4)}(2)$	1.72	1.32 (23.43)	1.30 (24.39)	1.11 (35.47)
$\boldsymbol{p}^{(4)}(3)$	1.36	1.28 (5.78)	1.23 (9.62)	0.97 (28.42)

3.5 Discussion

In this chapter, we considered the simultaneous estimation of negative multinomial parameter vectors and in particular derived empirical Bayes and hierarchical Bayes estimators which, under suitable conditions, dominate the UMVU estimator for the loss (3.1.3). The focus was on their basic properties in the relatively simple setting of this chapter. There are several related problems that need to be further addressed and some of which are briefly discussed in this section.

3.5.1 Inadmissibility of the UMVU estimator

Corollary 3.2.1 shows that the UMVU estimator $\hat{\boldsymbol{p}}^{\mathrm{U}}$ is inadmissible under the loss (3.1.3) for every $n \in \{1, \dots, N\}$ whenever $m \geq 7$ and $r \geq 5/2$. On the other hand, Theorem 3.3.1 shows that $\hat{\boldsymbol{p}}^{\mathrm{U}}$ is inadmissible for the loss (3.1.3) if either $r > m$ and $n(m-2) > 1$ or $r = m$ and $n(m-2) > N+1$. Thus, if $n \geq 2$, the UMVU estimator $\hat{\boldsymbol{p}}^{\mathrm{U}}$ is inadmissible for large m when $r \geq 5/2$ while it is inadmissible for large r when $m \geq 3$. It is also interesting to investigate admissibility of $\hat{\boldsymbol{p}}^{\mathrm{U}}$ for small r and m, which will be studied in a future paper.

3.5.2 Empirical Bayes estimation under the loss (3.1.6)

One empirical Bayes estimator under the loss (3.1.6) would be

$$\tilde{\boldsymbol{p}}^{(a_0,\boldsymbol{a})} = \Big(\frac{X_{i,\nu} + a_i}{r + a_0 + X_{\cdot,\nu} + a_\cdot + \tilde{\delta}^{(a_0,\boldsymbol{a})}(X_{\cdot,\cdot})} \Big)_{1 \leq i \leq m,\, 1 \leq \nu \leq N},$$

where

$$\tilde{\delta}^{(a_0,\boldsymbol{a})}(X_{\cdot,\cdot}) = \begin{cases} 1 - a_0 + Nra_\cdot / X_{\cdot,\cdot}, & \text{if } X_{\cdot,\cdot} \geq 1, \\ 0, & \text{if } X_{\cdot,\cdot} = 0. \end{cases}$$

Although the empirical Bayes estimator (3.2.3) performed better than the UMVU estimator in all of the four cases of Section 3.4, the risk values of the above empirical Bayes estimator $\tilde{\boldsymbol{p}}^{(a_0,\boldsymbol{a})}$ were larger than those of $\hat{\boldsymbol{p}}^{(a_0,\boldsymbol{a})}$ defined in (3.3.9) when we conducted a similar simulation study, where $a_0 = (1-m)/2$ and $\boldsymbol{a} = \boldsymbol{j}^{(m)}/2$ as in Lemma 3.3.3. This might be partly because $\tilde{\delta}^{(a_0,\boldsymbol{a})}(X_{\cdot,\cdot})$ is too large when $X_{\cdot,\cdot} \geq 1$, since the loss (3.1.6) penalizes small components of each estimator. On the other hand, if we set $a_0 = -m$ and $\boldsymbol{a} = \boldsymbol{j}^{(m)}$ as in Section 3.4, $\tilde{\boldsymbol{p}}^{(a_0,\boldsymbol{a})}$ performs well compared with $\hat{\boldsymbol{p}}^{(a_0,\boldsymbol{a})}$ in several cases as shown in Table 3.6, which reports values of the

risks and PRIAL of these two estimators. In Table 3.6, the settings are as in Section 3.4, B and EB† denote $\hat{\boldsymbol{p}}^{(a_0,\boldsymbol{a})}$ and $\tilde{\boldsymbol{p}}^{(a_0,\boldsymbol{a})}$, respectively, and the PRIAL of $\tilde{\boldsymbol{p}}^{(a_0,\boldsymbol{a})}$ over $\hat{\boldsymbol{p}}^{(a_0,\boldsymbol{a})}$ is defined analogously to that considered in Section 3.4.

Table 3.6: Risks of the Bayes estimator $\hat{\boldsymbol{p}}^{(a_0,\boldsymbol{a})}$ with $a_0 = -m$ and $\boldsymbol{a} = \boldsymbol{j}^{(m)}$ (B) and the empirical Bayes estimator $\tilde{\boldsymbol{p}}^{(a_0,\boldsymbol{a})}$ with $a_0 = -m$ and $\boldsymbol{a} = \boldsymbol{j}^{(m)}$ (EB†) for the four cases of Section 3.4. Values of PRIAL of EB† are given in parentheses.

$\boldsymbol{p}$	B	EB†
$\boldsymbol{p}^{(1)}(1)$	0.05	0.05 (8.20)
$\boldsymbol{p}^{(1)}(2)$	0.06	0.06 (11.30)
$\boldsymbol{p}^{(1)}(3)$	0.03	0.03 (4.86)
$\boldsymbol{p}^{(2)}(1)$	0.15	0.14 (8.50)
$\boldsymbol{p}^{(2)}(2)$	0.18	0.16 (11.78)
$\boldsymbol{p}^{(2)}(3)$	0.15	0.13 (10.31)
$\boldsymbol{p}^{(3)}(1)$	0.47	0.43 (7.48)
$\boldsymbol{p}^{(3)}(2)$	0.54	0.45 (16.31)
$\boldsymbol{p}^{(3)}(3)$	0.47	0.41 (12.74)
$\boldsymbol{p}^{(4)}(1)$	0.27	0.43 (−57.03)
$\boldsymbol{p}^{(4)}(2)$	0.46	0.38 (18.15)
$\boldsymbol{p}^{(4)}(3)$	0.33	0.37 (−12.12)

3.5.3 Extensions to unbalanced models

So far, we have mostly considered symmetric or balanced cases except that n can be smaller than N and that $a_1, \ldots, a_m$ might differ in general. More general unbalanced models will also be important. As an example, suppose that $\boldsymbol{X}_1 \sim \mathrm{NM}_m(r_1, \boldsymbol{p}_1), \ldots, \boldsymbol{X}_N \sim \mathrm{NM}_m(r_N, \boldsymbol{p}_N)$ for $r_1, \ldots, r_N > 1$. Then a straightforward generalization of Theorem 3.2.1 is that $(X_{i,\nu}/(r_\nu + X_{\cdot,\nu} - 1))_{1\le i\le m,\, 1\le \nu\le N}$ is dominated by $(X_{i,\nu}/(r_\nu + X_{\cdot,\nu} - 1 + \delta(X_{\cdot,\cdot})))_{1\le i\le m,\, 1\le \nu\le N}$ under the loss (3.1.3) if $r_1, \ldots, r_N \ge 5/2$, $z\delta(z) \le (z+1)\delta(z+1)$ for all $z \in \mathbb{N}$, and

- $(\overline{r}/\underline{r})^2\delta(z) \le 2(m-3)$ implies $\{2m - 6 - (\overline{r}/\underline{r})^2 m\}\delta(z) + 2(m-3)\underline{r} \ge 0$,
- $(\overline{r}/\underline{r})^2\delta(z) > 2(m-3)$ implies $n[\{2m-6-(\overline{r}/\underline{r})^2 m\}\delta(z)+2(m-3)\underline{r}] \ge (z-1)\{(\overline{r}/\underline{r})^2\delta(z) - 2(m-3)\}$

for all $z \in \mathbb{N} \setminus \{1\}$, where $\underline{r} = \min\{r_1, \ldots, r_N\}$ and $\overline{r} = \max\{r_1, \ldots, r_N\}$.

There are other possible extensions. For example, since the marginal distribution of any set of components of a negative multinomial random vector is also negative multinomial, it will be worthwhile to consider the more general case where the lengths of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ may not be the same. Weighted loss functions could be used. Also, we could use more than one function instead of δ and replace $X_{\cdot,\cdot}$ with some linear combination of components of $\boldsymbol{X}$. A generalization of the prior (3.3.1) is obtained by introducing another hyperparameter, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)^\top \in (0,\infty)^N$, and replacing ${p_{0,\nu}}^t$ with ${p_{0,\nu}}^{\gamma_\nu t}$ for $t \in (0,\infty)$ for each $\nu \in \{1, \ldots, N\}$.

3.6 Appendix: Proofs

Let $\mathbf{0}^{(m)} = (0, \ldots, 0)^\top \in \mathbb{R}^m$ and $\mathbf{0}^{(m,N)} = \mathbf{0}^{(m)}\mathbf{0}^{(N)\top} \in \mathbb{R}^{m\times N}$. Let $\boldsymbol{e}_i^{(m)}$ be the ith unit vector in $\mathbb{R}^m$, namely the ith column of the $m \times m$ identity matrix, for $i \in \{1, \ldots, m\}$. Let $\boldsymbol{e}_{i,\nu}^{(m,N)} = \boldsymbol{e}_i^{(m)}(\boldsymbol{e}_\nu^{(N)})^\top \in \mathbb{R}^{m\times N}$ for $i \in \{1, \ldots, m\}$ and $\nu \in \{1, \ldots, N\}$. Further let $\delta_{i,j}^{(m)} = \boldsymbol{e}_i^{(m)\top}\boldsymbol{e}_j^{(m)}$ for $i, j \in \{1, \ldots, m\}$ and let $\delta_{\nu,\nu'}^{(N)} = \boldsymbol{e}_\nu^{(N)\top}\boldsymbol{e}_{\nu'}^{(N)}$ for $\nu, \nu' \in \{1, \ldots, N\}$. For $\boldsymbol{v} = (v_{i,\nu})_{1\le i\le m, 1\le\nu\le N} \in \mathbb{R}^{m\times N}$ and $\tilde{\boldsymbol{v}} = (\tilde{v}_{i,\nu})_{1\le i\le m, 1\le\nu\le N} \in \mathbb{R}^{m\times N}$, we write the inner product $\sum_{\nu=1}^N \sum_{i=1}^m v_{i,\nu}\tilde{v}_{i,\nu}$ as $\boldsymbol{v}\cdot\tilde{\boldsymbol{v}}$. The following result is due to Hudson (1978).

Lemma 3.6.1 *Let $h\colon {\mathbb{N}_0}^{m\times N} \to \mathbb{R}$ and suppose that either $h(\boldsymbol{x}) \ge 0$ for all $\boldsymbol{x} \in {\mathbb{N}_0}^{m\times N}$ or $E[|h(\boldsymbol{X})|] < \infty$. Then for all $i \in \{1, \ldots, m\}$ and all $\nu \in \{1, \ldots, N\}$, if $h(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} = (x_{j,\nu'})_{1\le j\le m, 1\le\nu'\le N} \in {\mathbb{N}_0}^{m\times N}$ such that $x_{i,\nu} = 0$, we have*

$$E\Big[\frac{h(\boldsymbol{X})}{p_{i,\nu}}\Big] = E\Big[\frac{r + X_{\cdot,\nu}}{X_{i,\nu}+1} h(\boldsymbol{X} + \boldsymbol{e}_{i,\nu}^{(m,N)})\Big].$$

Proof. We have

$$\begin{aligned}
E\Big[\frac{h(\boldsymbol{X})}{p_{i,\nu}}\Big] &= \sum_{\boldsymbol{x}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)\in{\mathbb{N}_0}^{m\times N},\, \boldsymbol{x}\cdot\boldsymbol{e}_{i,\nu}^{(m,N)}\neq 0} \frac{h(\boldsymbol{x})}{p_{i,\nu}} \prod_{\nu'=1}^N \mathrm{NM}_m(\boldsymbol{x}_{\nu'}|r, \boldsymbol{p}_{\nu'}) \\
&= \sum_{\boldsymbol{x}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)\in{\mathbb{N}_0}^{m\times N}} \frac{h(\boldsymbol{x} + \boldsymbol{e}_{i,\nu}^{(m,N)})}{p_{i,\nu}} \frac{\mathrm{NM}_m(\boldsymbol{x}_\nu + \boldsymbol{e}_i^{(m)}|r, \boldsymbol{p}_\nu)}{\mathrm{NM}_m(\boldsymbol{x}_\nu|r, \boldsymbol{p}_\nu)} \prod_{\nu'=1}^N \mathrm{NM}_m(\boldsymbol{x}_{\nu'}|r, \boldsymbol{p}_{\nu'}) \\
&= E\Big[\frac{r + X_{\cdot,\nu}}{X_{i,\nu}+1} h(\boldsymbol{X} + \boldsymbol{e}_{i,\nu}^{(m,N)})\Big],
\end{aligned}$$

which proves the desired result. □

Proof of Theorem 3.2.1. Let $\Delta_n^{(\delta)} = E[L_n(\hat{\boldsymbol{p}}^{(\delta)}, \boldsymbol{p})] - E[L_n(\hat{\boldsymbol{p}}^{\mathrm{U}}, \boldsymbol{p})]$. For $\nu \in \{1, \ldots, N\}$, let

$$\phi_\nu^{(\delta)}(\boldsymbol{X}) = \frac{\delta(X_{\cdot,\cdot})}{r + X_{\cdot,\nu} - 1 + \delta(X_{\cdot,\cdot})}$$

so that for every $i \in \{1, \ldots, m\}$,

$$\hat{p}_{i,\nu}^{(\delta)} = \hat{p}_{i,\nu}^{\mathrm{U}} - \hat{p}_{i,\nu}^{\mathrm{U}}\phi_\nu^{(\delta)}(\boldsymbol{X}).$$

Then, by Lemma 3.6.1, we have

$$\begin{aligned}
\Delta_n^{(\delta)} &= E\Big[\sum_{\nu=1}^n \sum_{i=1}^m \Big(\frac{1}{p_{i,\nu}}[(\hat{p}_{i,\nu}^{\mathrm{U}})^2\{\phi_\nu^{(\delta)}(\boldsymbol{X})\}^2 - 2(\hat{p}_{i,\nu}^{\mathrm{U}})^2\phi_\nu^{(\delta)}(\boldsymbol{X})] + 2\hat{p}_{i,\nu}^{\mathrm{U}}\phi_\nu^{(\delta)}(\boldsymbol{X})\Big)\Big] \\
&= E\Big[\sum_{\nu=1}^n \sum_{i=1}^m \Big[\frac{X_{i,\nu}+1}{r + X_{\cdot,\nu}}\{\phi_\nu^{(\delta)}(\boldsymbol{X} + \boldsymbol{e}_{i,\nu}^{(m,N)})\}^2 - 2\frac{X_{i,\nu}+1}{r + X_{\cdot,\nu}}\phi_\nu^{(\delta)}(\boldsymbol{X} + \boldsymbol{e}_{i,\nu}^{(m,N)}) + 2\hat{p}_{i,\nu}^{\mathrm{U}}\phi_\nu^{(\delta)}(\boldsymbol{X})\Big]\Big] \\
&= E\Big[\sum_{\nu=1}^n \{I_{1,\nu}^{(\delta)}(\boldsymbol{X}) - 2I_{2,\nu}^{(\delta)}(\boldsymbol{X}) + 2I_{3,\nu}^{(\delta)}(\boldsymbol{X})\}\Big],
\end{aligned}$$

where

$$I_{1,\nu}^{(\delta)}(\boldsymbol{x}) = \frac{\sum_{i=1}^m x_{i,\nu} + m}{r + \sum_{i=1}^m x_{i,\nu}} \Big\{ \frac{\delta\big(\sum_{\nu'=1}^N \sum_{i=1}^m x_{i,\nu'} + 1\big)}{r + \sum_{i=1}^m x_{i,\nu} + \delta\big(\sum_{\nu'=1}^N \sum_{i=1}^m x_{i,\nu'} + 1\big)} \Big\}^2,$$

$$I_{2,\nu}^{(\delta)}(\boldsymbol{x}) = \frac{\sum_{i=1}^m x_{i,\nu} + m}{r + \sum_{i=1}^m x_{i,\nu}} \frac{\delta\big(\sum_{\nu'=1}^N \sum_{i=1}^m x_{i,\nu'} + 1\big)}{r + \sum_{i=1}^m x_{i,\nu} + \delta\big(\sum_{\nu'=1}^N \sum_{i=1}^m x_{i,\nu'} + 1\big)},$$

$$I_{3,\nu}^{(\delta)}(\boldsymbol{x}) = \frac{\sum_{i=1}^m x_{i,\nu}}{r + \sum_{i=1}^m x_{i,\nu} - 1} \frac{\delta\big(\sum_{\nu'=1}^N \sum_{i=1}^m x_{i,\nu'}\big)}{r + \sum_{i=1}^m x_{i,\nu} - 1 + \delta\big(\sum_{\nu'=1}^N \sum_{i=1}^m x_{i,\nu'}\big)},$$

for $\boldsymbol{x} = (x_{i,\nu'})_{1\le i\le m, 1\le \nu'\le N} \in {\mathbb{N}_0}^{m\times N}$ for each $\nu \in \{1,\dots,N\}$. Since $\sum_{\nu=1}^n \{I_{1,\nu}^{(\delta)}(\mathbf{0}^{(m,N)}) - 2I_{2,\nu}^{(\delta)}(\mathbf{0}^{(m,N)}) + 2I_{3,\nu}^{(\delta)}(\mathbf{0}^{(m,N)})\} < 0$, it is sufficient to show that $\sum_{\nu=1}^n \{I_{1,\nu}^{(\delta)}(\boldsymbol{x}) - 2I_{2,\nu}^{(\delta)}(\boldsymbol{x}) + 2I_{3,\nu}^{(\delta)}(\boldsymbol{x})\} \le 0$ for all $\boldsymbol{x} \in {\mathbb{N}_0}^{m\times N} \setminus \{\mathbf{0}^{(m,N)}\}$.

Fix $\boldsymbol{x} = (x_{i,\nu})_{1\le i\le m, 1\le \nu\le N} \in {\mathbb{N}_0}^{m\times N} \setminus \{\mathbf{0}^{(m,N)}\}$. For notational simplicity, let $z_\nu = \sum_{i=1}^m x_{i,\nu}$ for $\nu \in \{1,\dots,N\}$ and let $z = \sum_{\nu=1}^N z_\nu$. Then for all $\nu \in \{1,\dots,N\}$ such that $z_\nu \ne 0$, since

$$\delta(z) \le \frac{z+1}{z}\delta(z+1) \le \frac{z_\nu+1}{z_\nu}\delta(z+1),$$

we have

$$\begin{aligned} I_{3,\nu}^{(\delta)}(\boldsymbol{x}) &= \frac{z_\nu}{r+z_\nu-1}\frac{\delta(z)}{r+z_\nu-1+\delta(z)} \\ &\le \frac{z_\nu}{r+z_\nu-1}\frac{(z_\nu+1)\delta(z+1)}{z_\nu(r+z_\nu-1)+(z_\nu+1)\delta(z+1)} \le \frac{z_\nu+3}{r+z_\nu}\frac{\delta(z+1)}{r+z_\nu+\delta(z+1)}, \end{aligned}$$

where the second inequality follows from the assumption that $r \ge 5/2$. Therefore,

$$\sum_{\nu=1}^n \{I_{1,\nu}^{(\delta)}(\boldsymbol{x}) - 2I_{2,\nu}^{(\delta)}(\boldsymbol{x}) + 2I_{3,\nu}^{(\delta)}(\boldsymbol{x})\} \le I^{(\delta)}(\boldsymbol{x}),$$

where

$$I^{(\delta)}(\boldsymbol{x}) = \sum_{\nu=1}^n \Big(\frac{1}{r+z_\nu} \frac{\delta(z+1)}{\{r+z_\nu+\delta(z+1)\}^2} [z_\nu\{\delta(z+1)-2(m-3)\} - (m-6)\delta(z+1) - 2(m-3)r] \Big).$$

Suppose first that $\delta(z+1) \le 2(m-3)$. Then $I^{(\delta)}(\boldsymbol{x}) \le 0$ by assumption since $z+1 \ge 2$. On the other hand, if $\delta(z+1) > 2(m-3)$, then, by the covariance inequality,

$$\begin{aligned} I^{(\delta)}(\boldsymbol{x}) &\le \frac{1}{n}\Big[\sum_{\nu=1}^n \frac{1}{r+z_\nu}\frac{\delta(z+1)}{\{r+z_\nu+\delta(z+1)\}^2}\Big] \\ &\quad \times \Big[\Big(\sum_{\nu=1}^n z_\nu\Big)\{\delta(z+1)-2(m-3)\} - n\{(m-6)\delta(z+1)+2(m-3)r\}\Big] \\ &\le \frac{1}{n}\Big[\sum_{\nu=1}^n \frac{1}{r+z_\nu}\frac{\delta(z+1)}{\{r+z_\nu+\delta(z+1)\}^2}\Big] \\ &\quad \times [z\{\delta(z+1)-2(m-3)\} - n\{(m-6)\delta(z+1)+2(m-3)r\}]. \end{aligned}$$

The right-hand side of the above inequality is nonpositive by assumption. This completes the proof. □

Remark 3.6.1 In the above proof, we have shown that $I_n^{(\delta)}(\boldsymbol{x}) = \sum_{\nu=1}^n \{I_{1,\nu}^{(\delta)}(\boldsymbol{x}) - 2I_{2,\nu}^{(\delta)}(\boldsymbol{x}) + 2I_{3,\nu}^{(\delta)}(\boldsymbol{x})\} \le 0$ for all $\boldsymbol{x} \in {\mathbb{N}_0}^{m\times N}$. Conversely, this condition implies that $m \ge 2 + \delta(\infty)/2$ when $\lim_{z\to\infty} \delta(z) = \delta(\infty) \in (0,\infty)$ and $\lim_{z\to\infty} z\{\delta(z) - \delta(z+1)\} = 0$, which can be verified by considering $x^2 I_n^{(\delta)}(x\boldsymbol{j}^{(m)}{\boldsymbol{j}^{(N)}}^\top)$ for $x \in \mathbb{N}_0$ and taking the limit as $x \to \infty$. (The proof is omitted.) In particular, in the case of the empirical Bayes estimator $\hat{\boldsymbol{p}}^{\mathrm{EB}}$, the condition that $I_n^{(\delta^{\mathrm{EB}})}(\boldsymbol{x}) \le 0$ for all $\boldsymbol{x} \in {\mathbb{N}_0}^{m\times N}$ implies that $m \ge 5$, while it was assumed in Corollary 3.2.1 that $m \ge 7$.

Proof of Lemma 3.2.1. Let $\boldsymbol{x} = (x_{i,\nu})_{1\le i\le m,\, 1\le \nu\le N} \in {\mathbb{N}_0}^{m\times N}$ and fix $i \in \{1,\dots,m\}$ and $\nu \in \{1,\dots,N\}$. The posterior mean of $1/p_{i,\nu}$ with respect to the observation $\boldsymbol{X} = \boldsymbol{x}$ and the prior $\boldsymbol{p} \sim \prod_{\nu'=1}^N \mathrm{Dir}_m(\boldsymbol{p}_{\nu'}|a_0, \boldsymbol{j}^{(m)}) \propto \prod_{\nu'=1}^N {p_{0,\nu'}}^{a_0-1}$ is given by

$$\begin{aligned} E[1/p_{i,\nu}|\boldsymbol{X} = \boldsymbol{x}] &= \frac{\int_{{D_m}^N} (1/p_{i,\nu})\big\{\prod_{\nu'=1}^N \big({p_{0,\nu'}}^{r+a_0-1} \prod_{j=1}^m {p_{j,\nu'}}^{x_{j,\nu'}}\big)\big\} d\boldsymbol{p}}{\int_{{D_m}^N} \big\{\prod_{\nu'=1}^N \big({p_{0,\nu'}}^{r+a_0-1} \prod_{j=1}^m {p_{j,\nu'}}^{x_{j,\nu'}}\big)\big\} d\boldsymbol{p}} \\ &= \begin{cases} \dfrac{r + a_0 + x_{\cdot,\nu} + m - 1}{x_{i,\nu}}, & \text{if } x_{i,\nu} \ge 1, \\ \infty, & \text{if } x_{i,\nu} = 0, \end{cases} \end{aligned}$$

where $x_{\cdot,\nu} = \sum_{j=1}^m x_{j,\nu}$. Similarly, the posterior mean of $p_{i,\nu}$ is

$$E[p_{i,\nu}|\boldsymbol{X} = \boldsymbol{x}] = \frac{\int_{{D_m}^N} p_{i,\nu}\big\{\prod_{\nu'=1}^N \big({p_{0,\nu'}}^{r+a_0-1} \prod_{j=1}^m {p_{j,\nu'}}^{x_{j,\nu'}}\big)\big\} d\boldsymbol{p}}{\int_{{D_m}^N} \big\{\prod_{\nu'=1}^N \big({p_{0,\nu'}}^{r+a_0-1} \prod_{j=1}^m {p_{j,\nu'}}^{x_{j,\nu'}}\big)\big\} d\boldsymbol{p}} = \frac{x_{i,\nu} + 1}{r + a_0 + x_{\cdot,\nu} + m} < \infty.$$

Therefore, for any $d \in \mathbb{R}$, the posterior expectation of the loss $(d - p_{i,\nu})^2/p_{i,\nu}$ can be expressed as

$$E[(d - p_{i,\nu})^2/p_{i,\nu}|\boldsymbol{X} = \boldsymbol{x}] = d^2 E[1/p_{i,\nu}|\boldsymbol{X} = \boldsymbol{x}] - 2d + E[p_{i,\nu}|\boldsymbol{X} = \boldsymbol{x}],$$

which is minimized at

$$d = \frac{1}{E[1/p_{i,\nu}|\boldsymbol{X} = \boldsymbol{x}]} = \frac{x_{i,\nu}}{r + a_0 + x_{\cdot,\nu} + m - 1}.$$

Hence, $\hat{\boldsymbol{p}}^{(\delta^{(a_0)})} = (X_{i,\nu}/(r + a_0 + X_{\cdot,\nu} + m - 1))_{1\le i\le m,\, 1\le \nu\le N}$ is a Bayes solution. □

Proof of Lemma 3.3.1. Part (ii) follows immediately from part (i) since the posterior given $\boldsymbol{X} = (\boldsymbol{x}_1,\dots,\boldsymbol{x}_N)$ is proper for all $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N \in {\mathbb{N}_0}^m$ if and only if that given $\boldsymbol{X} = (\boldsymbol{0}^{(m)},\dots,\boldsymbol{0}^{(m)})$, namely $\boldsymbol{p} \sim \pi_{\alpha,\beta,g,r+a_0,\boldsymbol{a}}(\boldsymbol{p})$, is proper. For part (i), let $J^{(\alpha,\beta,g,a_0,\boldsymbol{a})} = \int_{{D_m}^N} \pi_{\alpha,\beta,g,a_0,\boldsymbol{a}}(\boldsymbol{p}) d\boldsymbol{p}$. Then we have

$$J^{(\alpha,\beta,g,a_0,\boldsymbol{a})} = \int_0^\infty t^{\alpha-1} e^{-\beta t} g(t) \{B_m(t + a_0, \boldsymbol{a})\}^N dt,$$

where

$$B_m(t + a_0, \boldsymbol{a}) = \int_{D_m} \Big({\mathring{p}_0}^{t+a_0-1} \prod_{i=1}^m {\mathring{p}_i}^{a_i-1}\Big) d\mathring{\boldsymbol{p}}_\nu = \begin{cases} \dfrac{\Gamma(t + a_0) \prod_{i=1}^m \Gamma(a_i)}{\Gamma(t + a_0 + a_\cdot)}, & \text{if } t + a_0 > 0, \\ \infty, & \text{if } t + a_0 \le 0, \end{cases}$$

for $t \in (0, \infty)$. Therefore, a necessary condition for the prior to be proper is that $a_0 \geq 0$. Suppose that $a_0 \geq 0$. Then

$$J^{(\alpha,\beta,g,a_0,\boldsymbol{a})} / \Big\{ \prod_{i=1}^{m} \Gamma(a_i) \Big\}^N = J_1^{(\alpha,\beta,g,a_0,\boldsymbol{a})} + J_2^{(\alpha,\beta,g,a_0,\boldsymbol{a})},$$

where

$$J_1^{(\alpha,\beta,g,a_0,\boldsymbol{a})} = \int_0^1 t^{\alpha-1} e^{-\beta t} g(t) \Big\{ \frac{\Gamma(t+a_0)}{\Gamma(t+a_0+a_\cdot)} \Big\}^N dt$$

and

$$J_2^{(\alpha,\beta,g,a_0,\boldsymbol{a})} = \int_1^\infty t^{\alpha-1} e^{-\beta t} g(t) \Big\{ \frac{\Gamma(t+a_0)}{\Gamma(t+a_0+a_\cdot)} \Big\}^N dt.$$

The term $J_1^{(\alpha,\beta,g,a_0,\boldsymbol{a})}$ is finite if and only if either $a_0 = 0$ and $\int_0^1 t^{\alpha-N-1} e^{-\beta t} g(t) dt < \infty$ or $a_0 > 0$ since $\lim_{t\to 0} \Gamma(t+a_0)/\Gamma(t+a_0+a_\cdot) = \Gamma(a_0)/\Gamma(a_0+a_\cdot)$ when $a_0 > 0$ and since $\Gamma(t+0)/\Gamma(t+0+a_\cdot) \sim t^{-1}/\Gamma(a_\cdot)$ as $t \to 0$ when $a_0 = 0$. The term $J_2^{(\alpha,\beta,g,a_0,\boldsymbol{a})}$ is finite if and only if $\int_1^\infty t^{\alpha-Na_\cdot-1} e^{-\beta t} g(t) dt < \infty$ since $\Gamma(t+a_0)/\Gamma(t+a_0+a_\cdot) \sim t^{-a_\cdot}$ as $t \to \infty$. This completes the proof of part (i). □

Proof of Lemma 3.3.2. Let $\boldsymbol{x} = (x_{i,\nu})_{1\leq i\leq m,\, 1\leq \nu\leq N} \in {\mathbb{N}_0}^{m\times N}$ and fix $i \in \{1, \ldots, m\}$ and $\nu \in \{1, \ldots, N\}$. Then it can be verified that the reciprocal of the posterior mean of $1/p_{i,\nu}$ with respect to the observation $\boldsymbol{X} = \boldsymbol{x}$ and the prior $\boldsymbol{p} \sim \pi_{\alpha,\beta,g,-m,\boldsymbol{j}^{(m)}}(\boldsymbol{p})$ is given by

$$\frac{1}{E[1/p_{i,\nu} | \boldsymbol{X} = \boldsymbol{x}]} = \begin{cases} \dfrac{x_{i,\nu}}{r + x_{\cdot,\nu} - 1 + \delta^{(\alpha,\beta,g)}(\boldsymbol{x}_\cdot)}, & \text{if } x_{i,\nu} \geq 1, \\ 0, & \text{if } x_{i,\nu} = 0, \end{cases}$$

where $x_{\cdot,\nu} = \sum_{j=1}^m x_{j,\nu}$ and $\boldsymbol{x}_\cdot = \big(\sum_{j=1}^m x_{j,1}, \ldots, \sum_{j=1}^m x_{j,N} \big)^\top$. Also, the posterior mean of $p_{i,\nu}$ is finite since $0 \leq p_{i,\nu} \leq 1$ and the posterior is proper. Therefore, for any $d \in \mathbb{R}$, the posterior expectation of the loss $(d - p_{i,\nu})^2/p_{i,\nu}$ can be expressed as

$$E[(d - p_{i,\nu})^2/p_{i,\nu} | \boldsymbol{X} = \boldsymbol{x}] = d^2 E[1/p_{i,\nu} | \boldsymbol{X} = \boldsymbol{x}] - 2d + E[p_{i,\nu} | \boldsymbol{X} = \boldsymbol{x}]$$

and is minimized at $d = 1/E[1/p_{i,\nu} | \boldsymbol{X} = \boldsymbol{x}]$. Hence, $\hat{\boldsymbol{p}}^{(\alpha,\beta,g)}$ is a Bayes solution. □

Proof of Proposition 3.3.1. Part (i) follows from the definition of the function $\delta^{(\alpha,\beta,g)}$. Let, for $t \in (0, \infty)$,

$$f_{\alpha,\beta,g}(t) = t^{\alpha-1} e^{-\beta t} g(t) \prod_{\nu'=1}^{N} \frac{\Gamma(t+r-m)}{\Gamma(t+r+z_{\nu'})}.$$

For part (ii), suppose that $\delta^{(\alpha,\beta,g)}(\boldsymbol{z}) < \infty$. Then by the covariance inequality we have

$$\delta^{(\alpha,\beta,g)}(\boldsymbol{z}) / \delta^{(\alpha,\beta,g)}(\boldsymbol{z} + \boldsymbol{e}_\nu^{(N)}) = \frac{\int_0^\infty t f_{\alpha,\beta,g}(t) dt}{\int_0^\infty f_{\alpha,\beta,g}(t) dt} \Big/ \frac{\int_0^\infty \frac{t}{t+r+z_\nu} f_{\alpha,\beta,g}(t) dt / \int_0^\infty f_{\alpha,\beta,g}(t) dt}{\int_0^\infty \frac{1}{t+r+z_\nu} f_{\alpha,\beta,g}(t) dt / \int_0^\infty f_{\alpha,\beta,g}(t) dt} \geq 1.$$

For part (iii), let $k \in \mathbb{N}$. Then

$$\delta^{(\alpha,\beta,g)}(\boldsymbol{z}+k\boldsymbol{e}_\nu^{(N)}) = \frac{\int_0^\infty \frac{t}{(t+r+z_\nu)\cdots(t+r+z_\nu+k-1)} f_{\alpha,\beta,g}(t)dt}{\int_0^\infty \frac{1}{(t+r+z_\nu)\cdots(t+r+z_\nu+k-1)} f_{\alpha,\beta,g}(t)dt}.$$

Fix $\varepsilon > 0$. Then it follows that for each $l \in \{0,1\}$,

$$\begin{aligned}
&\Big| \frac{\int_0^\infty \frac{t^l}{(t+r+z_\nu)\cdots(t+r+z_\nu+k-1)} f_{\alpha,\beta,g}(t)dt}{\int_0^\varepsilon \frac{t^l}{(t+r+z_\nu)\cdots(t+r+z_\nu+k-1)} f_{\alpha,\beta,g}(t)dt} - 1 \Big| \le \frac{\int_\varepsilon^\infty \frac{t^l}{(t+r+z_\nu)\cdots(t+r+z_\nu+k-1)} f_{\alpha,\beta,g}(t)dt}{\int_0^{\varepsilon/2} \frac{t^l}{(t+r+z_\nu)\cdots(t+r+z_\nu+k-1)} f_{\alpha,\beta,g}(t)dt} \\
&\le \frac{(\varepsilon/2+r+z_\nu)\cdots(\varepsilon/2+r+z_\nu+k-1)}{(\varepsilon+r+z_\nu)\cdots(\varepsilon+r+z_\nu+k-1)} \frac{\int_\varepsilon^\infty t^l f_{\alpha,\beta,g}(t)dt}{\int_0^{\varepsilon/2} t^l f_{\alpha,\beta,g}(t)dt} \\
&= \frac{\Gamma(\varepsilon/2+r+z_\nu+k)/\Gamma(\varepsilon/2+r+z_\nu)}{\Gamma(\varepsilon+r+z_\nu+k)/\Gamma(\varepsilon+r+z_\nu)} \frac{\int_\varepsilon^\infty t^l f_{\alpha,\beta,g}(t)dt}{\int_0^{\varepsilon/2} t^l f_{\alpha,\beta,g}(t)dt},
\end{aligned}$$

the right-hand side of which converges to zero as $k \to \infty$ since $\Gamma(\varepsilon/2+r+z_\nu+k)/\Gamma(\varepsilon+r+z_\nu+k) \sim 1/(\varepsilon+r+z_\nu+k)^{\varepsilon/2}$ as $k \to \infty$. Therefore, as $k \to \infty$,

$$\delta^{(\alpha,\beta,g)}(\boldsymbol{z}+k\boldsymbol{e}_\nu^{(N)}) \sim \frac{\int_0^\varepsilon \frac{t}{(t+r+z_\nu)\cdots(t+r+z_\nu+k-1)} f_{\alpha,\beta,g}(t)dt}{\int_0^\varepsilon \frac{1}{(t+r+z_\nu)\cdots(t+r+z_\nu+k-1)} f_{\alpha,\beta,g}(t)dt} \le \varepsilon.$$

Since ε is arbitrarily chosen, we conclude that $\lim_{\mathbb{N}\ni k\to\infty} \delta^{(\alpha,\beta,g)}(\boldsymbol{z}+k\boldsymbol{e}_\nu^{(N)}) = 0$. For part (iv), let $k \in \mathbb{N}\setminus\{1\}$. Then

$$\begin{aligned}
\frac{\delta^{(\alpha,\beta,g)}(\boldsymbol{z}+k\boldsymbol{j}^{(N)})}{1/\log k} &= \frac{\int_0^\infty (\log k) t^\alpha e^{-\beta t} g(t) \big\{ \prod_{\nu=1}^N \frac{\Gamma(t+r-m)}{\Gamma(t+r+z_\nu+k)} \big\} dt}{\int_0^\infty t^{\alpha-1} e^{-\beta t} g(t) \big\{ \prod_{\nu=1}^N \frac{\Gamma(t+r-m)}{\Gamma(t+r+z_\nu+k)} \big\} dt} \\
&= \frac{\int_0^\infty u^\alpha e^{-\beta u/\log k} g\big(\frac{u}{\log k}\big) \big\{ \prod_{\nu=1}^N \frac{\Gamma(u/\log k+r-m)\Gamma(r+z_\nu+k)}{\Gamma(u/\log k+r+z_\nu+k)\Gamma(r+z_\nu)} \big\} du}{\int_0^\infty u^{\alpha-1} e^{-\beta u/\log k} g\big(\frac{u}{\log k}\big) \big\{ \prod_{\nu=1}^N \frac{\Gamma(u/\log k+r-m)\Gamma(r+z_\nu+k)}{\Gamma(u/\log k+r+z_\nu+k)\Gamma(r+z_\nu)} \big\} du}.
\end{aligned}$$

Now for each $l \in \{0,1\}$ and all $u \in (0,\infty)$, we have that

$$\begin{aligned}
&u^{\alpha+l-1} e^{-\beta u/\log k} g\Big(\frac{u}{\log k}\Big) \prod_{\nu=1}^N \frac{\Gamma(u/\log k+r-m)\Gamma(r+z_\nu+k)}{\Gamma(u/\log k+r+z_\nu+k)\Gamma(r+z_\nu)} \\
&\le \frac{\big[\sup_{t\in(0,\infty)} \big\{ g(t) \prod_{\nu=1}^N \frac{\Gamma(t+r-m)}{\Gamma(t+r+z_\nu)} \big\}\big] u^{\alpha+l-1}}{\prod_{\nu=1}^N \big\{ \big(1+\frac{u/\log k}{r+z_\nu}\big)\cdots\big(1+\frac{u/\log k}{r+z_\nu+k-1}\big) \big\}} \\
&\le \frac{\big[\sup_{t\in(0,\infty)} \big\{ g(t) \prod_{\nu=1}^N \frac{\Gamma(t+r-m)}{\Gamma(t+r+z_\nu)} \big\}\big] u^{\alpha+l-1}}{\prod_{\nu=1}^N \big\{ 1+u\big(\log\frac{r+z_\nu+k}{r+z_\nu}\big)/\log k \big\}} \\
&\le \frac{\big[\sup_{t\in(0,\infty)} \big\{ g(t) \prod_{\nu=1}^N \frac{\Gamma(t+r-m)}{\Gamma(t+r+z_\nu)} \big\}\big] u^{\alpha+l-1}}{\prod_{\nu=1}^N \big[1+u \inf_{k'\in\mathbb{N}\setminus\{1\}} \big\{ \big(\log\frac{r+z_\nu+k'}{r+z_\nu}\big)/\log k' \big\} \big]},
\end{aligned}$$

where the second inequality follows since

$$\Big(1+\frac{u/\log k}{r+z_\nu}\Big)\times\cdots\times\Big(1+\frac{u/\log k}{r+z_\nu+k-1}\Big)\geq 1+\frac{u}{\log k}\Big(\frac{1}{r+z_\nu}+\cdots+\frac{1}{r+z_\nu+k-1}\Big)$$
$$\geq 1+\frac{u}{\log k}\log\frac{r+z_\nu+k}{r+z_\nu}$$

for every $\nu\in\{1,\dots,N\}$, and that

$$\lim_{\mathbb{N}\setminus\{1\}\ni k\to\infty}\Big\{u^{\alpha+l-1}e^{-\beta u/\log k}g\Big(\frac{u}{\log k}\Big)\prod_{\nu=1}^{N}\frac{\Gamma(u/\log k+r-m)\Gamma(r+z_\nu+k)}{\Gamma(u/\log k+r+z_\nu+k)\Gamma(r+z_\nu)}\Big\}$$
$$=g(0)\Big\{\prod_{\nu=1}^{N}\frac{\Gamma(r-m)}{\Gamma(r+z_\nu)}\Big\}u^{\alpha+l-1}\prod_{\nu=1}^{N}\lim_{\mathbb{N}\setminus\{1\}\ni k\to\infty}\frac{\Gamma(r+z_\nu+k)}{\Gamma(u/\log k+r+z_\nu+k)}$$
$$=g(0)\Big\{\prod_{\nu=1}^{N}\frac{\Gamma(r-m)}{\Gamma(r+z_\nu)}\Big\}u^{\alpha+l-1}e^{-Nu}.$$

Thus,

$$\lim_{\mathbb{N}\setminus\{1\}\ni k\to\infty}\frac{\delta^{(\alpha,\beta,g)}(\boldsymbol{z}+k\boldsymbol{j}^{(N)})}{1/\log k}=\frac{\int_0^\infty u^\alpha e^{-Nu}du}{\int_0^\infty u^{\alpha-1}e^{-Nu}du}=\frac{\alpha}{N},$$

and the result follows. □

Proof of Theorem 3.3.1. First, note that $r\geq m\geq 3$ by assumption. Let $\Delta_n^{(\alpha,\beta,g)}=E[L_n(\hat{\boldsymbol{p}}^{(\alpha,\beta,g)},\boldsymbol{p})]-E[L_n(\hat{\boldsymbol{p}}^{\mathrm{U}},\boldsymbol{p})]$. For $\nu\in\{1,\dots,N\}$, let

$$\phi_\nu^{(\alpha,\beta,g)}(\boldsymbol{X})=\begin{cases}\dfrac{K(\alpha+1,\beta,g,r-m,\boldsymbol{X}_{\cdot}+m\boldsymbol{j}^{(N)})}{K(\alpha,\beta,g,r-m,\boldsymbol{X}_{\cdot}+m\boldsymbol{j}^{(N)}-\boldsymbol{e}_\nu^{(N)})}, & \text{if } X_{\cdot,\nu}\geq 1,\\ 0, & \text{if } X_{\cdot,\nu}=0,\end{cases}$$

so that for every $i\in\{1,\dots,m\}$,

$$\hat{p}_{i,\nu}^{(\alpha,\beta,g)}=\hat{p}_{i,\nu}^{\mathrm{U}}-\hat{p}_{i,\nu}^{\mathrm{U}}\phi_\nu^{(\alpha,\beta,g)}(\boldsymbol{X}).$$

Then, by Lemma 3.6.1, we have

$$\Delta_n^{(\alpha,\beta,g)}=E\Big[\sum_{\nu=1}^{n}\sum_{i=1}^{m}\Big(\frac{1}{p_{i,\nu}}[(\hat{p}_{i,\nu}^{\mathrm{U}})^2\{\phi_\nu^{(\alpha,\beta,g)}(\boldsymbol{X})\}^2-2(\hat{p}_{i,\nu}^{\mathrm{U}})^2\phi_\nu^{(\alpha,\beta,g)}(\boldsymbol{X})]+2\hat{p}_{i,\nu}^{\mathrm{U}}\phi_\nu^{(\alpha,\beta,g)}(\boldsymbol{X})\Big)\Big]$$
$$=E\Big[\sum_{\nu=1}^{n}\sum_{i=1}^{m}\Big[\frac{X_{i,\nu}+1}{r+X_{\cdot,\nu}}\{\phi_\nu^{(\alpha,\beta,g)}(\boldsymbol{X}+\boldsymbol{e}_{i,\nu}^{(m,N)})\}^2$$
$$-2\frac{X_{i,\nu}+1}{r+X_{\cdot,\nu}}\phi_\nu^{(\alpha,\beta,g)}(\boldsymbol{X}+\boldsymbol{e}_{i,\nu}^{(m,N)})+2\hat{p}_{i,\nu}^{\mathrm{U}}\phi_\nu^{(\alpha,\beta,g)}(\boldsymbol{X})\Big]\Big]$$
$$=E[I_{1,n}^{(\alpha,\beta,g)}(\boldsymbol{X})-2I_{2,n}^{(\alpha,\beta,g)}(\boldsymbol{X})+2I_{3,n}^{(\alpha,\beta,g)}(\boldsymbol{X})],$$

where

$$I_{1,n}^{(\alpha,\beta,g)}(\boldsymbol{x}) = \sum_{\nu=1}^{n} \frac{x_{\cdot,\nu}+m}{r+x_{\cdot,\nu}} \Big\{ \frac{K(\alpha+1,\beta,g,r-m,\boldsymbol{x}_{\cdot}+m\boldsymbol{j}^{(N)}+\boldsymbol{e}_{\nu}^{(N)})}{K(\alpha,\beta,g,r-m,\boldsymbol{x}_{\cdot}+m\boldsymbol{j}^{(N)})} \Big\}^2,$$

$$I_{2,n}^{(\alpha,\beta,g)}(\boldsymbol{x}) = \sum_{\nu=1}^{n} \frac{x_{\cdot,\nu}+m}{r+x_{\cdot,\nu}} \frac{K(\alpha+1,\beta,g,r-m,\boldsymbol{x}_{\cdot}+m\boldsymbol{j}^{(N)}+\boldsymbol{e}_{\nu}^{(N)})}{K(\alpha,\beta,g,r-m,\boldsymbol{x}_{\cdot}+m\boldsymbol{j}^{(N)})},$$

$$I_{3,n}^{(\alpha,\beta,g)}(\boldsymbol{x}) = \sum_{\nu=1}^{n} \frac{x_{\cdot,\nu}}{r+x_{\cdot,\nu}-1} \phi_{\nu}^{(\alpha,\beta,g)}(\boldsymbol{x}),$$

and $\boldsymbol{x}_{\cdot} = (x_{\cdot,1},\ldots,x_{\cdot,N})^{\top} = \big(\sum_{i=1}^{m} x_{i,1},\ldots,\sum_{i=1}^{m} x_{i,N}\big)^{\top}$ for $\boldsymbol{x} = (x_{i,\nu})_{1\le i\le m,\,1\le\nu\le N} \in {\mathbb{N}_0}^{m\times N}$. Since $I_{1,n}^{(\alpha,\beta,g)}(\boldsymbol{0}^{(m,N)}) - 2I_{2,n}^{(\alpha,\beta,g)}(\boldsymbol{0}^{(m,N)}) + 2I_{3,n}^{(\alpha,\beta,g)}(\boldsymbol{0}^{(m,N)}) < 0$, it is sufficient to show that $I_{1,n}^{(\alpha,\beta,g)}(\boldsymbol{x}) - 2I_{2,n}^{(\alpha,\beta,g)}(\boldsymbol{x}) + 2I_{3,n}^{(\alpha,\beta,g)}(\boldsymbol{x}) \le 0$ for all $\boldsymbol{x} \in {\mathbb{N}_0}^{m\times N} \setminus \{\boldsymbol{0}^{(m,N)}\}$.

Fix $\boldsymbol{x} = (x_{i,\nu})_{1\le i\le m,\,1\le\nu\le N} \in {\mathbb{N}_0}^{m\times N}\setminus\{\boldsymbol{0}^{(m,N)}\}$. For notational simplicity, let $z_{\nu} = \sum_{i=1}^{m} x_{i,\nu}$ for $\nu \in \{1,\ldots,N\}$ and let $\boldsymbol{z} = (z_1,\ldots,z_N)^{\top}$ and $z = \sum_{\nu=1}^{N} z_{\nu}$. In addition, we use the abbreviated notation

$$I_1 = I_{1,n}^{(\alpha,\beta,g)}(\boldsymbol{x}), \quad I_2 = I_{2,n}^{(\alpha,\beta,g)}(\boldsymbol{x}), \quad I_3 = I_{3,n}^{(\alpha,\beta,g)}(\boldsymbol{x}), \quad I = I_1 - 2I_2 + 2I_3,$$

$$H(l) = \frac{K(\alpha+l,\beta,g,r-m,\boldsymbol{x}_{\cdot}+m\boldsymbol{j}^{(N)})}{K(\alpha,\beta,g,r-m,\boldsymbol{x}_{\cdot}+m\boldsymbol{j}^{(N)})}, \quad H(l,\pm\nu) = \frac{K(\alpha+l,\beta,g,r-m,\boldsymbol{x}_{\cdot}+m\boldsymbol{j}^{(N)}\pm\boldsymbol{e}_{\nu}^{(N)})}{K(\alpha,\beta,g,r-m,\boldsymbol{x}_{\cdot}+m\boldsymbol{j}^{(N)})},$$

for $l \in \{0,1,2\}$ and $\nu \in \{1,\ldots,N\}$. Also, let, for $t \in (0,\infty)$,

$$f_{\alpha,\beta,g}(t) = t^{\alpha-1}e^{-\beta t}g(t)\prod_{\nu=1}^{N}\frac{\Gamma(t+r-m)}{\Gamma(t+r+z_{\nu})}$$

so that, for example, $K(\alpha,\beta,g,r-m,\boldsymbol{x}_{\cdot}+m\boldsymbol{j}^{(N)}) = \int_0^{\infty} f_{\alpha,\beta,g}(t)dt$ and let, for $t \in (0,\infty)$,

$$f_{\alpha,\beta,g}^{*}(t) = \frac{f_{\alpha,\beta,g}(t)}{K(\alpha,\beta,g,r-m,\boldsymbol{x}_{\cdot}+m\boldsymbol{j}^{(N)})} = \frac{f_{\alpha,\beta,g}(t)}{\int_0^{\infty} f_{\alpha,\beta,g}(t')dt'}.$$

For all $\nu \in \{1,\ldots,N\}$ such that $z_{\nu} \neq 0$, we have that

$$\begin{aligned}
\phi_{\nu}^{(\alpha,\beta,g)}(\boldsymbol{x}) &= \frac{H(1)}{H(0,-\nu)} = \frac{\int_0^{\infty} t f_{\alpha,\beta,g}(t)dt}{\int_0^{\infty}(t+r+z_{\nu}-1)f_{\alpha,\beta,g}(t)dt} \\
&= \frac{\int_0^{\infty} t f_{\alpha,\beta,g}(t)dt}{\int_0^{\infty} f_{\alpha,\beta,g}(t)dt} \frac{\int_0^{\infty}(t+r+z_{\nu}-1)f_{\alpha,\beta,g}(t)dt - \int_0^{\infty} t f_{\alpha,\beta,g}(t)dt}{(r+z_{\nu}-1)\int_0^{\infty}(t+r+z_{\nu}-1)f_{\alpha,\beta,g}(t)dt} \\
&= \frac{1}{r+z_{\nu}-1}\Big\{H(1) - \frac{H(1)}{H(0,-\nu)}H(1)\Big\}
\end{aligned}$$

and that

$$\begin{aligned}
\frac{H(1)}{H(0,-\nu)}/H(1,\nu) &= \frac{\int_0^{\infty} t f_{\alpha,\beta,g}^{*}(t)dt}{\int_0^{\infty}(t+r+z_{\nu}-1)f_{\alpha,\beta,g}^{*}(t)dt\int_0^{\infty}\frac{t}{t+r+z_{\nu}}f_{\alpha,\beta,g}^{*}(t)dt} \\
&\ge \frac{\int_0^{\infty} t f_{\alpha,\beta,g}^{*}(t)dt}{\int_0^{\infty} t\frac{t+r+z_{\nu}-1}{t+r+z_{\nu}}f_{\alpha,\beta,g}^{*}(t)dt} \ge 1
\end{aligned}$$

by the covariance inequality. Therefore,

$$I_3 \le \sum_{\nu=1}^{n} \frac{z_\nu}{(r+z_\nu-1)^2}\{H(1)-H(1,\nu)H(1)\} \le \sum_{\nu=1}^{n} \frac{z_\nu+2}{(r+z_\nu)^2}\{H(1)-H(1,\nu)H(1)\}$$
$$= \sum_{\nu=1}^{n} \frac{z_\nu+2}{(r+z_\nu)^2}H(1) - \sum_{\nu=1}^{n} \frac{z_\nu+2}{(r+z_\nu)^2}H(1,\nu)H(1). \tag{3.6.1}$$

Since

$$H(1,\nu) = \int_0^\infty \frac{t}{t+r+z_\nu} f^*_{\alpha,\beta,g}(t)dt$$
$$= \int_0^\infty \frac{t}{r+z_\nu}\Big(1-\frac{t}{t+r+z_\nu}\Big) f^*_{\alpha,\beta,g}(t)dt = \frac{1}{r+z_\nu}\{H(1)-H(2,\nu)\},$$

for all $\nu \in \{1,\dots,N\}$, it follows that

$$I_2 = \sum_{\nu=1}^{n} \frac{z_\nu+m}{r+z_\nu}H(1,\nu) = \sum_{\nu=1}^{n} \frac{z_\nu+m}{(r+z_\nu)^2}H(1) - \sum_{\nu=1}^{n} \frac{z_\nu+m}{(r+z_\nu)^2}H(2,\nu). \tag{3.6.2}$$

Now, by the covariance inequality,

$$\sum_{\nu=1}^{n} \frac{z_\nu+m}{(r+z_\nu)^2}H(2,\nu) \le \frac{1}{n}\Big\{\sum_{\nu=1}^{n} \frac{1}{(r+z_\nu)^2}\Big\}\sum_{\nu=1}^{n}(z_\nu+m)H(2,\nu). \tag{3.6.3}$$

By integration by parts,

$$\infty > (\alpha+1)\int_0^\infty t f_{\alpha,\beta,g}(t)dt = \int_0^\infty (\alpha+1)t^\alpha e^{-\beta t}g(t)\Big\{\prod_{\nu'=1}^{N}\frac{\Gamma(t+r-m)}{\Gamma(t+r+z_{\nu'})}\Big\}dt$$
$$= \lim_{\varepsilon\to 0}\Big(\Big[t^{\alpha+1}e^{-\beta t}g(t)\Big\{\prod_{\nu'=1}^{N}\frac{\Gamma(t+r-m)}{\Gamma(t+r+z_{\nu'})}\Big\}\Big]_\varepsilon^{1/\varepsilon}$$
$$- \int_\varepsilon^{1/\varepsilon} t^{\alpha+1}\Big[\frac{\partial}{\partial t}\Big\{e^{-\beta t}g(t)\prod_{\nu'=1}^{N}\frac{\Gamma(t+r-m)}{\Gamma(t+r+z_{\nu'})}\Big\}\Big]dt\Big)$$
$$= \int_0^\infty t^{\alpha+1}e^{-\beta t}g(t)\Big\{\prod_{\nu'=1}^{N}\frac{\Gamma(t+r-m)}{\Gamma(t+r+z_{\nu'})}\Big\}\Big\{\beta+\frac{-g'(t)}{g(t)}\Big\}dt$$
$$+ \sum_{\nu=1}^{N}\sum_{k=1}^{z_\nu+m}\int_0^\infty t^{\alpha+1}e^{-\beta t}g(t)\Big\{\prod_{\nu'=1}^{N}\frac{\Gamma(t+r-m)}{\Gamma(t+r+z_{\nu'})}\Big\}\frac{1}{t+r-m+k-1}dt,$$

where the last equality follows from the assumptions of the theorem since $\Gamma(t) \sim t^{-1}$ as $t \to 0$ while $\prod_{\nu'=1}^{N}\{\Gamma(t+r-m)/\Gamma(t+r+z_{\nu'})\} \sim t^{-z-Nm}$ as $t \to \infty$ and since $-g'(t) \ge 0$ and

$|t/(t+r-m+k-1)| \le 1$ for all $t \in (0,\infty)$ and $k \in \{1,2,\dots\}$. Therefore,

$$\begin{aligned}
&\sum_{\nu=1}^{n}(z_\nu+m)H(2,\nu)\\
&=\sum_{\nu=1}^{n}\sum_{k=1}^{z_\nu+m}\int_0^\infty t^{\alpha+1}e^{-\beta t}g(t)\Big\{\prod_{\nu'=1}^{N}\frac{\Gamma(t+r-m)}{\Gamma(t+r+z_{\nu'})}\Big\}\frac{1}{t+r+z_\nu}dt\Big/\int_0^\infty f_{\alpha,\beta,g}(t)dt\\
&\le\sum_{\nu=1}^{N}\sum_{k=1}^{z_\nu+m}\int_0^\infty t^{\alpha+1}e^{-\beta t}g(t)\Big\{\prod_{\nu'=1}^{N}\frac{\Gamma(t+r-m)}{\Gamma(t+r+z_{\nu'})}\Big\}\frac{1}{t+r-m+k-1}dt\Big/\int_0^\infty f_{\alpha,\beta,g}(t)dt\\
&\le\Big[(\alpha+1)\int_0^\infty tf_{\alpha,\beta,g}(t)dt-\beta\int_0^\infty t^{\alpha+1}e^{-\beta t}g(t)\Big\{\prod_{\nu'=1}^{N}\frac{\Gamma(t+r-m)}{\Gamma(t+r+z_{\nu'})}\Big\}dt\Big]\Big/\int_0^\infty f_{\alpha,\beta,g}(t)dt\\
&=(\alpha+1)H(1)-\beta H(2)\le(\alpha+1)H(1)-\beta H(1)H(1),
\end{aligned}\tag{3.6.4}$$

where the last inequality follows since, by the covariance inequality,

$$H(2)=\int_0^\infty t^2f^*_{\alpha,\beta,g}(t)dt\ge\Big\{\int_0^\infty tf^*_{\alpha,\beta,g}(t)dt\Big\}^2=\{H(1)\}^2.$$

Combining (3.6.2), (3.6.3), and (3.6.4) gives

$$\begin{aligned}
I_2&\ge\sum_{\nu=1}^{n}\frac{z_\nu+m}{(r+z_\nu)^2}H(1)-\frac{1}{n}\Big\{\sum_{\nu=1}^{n}\frac{1}{(r+z_\nu)^2}\Big\}\sum_{\nu=1}^{n}(z_\nu+m)H(2,\nu)\\
&\ge\sum_{\nu=1}^{n}\frac{z_\nu+m-(\alpha+1)/n}{(r+z_\nu)^2}H(1)+\frac{\beta}{n}\sum_{\nu=1}^{n}\frac{1}{(r+z_\nu)^2}H(1)H(1)\\
&\ge\sum_{\nu=1}^{n}\frac{z_\nu+m-(\alpha+1)/n}{(r+z_\nu)^2}H(1)+\frac{\beta}{n}\sum_{\nu=1}^{n}\frac{r+z_\nu}{(r+z_\nu)^2}H(1,\nu)H(1)
\end{aligned}\tag{3.6.5}$$

since, for all $\nu\in\{1,\dots,N\}$,

$$H(1)=\int_0^\infty tf^*_{\alpha,\beta,g}(t)dt\ge\int_0^\infty\frac{r+z_\nu}{t+r+z_\nu}tf^*_{\alpha,\beta,g}(t)dt=(r+z_\nu)H(1,\nu).$$

Finally,

$$I_1=\sum_{\nu=1}^{n}\frac{z_\nu+m}{r+z_\nu}\{H(1,\nu)\}^2\le\sum_{\nu=1}^{n}\frac{z_\nu+m}{(r+z_\nu)^2}H(1,\nu)H(1).\tag{3.6.6}$$

Also, for all $\nu\in\{1,\dots,N\}$,

$$H(1,\nu)=\int_0^\infty\frac{t}{t+r+z_\nu}f^*_{\alpha,\beta,g}(t)dt\le\int_0^\infty f^*_{\alpha,\beta,g}(t)dt=1.$$

Hence, combining (3.6.1), (3.6.5), and (3.6.6), we obtain

$$\begin{aligned}
I&\le\sum_{\nu=1}^{n}\frac{-z_\nu+m-4}{(r+z_\nu)^2}H(1,\nu)H(1)-2\sum_{\nu=1}^{n}\frac{m-2-(\alpha+1)/n}{(r+z_\nu)^2}H(1)-2\frac{\beta}{n}\sum_{\nu=1}^{n}\frac{r+z_\nu}{(r+z_\nu)^2}H(1,\nu)H(1)\\
&\le\sum_{\nu=1}^{n}\frac{-z_\nu-m+2(\alpha+1)/n-2(\beta/n)(r+z_\nu)}{(r+z_\nu)^2}H(1,\nu)H(1)\le 0,
\end{aligned}$$

where the second and third inequalities follow from (3.3.8), and this completes the proof. □

Proof of Lemma 3.3.3. Let $\mathring{\boldsymbol{X}} = (\mathring{X}_1, \dots, \mathring{X}_m)^\top \sim \mathrm{NM}_m(r, \mathring{\boldsymbol{p}})$. Then the square root of the determinant of the information matrix corresponding to this distribution is

$$\begin{aligned}
&\sqrt{\Big|\Big(E\Big[-\frac{\partial^2}{\partial \mathring{p}_i \partial \mathring{p}_j}\log \mathrm{NM}_m(\mathring{\boldsymbol{X}}|r,\mathring{p})\Big]\Big)_{1\le i,j\le m}\Big|} = \sqrt{\Big|\Big(E\Big[\frac{r}{\mathring{p}_0^2}+\delta_{i,j}^{(m)}\frac{\mathring{X}_i}{\mathring{p}_i^2}\Big]\Big)_{1\le i,j\le m}\Big|} \\
&= \sqrt{|\boldsymbol{D}(\mathring{\boldsymbol{p}}) + (r/\mathring{p}_0^2)\boldsymbol{j}^{(m)}\boldsymbol{j}^{(m)\top}|} = \sqrt{|\boldsymbol{D}(\mathring{\boldsymbol{p}})|[1+(r/\mathring{p}_0^2)\boldsymbol{j}^{(m)\top}\{\boldsymbol{D}(\mathring{\boldsymbol{p}})\}^{-1}\boldsymbol{j}^{(m)}]} \qquad (3.6.7)\\
&= \sqrt{\frac{r^m}{\mathring{p}_0^m}\Big(\prod_{i=1}^m \frac{1}{\mathring{p}_i}\Big)\Big(1+\frac{\mathring{p}_\cdot}{\mathring{p}_0}\Big)} \propto \mathrm{Dir}_m\Big(\mathring{\boldsymbol{p}}\Big|\frac{1-m}{2}, \frac{1}{2}\boldsymbol{j}^{(m)}\Big),
\end{aligned}$$

where $\boldsymbol{D}(\boldsymbol{p}) = (r/\mathring{p}_0)\mathrm{diag}\,(1/\mathring{p}_1, \dots, 1/\mathring{p}_m)$. This is the desired result. □

Proof of Lemma 3.3.4. Let $\boldsymbol{x} = (x_{i,\nu})_{1\le i\le m,\,1\le \nu\le N} \in {\mathbb{N}_0}^{m\times N}$ and fix $i \in \{1,\dots,m\}$ and $\nu \in \{1,\dots,N\}$. Since the prior density is strictly positive and the posterior is proper, the posterior mean of $p_{i,\nu}$ with respect to the observation $\boldsymbol{X} = \boldsymbol{x}$, denoted $E[p_{i,\nu}|\boldsymbol{X}=\boldsymbol{x}]$, satisfies $E[p_{i,\nu}|\boldsymbol{X}=\boldsymbol{x}] \in (0,\infty)$. Also, $E[p_{i,\nu}\log(1/p_{i,\nu})|\boldsymbol{X}=\boldsymbol{x}] \in (0,\infty)$. Therefore, for any $\tilde{d} \in (0,\infty)$, the posterior expectation of the loss $\tilde{d} - p_{i,\nu} - p_{i,\nu}\log(\tilde{d}/p_{i,\nu})$ can be expressed as

$$\begin{aligned}
&E[\tilde{d} - p_{i,\nu} - p_{i,\nu}\log(\tilde{d}/p_{i,\nu})|\boldsymbol{X}=\boldsymbol{x}] \\
&= \tilde{d} - E[p_{i,\nu}|\boldsymbol{X}=\boldsymbol{x}]\log\tilde{d} - E[p_{i,\nu}|\boldsymbol{X}=\boldsymbol{x}] - E[p_{i,\nu}\log(1/p_{i,\nu})|\boldsymbol{X}=\boldsymbol{x}]
\end{aligned}$$

and thus is minimized at $\tilde{d} = E[p_{i,\nu}|\boldsymbol{X}=\boldsymbol{x}]$, which yields the desired result. □

Proof of Proposition 3.3.2. The proof is similar to that of Proposition 3.3.1. Part (i) follows from the definition. Let

$$f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t) = t^{\alpha-1}e^{-\beta t}g(t)\prod_{\nu''=1}^N \frac{\Gamma(t+r+a_0)}{\Gamma(t+r+a_0+z_{\nu''}+a_\cdot)}$$

for $t \in (0,\infty)$ so that $K(\alpha,\beta,g,r+a_0,\boldsymbol{z}+a_\cdot\boldsymbol{j}^{(N)}) = \int_0^\infty f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)dt$. Then part (ii) follows since

$$\frac{\delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z})}{\delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z}+\boldsymbol{e}_{\nu'}^{(N)})} = \frac{\dfrac{\int_0^\infty t\frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot}dt}{\int_0^\infty \frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot}dt}}{\dfrac{\int_0^\infty \frac{t}{t+r+a_0+z_{\nu'}+a_\cdot+\delta_{\nu,\nu'}^{(N)}}\frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot}dt \Big/ \int_0^\infty \frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot}dt}{\int_0^\infty \frac{1}{t+r+a_0+z_{\nu'}+a_\cdot+\delta_{\nu,\nu'}^{(N)}}\frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot}dt \Big/ \int_0^\infty \frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot}dt}} \ge 1$$

by the covariance inequality. For part (iii), let $k \in \mathbb{N}$. Then

$$\delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z}+k\boldsymbol{e}_{\nu'}^{(N)}) = \frac{\int_0^\infty \frac{t}{(t+r+a_0+z_{\nu'}+a_\cdot+\delta_{\nu,\nu'}^{(N)})\cdots(t+r+a_0+z_{\nu'}+a_\cdot+\delta_{\nu,\nu'}^{(N)}+k-1)}\frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot}dt}{\int_0^\infty \frac{1}{(t+r+a_0+z_{\nu'}+a_\cdot+\delta_{\nu,\nu'}^{(N)})\cdots(t+r+a_0+z_{\nu'}+a_\cdot+\delta_{\nu,\nu'}^{(N)}+k-1)}\frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot}dt}.$$

Fix $\varepsilon > 0$. Then it follows that for each $l \in \{0, 1\}$,

$$\begin{aligned}
&\Big| \frac{\int_0^\infty \frac{t^l}{(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'})\cdots(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k-1)} \frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot} dt}{\int_0^\varepsilon \frac{t^l}{(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'})\cdots(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k-1)} \frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot} dt} - 1 \Big| \\
&\le \frac{\int_\varepsilon^\infty \frac{t^l}{(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'})\cdots(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k-1)} \frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot} dt}{\int_0^{\varepsilon/2} \frac{t^l}{(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'})\cdots(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k-1)} \frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot} dt} \\
&\le \frac{(\varepsilon/2+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'})\cdots(\varepsilon/2+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k-1)}{(\varepsilon+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'})\cdots(\varepsilon+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k-1)} \frac{\int_\varepsilon^\infty t^l \frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot} dt}{\int_0^{\varepsilon/2} t^l \frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot} dt} \\
&= \frac{\Gamma(\varepsilon/2+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k)/\Gamma(\varepsilon/2+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'})}{\Gamma(\varepsilon+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k)/\Gamma(\varepsilon+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'})} \frac{\int_\varepsilon^\infty t^l \frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot} dt}{\int_0^{\varepsilon/2} t^l \frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot} dt},
\end{aligned}$$

the right-hand side of which converges to zero as $k \to \infty$ since $\Gamma(\varepsilon/2+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k)/\Gamma(\varepsilon+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k) \sim 1/(\varepsilon+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k)^{\varepsilon/2}$ as $k \to \infty$. Therefore, as $k \to \infty$,

$$\delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z}+k\boldsymbol{e}^{(N)}_{\nu'}) \sim \frac{\int_0^\varepsilon \frac{t}{(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'})\cdots(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k-1)} \frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot} dt}{\int_0^\varepsilon \frac{1}{(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'})\cdots(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k-1)} \frac{f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t)}{t+r+a_0+z_\nu+a_\cdot} dt} \le \varepsilon.$$

Since ε is arbitrarily chosen, we conclude that $\lim_{\mathbb{N}\ni k\to\infty} \delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z}+k\boldsymbol{e}^{(N)}_{\nu'}) = 0$. For part (iv), let $k \in \mathbb{N} \setminus \{1\}$. Then

$$\begin{aligned}
\frac{\delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z}+k\boldsymbol{j}^{(N)})}{1/\log k} &= \frac{\int_0^\infty (\log k) t^\alpha e^{-\beta t} g(t) \big\{ \prod_{\nu'=1}^N \frac{\Gamma(t+r+a_0)}{\Gamma(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k)} \big\} dt}{\int_0^\infty t^{\alpha-1} e^{-\beta t} g(t) \big\{ \prod_{\nu'=1}^N \frac{\Gamma(t+r+a_0)}{\Gamma(t+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k)} \big\} dt} \\
&= \frac{\int_0^\infty u^\alpha e^{-\beta u/\log k} g\big(\frac{u}{\log k}\big) \big\{ \prod_{\nu'=1}^N \frac{\Gamma(u/\log k+r+a_0)\Gamma(r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k)}{\Gamma(u/\log k+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k)\Gamma(r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'})} \big\} du}{\int_0^\infty u^{\alpha-1} e^{-\beta u/\log k} g\big(\frac{u}{\log k}\big) \big\{ \prod_{\nu'=1}^N \frac{\Gamma(u/\log k+r+a_0)\Gamma(r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k)}{\Gamma(u/\log k+r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'}+k)\Gamma(r+a_0+z_{\nu'}+a_\cdot+\delta^{(N)}_{\nu,\nu'})} \big\} du}.
\end{aligned}$$

Now for each $l \in \{0,1\}$ and all $u \in (0,\infty)$, we have that

$$
\begin{aligned}
&u^{\alpha+l-1}e^{-\beta u/\log k}g\Big(\frac{u}{\log k}\Big)\prod_{\nu'=1}^{N}\frac{\Gamma(u/\log k+r+a_0)\Gamma(r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}+k)}{\Gamma(u/\log k+r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}+k)\Gamma(r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)})}\\
&\le\frac{\big[\sup_{t\in(0,\infty)}\big\{g(t)\prod_{\nu'=1}^{N}\frac{\Gamma(t+r+a_0)}{\Gamma(t+r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)})}\big\}\big]u^{\alpha+l-1}}{\prod_{\nu'=1}^{N}\big\{\big(1+\frac{u/\log k}{r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}}\big)\cdots\big(1+\frac{u/\log k}{r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}+k-1}\big)\big\}}\\
&\le\frac{\big[\sup_{t\in(0,\infty)}\big\{g(t)\prod_{\nu'=1}^{N}\frac{\Gamma(t+r+a_0)}{\Gamma(t+r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)})}\big\}\big]u^{\alpha+l-1}}{\prod_{\nu'=1}^{N}\big\{1+u\big(\log\frac{r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}+k}{r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}}\big)/\log k\big\}}\\
&\le\frac{\big[\sup_{t\in(0,\infty)}\big\{g(t)\prod_{\nu'=1}^{N}\frac{\Gamma(t+r+a_0)}{\Gamma(t+r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)})}\big\}\big]u^{\alpha+l-1}}{\prod_{\nu'=1}^{N}\big[1+u\inf_{k'\in\mathbb{N}\setminus\{1\}}\big\{\big(\log\frac{r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}+k'}{r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}}\big)/\log k'\big\}\big]},
\end{aligned}
$$

where the second inequality follows since

$$
\begin{aligned}
&\Big(1+\frac{u/\log k}{r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}}\Big)\times\cdots\times\Big(1+\frac{u/\log k}{r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}+k-1}\Big)\\
&\ge 1+\frac{u}{\log k}\Big(\frac{1}{r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}}+\cdots+\frac{1}{r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}+k-1}\Big)\\
&\ge 1+\frac{u}{\log k}\log\frac{r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}+k}{r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}}
\end{aligned}
$$

for every $\nu' \in \{1,\ldots,N\}$, and that

$$
\begin{aligned}
&\lim_{\mathbb{N}\setminus\{1\}\ni k\to\infty}\Big\{u^{\alpha+l-1}e^{-\beta u/\log k}g\Big(\frac{u}{\log k}\Big)\\
&\quad\times\prod_{\nu'=1}^{N}\frac{\Gamma(u/\log k+r+a_0)\Gamma(r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}+k)}{\Gamma(u/\log k+r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}+k)\Gamma(r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)})}\Big\}\\
&=g(0)\Big\{\prod_{\nu'=1}^{N}\frac{\Gamma(r+a_0)}{\Gamma(r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)})}\Big\}u^{\alpha+l-1}\\
&\quad\times\prod_{\nu'=1}^{N}\lim_{\mathbb{N}\setminus\{1\}\ni k\to\infty}\frac{\Gamma(r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}+k)}{\Gamma(u/\log k+r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)}+k)}\\
&=g(0)\Big\{\prod_{\nu'=1}^{N}\frac{\Gamma(r+a_0)}{\Gamma(r+a_0+z_{\nu'}+a.+\delta_{\nu,\nu'}^{(N)})}\Big\}u^{\alpha+l-1}e^{-Nu}.
\end{aligned}
$$

Thus,

$$
\lim_{\mathbb{N}\setminus\{1\}\ni k\to\infty}\frac{\delta_{\nu}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{z}+k\boldsymbol{j}^{(N)})}{1/\log k}=\frac{\int_0^\infty u^{\alpha}e^{-Nu}du}{\int_0^\infty u^{\alpha-1}e^{-Nu}du}=\frac{\alpha}{N},
$$

and the result follows. □

The following lemma will be used in the proof of Theorem 3.3.2.

Lemma 3.6.2 *Let $u_1, u_2 > 0$. Then*

$$\frac{\Gamma'(u_1+u_2)}{\Gamma(u_1+u_2)} - \frac{\Gamma'(u_1)}{\Gamma(u_1)} \geq \frac{u_2}{u_1+u_2}.$$

Proof. By Theorem 2.1 of Muldoon (1978), we have

$$\frac{\partial}{\partial u}\Big\{\frac{\Gamma'(u)}{\Gamma(u)} - \log u\Big\} > 0 \tag{3.6.8}$$

for all $u > 0$. Therefore,

$$\frac{\Gamma'(u_1+u_2)}{\Gamma(u_1+u_2)} - \frac{\Gamma'(u_1)}{\Gamma(u_1)} \geq \log\frac{u_1+u_2}{u_1} \geq \frac{u_2}{u_1+u_2}$$

as desired. □

Although the inequality (3.6.8) is used in the above proof of Lemma 3.6.2, we can prove the result directly.

Proof of Theorem 3.3.2. The proof is similar to that of Theorem 3.3.1. First, note that $r > 2$ and $a_\cdot > 1$ by assumption. Let $\Delta_n^{(\alpha,\beta,g,a_0,\boldsymbol{a})} = E[\tilde{L}_n(\hat{\boldsymbol{p}}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}, \boldsymbol{p})] - E[\tilde{L}_n(\hat{\boldsymbol{p}}^{(a_0,\boldsymbol{a})}, \boldsymbol{p})]$. For $\nu \in \{1,\ldots,N\}$, let

$$\phi_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{X}) = \frac{K(\alpha+1,\beta,g,r+a_0,\boldsymbol{X}_\cdot + a_\cdot \boldsymbol{j}^{(N)} + \boldsymbol{e}_\nu^{(N)})}{K(\alpha,\beta,g,r+a_0,\boldsymbol{X}_\cdot + a_\cdot \boldsymbol{j}^{(N)})}$$

so that for every $i \in \{1,\ldots,m\}$,

$$\hat{p}_{i,\nu}^{(\alpha,\beta,g,a_0,\boldsymbol{a})} = \hat{p}_{i,\nu}^{(a_0,\boldsymbol{a})} - \hat{p}_{i,\nu}^{(a_0,\boldsymbol{a})}\phi_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{X}).$$

By Lemma 3.6.1, we have

$$\begin{aligned}
\Delta_n^{(\alpha,\beta,g,a_0,\boldsymbol{a})} &= E\Big[\sum_{\nu=1}^{n}\sum_{i=1}^{m}\Big[-\hat{p}_{i,\nu}^{(a_0,\boldsymbol{a})}\phi_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{X}) + p_{i,\nu}\log\Big\{1+\frac{\delta_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{X}_\cdot)}{r+a_0+X_{\cdot,\nu}+a_\cdot}\Big\}\Big]\Big] \\
&= E\Big[\sum_{\nu=1}^{n}\sum_{i=1}^{m}\Big[-\hat{p}_{i,\nu}^{(a_0,\boldsymbol{a})}\phi_\nu^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{X}) \\
&\quad + \hat{p}_{i,\nu}^{\mathrm{U}}\log\Big\{1+\frac{1}{r+a_0+X_{\cdot,\nu}+a_\cdot-1}\frac{K(\alpha+1,\beta,g,r+a_0,\boldsymbol{X}_\cdot+a_\cdot\boldsymbol{j}^{(N)})}{K(\alpha,\beta,g,r+a_0,\boldsymbol{X}_\cdot+a_\cdot\boldsymbol{j}^{(N)})}\Big\}\Big]\Big] \\
&= E[-I_{1,n}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{x}) + I_{2,n}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{x})],
\end{aligned}$$

where

$$I_{1,n}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{x}) = \sum_{\nu=1}^{n} \frac{x_{\cdot,\nu} + a_\cdot}{r + a_0 + x_{\cdot,\nu} + a_\cdot} \frac{K(\alpha+1,\beta,g,r+a_0,\boldsymbol{x}_\cdot + a_\cdot \boldsymbol{j}^{(N)} + \boldsymbol{e}_\nu^{(N)})}{K(\alpha,\beta,g,r+a_0,\boldsymbol{x}_\cdot + a_\cdot \boldsymbol{j}^{(N)})},$$

$$I_{2,n}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{x}) = \sum_{\nu=1}^{n} \frac{x_{\cdot,\nu}}{r + x_{\cdot,\nu} - 1} \log\Big\{1 + \frac{1}{r + a_0 + x_{\cdot,\nu} + a_\cdot - 1} \frac{K(\alpha+1,\beta,g,r+a_0,\boldsymbol{x}_\cdot + a_\cdot \boldsymbol{j}^{(N)})}{K(\alpha,\beta,g,r+a_0,\boldsymbol{x}_\cdot + a_\cdot \boldsymbol{j}^{(N)})}\Big\},$$

and $\boldsymbol{x}_\cdot = (x_{\cdot,1},\ldots,x_{\cdot,N})^\top = \big(\sum_{i=1}^m x_{i,1},\ldots,\sum_{i=1}^m x_{i,N}\big)^\top$ for $\boldsymbol{x} = (x_{i,\nu})_{1\le i\le m,\,1\le\nu\le N} \in {\mathbb{N}_0}^{m\times N}$. Since $-I_{1,n}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{0}^{(m,N)}) + I_{2,n}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{0}^{(m,N)}) < 0$, it is sufficient to show that $-I_{1,n}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{x}) + I_{2,n}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{x}) \le 0$ for all $\boldsymbol{x} \in {\mathbb{N}_0}^{m\times N} \setminus \{\boldsymbol{0}^{(m,N)}\}$.

Fix $\boldsymbol{x} = (x_{i,\nu})_{1\le i\le m,\,1\le\nu\le N} \in {\mathbb{N}_0}^{m\times N} \setminus \{\boldsymbol{0}^{(m,N)}\}$. Let $z_\nu = \sum_{i=1}^m x_{i,\nu}$ for $\nu \in \{1,\ldots,N\}$ and let $\boldsymbol{z} = (z_1,\ldots,z_N)^\top$ and $z = \sum_{\nu=1}^N z_\nu$. We use the abbreviated notation

$$\tilde{I}_1 = I_{1,n}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{x}), \quad \tilde{I}_2 = I_{2,n}^{(\alpha,\beta,g,a_0,\boldsymbol{a})}(\boldsymbol{x}), \quad \tilde{I} = -\tilde{I}_1 + \tilde{I}_2,$$

$$\tilde{K}(l) = K(\alpha+l,\beta,g,r+a_0,\boldsymbol{x}_\cdot + a_\cdot \boldsymbol{j}^{(N)}), \quad \tilde{K}(l,\nu) = K(\alpha+l,\beta,g,r+a_0,\boldsymbol{x}_\cdot + a_\cdot \boldsymbol{j}^{(N)} + \boldsymbol{e}_\nu^{(N)}),$$

for $l \in \{0,1,2\}$ and $\nu \in \{1,\ldots,N\}$. Also, let

$$f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t) = t^{\alpha-1} e^{-\beta t} g(t) \prod_{\nu=1}^{N} \frac{\Gamma(t+r+a_0)}{\Gamma(t+r+a_0+z_\nu+a_\cdot)}$$

for $t \in (0,\infty)$.

Clearly,

$$\begin{aligned}
\tilde{I}_2 &= \sum_{\nu=1}^{n} \frac{z_\nu}{r+z_\nu-1} \log\Big\{1 + \frac{1}{r+a_0+z_\nu+a_\cdot-1} \frac{\tilde{K}(1)}{\tilde{K}(0)}\Big\} \\
&\le \sum_{\nu=1}^{n} \frac{z_\nu}{r+z_\nu-1} \frac{1}{r+a_0+z_\nu+a_\cdot-1} \frac{\tilde{K}(1)}{\tilde{K}(0)} \\
&\le \sum_{\nu=1}^{n} \frac{z_\nu+a_0+a_\cdot+2}{(r+a_0+z_\nu+a_\cdot)^2} \frac{\tilde{K}(1)}{\tilde{K}(0)} \qquad (3.6.9)
\end{aligned}$$

by assumption. On the other hand,

$$\begin{aligned}
\tilde{I}_1 &= \sum_{\nu=1}^{n} \frac{z_\nu+a_\cdot}{r+a_0+z_\nu+a_\cdot} \frac{\tilde{K}(1,\nu)}{\tilde{K}(0)} = \sum_{\nu=1}^{n} \frac{z_\nu+a_\cdot}{(r+a_0+z_\nu+a_\cdot)^2} \frac{\tilde{K}(1)}{\tilde{K}(0)} - \sum_{\nu=1}^{n} \frac{z_\nu+a_\cdot}{(r+a_0+z_\nu+a_\cdot)^2} \frac{\tilde{K}(2,\nu)}{\tilde{K}(0)} \\
&\ge \sum_{\nu=1}^{n} \frac{z_\nu+a_\cdot}{(r+a_0+z_\nu+a_\cdot)^2} \frac{\tilde{K}(1)}{\tilde{K}(0)} - \frac{1}{n} \sum_{\nu=1}^{n} \frac{1}{(r+a_0+z_\nu+a_\cdot)^2} \sum_{\nu=1}^{n} (z_\nu+a_\cdot) \frac{\tilde{K}(2,\nu)}{\tilde{K}(0)} \qquad (3.6.10)
\end{aligned}$$

by the covariance inequality. Furthermore, by integration by parts, we have

$$\begin{aligned}
(\alpha+1)\tilde{K}(1) &= \int_0^\infty t^2 f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t) \Big\{\beta + \frac{-g'(t)}{g(t)}\Big\} dt \\
&\quad + \sum_{\nu=1}^{N} \int_0^\infty t^2 f_{\alpha,\beta,g,a_0,\boldsymbol{a}}(t) \Big\{\frac{\Gamma'(t+r+a_0+z_\nu+a_\cdot)}{\Gamma(t+r+a_0+z_\nu+a_\cdot)} - \frac{\Gamma'(t+r+a_0)}{\Gamma(t+r+a_0)}\Big\} \\
&\ge \sum_{\nu=1}^{n} (z_\nu+a_\cdot) \tilde{K}(2,\nu), \qquad (3.6.11)
\end{aligned}$$

where the equality follows since $\Gamma(t) \sim t^{-1}$ as $t \to 0$ while $\prod_{\nu=1}^{N}\{\Gamma(t+r+a_0)/\Gamma(t+r+a_0+z_\nu+a_\cdot)\} \sim t^{-z-Na_\cdot}$ as $t \to \infty$ and where the inequality follows from Lemma 3.6.2. Hence, combining (3.6.9), (3.6.10), and (3.6.11), we obtain

$$\begin{aligned}\tilde{I} &\le -\sum_{\nu=1}^{n} \frac{z_\nu + a_\cdot - (\alpha+1)/n}{(r+a_0+z_\nu+a_\cdot)^2} \frac{\tilde{K}(1)}{\tilde{K}(0)} + \sum_{\nu=1}^{n} \frac{z_\nu + a_0 + a_\cdot + 2}{(r+a_0+z_\nu+a_\cdot)^2} \frac{\tilde{K}(1)}{\tilde{K}(0)} \\ &= -\sum_{\nu=1}^{n} \frac{-a_0 - 2 - (\alpha+1)/n}{(r+a_0+z_\nu+a_\cdot)^2} \frac{\tilde{K}(1)}{\tilde{K}(0)},\end{aligned}$$

the right-hand side of which is nonpositive by assumption (3.3.10), and the result follows. □

Proof of Proposition 3.3.3. Properties (ii) and (iv) are trivial. Property (iii) follows since

$$\int_0^\infty \pi(\boldsymbol{p}, t|\alpha, \beta, a_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_N)dt = \pi(\boldsymbol{p}|\alpha, \beta, a_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_N)$$

for $\boldsymbol{p} \in {D_m}^N$. For part (i), note that the integrals are finite only if $a_0 \ge 0$ since otherwise

$$\int_{{D_m}^N} \pi(\boldsymbol{p}|\alpha, \beta, a_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_N)d\boldsymbol{p} \ge \int_{{D_m}^N} \pi_{\alpha,\beta,g_1,a_0,\overline{a}\boldsymbol{j}^{(m)}}(\boldsymbol{p})d\boldsymbol{p} = \infty,$$

where $\overline{a} = \max\{\max\{a_{1,1}, \ldots, a_{m,1}\}, \ldots, \max\{a_{1,N}, \ldots, a_{m,N}\}\}$, by Lemma 3.3.1. Suppose that $a_0 \ge 0$. Then we have

$$\begin{aligned}&\int_{{D_m}^N \times (0,\infty)} \pi(\boldsymbol{p}, t|\alpha, \beta, a_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_N)d(\boldsymbol{p}, t) / \prod_{\nu=1}^{N}\prod_{i=1}^{m} \Gamma(a_{i,\nu}) \\ &= \int_0^1 t^{\alpha-1}e^{-\beta t}\Big\{\prod_{\nu=1}^{N} \frac{\Gamma(t+a_0)}{\Gamma(t+a_0+a_{\cdot,\nu})}\Big\}dt + \int_1^\infty t^{\alpha-1}e^{-\beta t}\Big\{\prod_{\nu=1}^{N} \frac{\Gamma(t+a_0)}{\Gamma(t+a_0+a_{\cdot,\nu})}\Big\}dt.\end{aligned}$$

The first term on the right side is finite if and only if $a_0 > 0$ or $\alpha > N$ since $\Gamma(t) \sim t^{-1}$ as $t \to 0$. The second term on the right side is finite if and only if $a_{\cdot,\cdot} > \alpha$ or $\beta > 0$ since $\prod_{\nu=1}^{N} \Gamma(t+a_0)/\Gamma(t+a_0+a_{\cdot,\nu}) \sim t^{-a_{\cdot,\cdot}}$ as $t \to \infty$. This completes the proof. □

Chapter 4

Bayesian Shrinkage Approaches to Unbalanced Problems of Estimation and Prediction on the Basis of Negative Multinomial Samples

4.1 Introduction

Properties of shrinkage estimators based on count variables have been extensively investigated within the decision-theoretic framework since the seminal work of Clevenson and Zidek (1975). For example, as briefly reviewed in Section 1 of Hamura and Kubokawa (2020b), estimation of Poisson parameters was studied by Ghosh and Parsian (1981), Tsui (1979b), Tsui and Press (1982), and Ghosh and Yang (1988) in various settings while Tsui (1979a), Hwang (1982), and Ghosh, Hwang, and Tsui (1983) showed that similar results hold for discrete exponential families. Extending the result of Tsui (1984) and Tsui (1986a), Tsui (1986b) proved that Clevenson–Zidek-type estimators dominate the usual estimator in the case of the negagive multinomial distribution, which is a generalization of the negative binomial distribution and is a special case of the general distributions of Chou (1991) and Dey and Chung (1992). More recent studies include Chang and Shinozaki (2019), Stoltenberg and Hjort (2019), and Hamura and Kubokawa (2019b, 2020b, 2020c). On the other hand, since Komaki (2001), Bayesian predictive densities with respect to shrinkage priors have been shown to dominate those based on noninformative priors and parallels between estimation and prediction have been noted in the literature. In particular, Komaki (2004, 2006, 2015) and Hamura and Kubokawa (2019b) obtained dominance conditions in the Poisson case.

There are still directions in which these results could be generalized further. First, although sample sizes will be unbalanced in many practical situations, some of the results are applicable only to the balanced case. Weights in loss functions may also be unbalanced in practice (see, for example, Section 7 of Stoltenberg and Hjort (2019)). Second, as pointed out by Hamura and Kubokawa (2020b), decision-theoretic properties of Bayesian procedures have not been fully studied for discrete distributions other than the Poisson distribution. Even in the Poisson case, it was only after the work of Komaki (2015) that many Bayesian shrinkage estimators were shown to dominate usual estimators in the presence of unbalanced sample sizes (Hamura and

Kubokawa (2019b, 2020c)). Third, while theoretical properties of Bayesian predictive densities for Poisson models have been investigated in several papers as mentioned earlier, relatively few researchers (Komaki (2012), Hamura and Kubokawa (2019a)) have considered predictive density estimation for other discrete exponential families. In this chapter, we treat these three issues when considering Bayesian estimators and predictive density estimators based on negative multinomial observations in unbalanced settings.

In Section 4.2, we consider the problem of estimating negative multinomial parameter vectors under the standardized squared error loss in the general case where sample sizes, lengths of observation vectors, and weights in the loss function may all be unbalanced. First, we generalize Theorem 1 of Hamura and Kubokawa (2020b) to this unbalanced case and also obtain another general sufficient condition for a general shrinkage estimator to dominate the UMVU estimator. Then, using the method of maximum likelihood, a new empirical Bayes estimator is derived which has a simple form as well as improves on the UMVU estimator. Finally, we present still another dominance condition, which is applicable specifically to empirical Bayes estimators including those based on the method of moments.

In Section 4.3, we consider the practically important problem of estimating the joint predictive density of several independent multinomial tables under the Kullback-Leibler divergence. The distribution of any one of them is specified by a set of negative multinomial probability vectors, with each cell probability given by the product of the corresponding elements of the vectors. The setting we consider is quite general in that two tables may be related through a set of common overlapping probability vectors. Two simple special cases are the prediction problems for independent multinomial vectors and for a single multinomial table. We show that the Bayesian predictive density with respect to the Jeffreys prior is dominated by that with respect to a generalization of the shrinkage prior considered by Hamura and Kubokawa (2020b) under suitable conditions. Whereas Komaki (2012) investigated asymptotic properties of Bayesian predictive densities for future multinomial observations based on current multinomial observations, the sample space is not a finite set in our setting and we investigate finite sample properties of Bayesian predictive densities. Although Hamura and Kubokawa (2019a) considered Bayesian predictive densities for a negative binomial model, where a future observation also is negative binomial and can take on an infinite number of values, they did not treat the problem of estimating the joint predictive density of multiple negative binomial observations.

In Section 4.4, simple and illustrative simulation studies are performed. In Section 4.4.1, our proposed empirical Bayes estimator and the UMVU estimator given in Section 4.2 are compared. In Section 4.4.2, the Bayesian predictive densities given in Section 4.3 are compared.

In Section 4.5, predictive density estimation for the negative multinomial distribution is discussed. Although no dominance conditions are obtained, generalizing Theorem 2.1 of Hamura and Kubokawa (2019a), we derive two kinds of identities which relate prediction to estimation in the negative multinomial case. In particular, the risk function of an arbitrary Bayesian predictive density under the Kullback-Leibler divergence is expressed using the risk functions of an infinite number of corresponding Bayes estimators under a weighted version of Stein's loss.

4.2 Empirical Bayes Point Estimation

Let $N \in \mathbb{N} = \{1, 2, \dots\}$, $m_1, \dots, m_N \in \mathbb{N}$, and $r_1, \dots, r_N > 0$. For $\nu = 1, \dots, N$, let $\boldsymbol{p}_\nu = (p_{i,\nu})_{i=1}^{m_\nu} \in D_{m_\nu} = \{(\mathring{p}_1, \dots, \mathring{p}_{m_\nu})^\top | \mathring{p}_1, \dots, \mathring{p}_{m_\nu} > 0, \sum_{i=1}^{m_\nu} \mathring{p}_i < 1\}$ and let $p_{0,\nu} = 1 - p_{\cdot,\nu} =$

$1-\sum_{i=1}^{m_\nu} p_{i,\nu}$. Let $\boldsymbol{X}_1,\dots,\boldsymbol{X}_N$ be independent negative multinomial variables such that for each $\nu=1,\dots,N$, the probability mass function of $\boldsymbol{X}_\nu$ is given by

$$\frac{\Gamma\bigl(r_\nu+\sum_{i=1}^{m_\nu}x_{i,\nu}\bigr)}{\Gamma(r_\nu)\prod_{i=1}^{m_\nu}x_{i,\nu}!}{p_{0,\nu}}^{r_\nu}\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{x_{i,\nu}}$$

for $\boldsymbol{x}_\nu=(x_{i,\nu})_{i=1}^{m_\nu}\in{\mathbb{N}_0}^{m_\nu}$, where $\mathbb{N}_0=\{0,1,2,\dots\}$. As pointed out by Hamura and Kubokawa (2020b), $m_1,\dots,m_N$ may be different for example when we consider marginal distributions of negative multinomial vectors of the same length. For some basic properties of the negative multinomial distribution, see Sibuya, Yoshimura, and Shimizu (1964) and Tsui (1986b).

Now we assume that all the elements of $\boldsymbol{p}=(\boldsymbol{p}_\nu)_{\nu=1,\dots,N}\in D=D_{m_1}\times\cdots\times D_{m_N}$ are unknown and consider the problem of estimating $\boldsymbol{p}$ on the basis of the minimal and complete sufficient statistic $\boldsymbol{X}=(\boldsymbol{X}_\nu)_{\nu=1,\dots,N}=((X_{i,\nu})_{i=1}^{m_\nu})_{\nu=1,\dots,N}$ under the standardized squared loss function given by

$$L_{n,\boldsymbol{c}}(\boldsymbol{d},\boldsymbol{p})=\sum_{\nu=1}^{n}\sum_{i=1}^{m_\nu}c_{i,\nu}\frac{(d_{i,\nu}-p_{i,\nu})^2}{p_{i,\nu}} \tag{4.2.1}$$

for $\boldsymbol{d}=((d_{i,\nu})_{i=1}^{m_\nu})_{\nu=1,\dots,N}\in\mathbb{R}^{m_1}\times\cdots\times\mathbb{R}^{m_N}$, where $n\in\{1,\dots,N\}$ and $\boldsymbol{c}=((c_{i,\nu})_{i=1}^{m_\nu})_{\nu=1,\dots,N}\in[0,\infty)^{m_1}\times\cdots\times[0,\infty)^{m_N}$.

For $\nu=1,\dots,N$, let $X_{\cdot,\nu}=\sum_{i=1}^{m_\nu}X_{i,\nu}$. Then the UMVU estimator of $\boldsymbol{p}$ is $\hat{\boldsymbol{p}}^{\mathrm{U}}=((\hat{p}_{i,\nu}^{\mathrm{U}})_{i=1}^{m_\nu})_{\nu=1,\dots,N}$, where

$$\hat{p}_{i,\nu}^{\mathrm{U}}=\frac{X_{i,\nu}}{r_\nu+X_{\cdot,\nu}-1} \tag{4.2.2}$$

for $i=1,\dots,m_\nu$ for $\nu=1,\dots,N$. (We write $0/0=0$.) We first derive a general sufficient condition for the shrinkage estimator

$$\hat{\boldsymbol{p}}^{(\boldsymbol{\delta})}=((\hat{p}_{i,\nu}^{(\boldsymbol{\delta})})_{i=1}^{m_\nu})_{\nu=1,\dots,N}=\Bigl(\Bigl(\frac{X_{i,\nu}}{r_\nu+X_{\cdot,\nu}-1+\delta_\nu(X_{\cdot,\cdot})}\Bigr)_{i=1}^{m_\nu}\Bigr)_{\nu=1,\dots,N} \tag{4.2.3}$$

to dominate $\hat{\boldsymbol{p}}^{\mathrm{U}}$, where $\boldsymbol{\delta}=(\delta_\nu)_{\nu=1}^{N}\colon\mathbb{N}_0\to(0,\infty)^N$ and $X_{\cdot,\cdot}=\sum_{\nu=1}^{N}X_{\cdot,\nu}=\sum_{\nu=1}^{N}\sum_{i=1}^{m_\nu}X_{i,\nu}$. For notational simplicity, let $\underline{r}=\min_{1\le\nu\le n}r_\nu$ and $\overline{r}=\max_{1\le\nu\le n}r_\nu$. For $\nu=1,\dots,N$, let $c_{\cdot,\nu}=\sum_{i=1}^{m_\nu}c_{i,\nu}$. Let $\underline{c}_{\cdot}=\min_{1\le\nu\le n}c_{\cdot,\nu}$ and $\overline{\overline{c}}=\max_{1\le\nu\le n}\max_{1\le i\le m_\nu}c_{i,\nu}$. Finally, let $\underline{\delta}(x)=\min_{1\le\nu\le n}\delta_\nu(x)$ and $\overline{\delta}(x)=\max_{1\le\nu\le n}\delta_\nu(x)$ for $x\in\mathbb{N}_0$ and let $\rho=\inf_{x\in\mathbb{N}\setminus\{1\}}\underline{\delta}(x)/\overline{\delta}(x)\in[0,1]$.

Theorem 4.2.1 *Assume that $r_\nu\ge5/2$ for all $\nu=1,\dots,n$ with $c_{\cdot,\nu}>0$ and that $0<3\overline{\overline{c}}\le\underline{c}_{\cdot}$. Suppose that for all $\nu=1,\dots,n$ such that $c_{\cdot,\nu}>0$ and for all $x\in\mathbb{N}$, we have*

$$x\delta_\nu(x)\le(x+1)\delta_\nu(x+1). \tag{4.2.4}$$

Suppose further that for all $x\in\mathbb{N}$, one of the following two conditions are satisfied:

(i) • $\overline{\overline{c}}\overline{\delta}(x+1)\le2(\underline{r}/\overline{r})^2(\underline{c}_{\cdot}-3\overline{\overline{c}})\rho$ *implies*

$$\Bigl\{2\Bigl(\frac{\underline{r}}{\overline{r}}\Bigr)^2(\underline{c}_{\cdot}-3\overline{\overline{c}})\rho-\underline{c}_{\cdot}\Bigr\}\overline{\delta}(x+1)+2\underline{r}\Bigl(\frac{\underline{r}}{\overline{r}}\Bigr)^2(\underline{c}_{\cdot}-3\overline{\overline{c}})\rho\ge0\quad\textit{and} \tag{4.2.5}$$

- $\overline{\underline{c}}\overline{\delta}(x+1) > 2(\underline{r}/\overline{r})^2(\underline{c}_\cdot - 3\overline{\underline{c}})\rho$ *implies*

$$n\Big[\Big\{2\Big(\frac{\underline{r}}{\overline{r}}\Big)^2(\underline{c}_\cdot - 3\overline{\underline{c}})\rho - \underline{c}_\cdot\Big\}\overline{\delta}(x+1) + 2\underline{r}\Big(\frac{\underline{r}}{\overline{r}}\Big)^2(\underline{c}_\cdot - 3\overline{\underline{c}})\rho\Big] \geq x\Big\{\overline{\underline{c}}\overline{\delta}(x+1) - 2\Big(\frac{\underline{r}}{\overline{r}}\Big)^2(\underline{c}_\cdot - 3\overline{\underline{c}})\rho\Big\}. \tag{4.2.6}$$

(ii) - $\overline{\underline{c}}\overline{\delta}(x+1) \leq 2(\underline{c}_\cdot - 3\overline{\underline{c}})\rho$ *implies*

$$2(\underline{c}_\cdot - 3\overline{\underline{c}})\rho - (\underline{c}_\cdot - \underline{r}\overline{\underline{c}}) \geq 0 \quad \textit{and} \tag{4.2.7}$$

- $\overline{\underline{c}}\overline{\delta}(x+1) > 2(\underline{c}_\cdot - 3\overline{\underline{c}})\rho$ *implies*

$$n\{2(\underline{c}_\cdot - 3\overline{\underline{c}})\rho - (\underline{c}_\cdot - \underline{r}\overline{\underline{c}})\}\overline{\delta}(x+1) \geq \Big(\sum_{\nu=1}^n r_\nu + x\Big)\{\overline{\underline{c}}\overline{\delta}(x+1) - 2(\underline{c}_\cdot - 3\overline{\underline{c}})\rho\}. \tag{4.2.8}$$

Then the shrinkage estimator $\hat{\boldsymbol{p}}^{(\boldsymbol{\delta})}$ given in (4.2.3) dominates the UMVU estimator $\hat{\boldsymbol{p}}^{\mathrm{U}}$ given by (4.2.2) under the standardized squared loss (4.2.1).

Part (i) of Theorem 4.2.1 is a generalization of Theorem 1 of Hamura and Kubokawa (2020b), who further obtained simpler conditions in specific cases. On the other hand, part (ii) is another result of this chapter. It is worth noting that under the setting of Theorem 4.2.1, there may exist $\nu = 1, \ldots, n$ such that $c_{i,\nu} = 0 < c_{i',\nu}$ for some $i, i' = 1, \ldots, m_\nu$.

Next, we derive an empirical Bayes estimator based on the method of maximum likelihood. Consider the conjugate Dirichlet prior distribution

$$\prod_{\nu=1}^N \mathrm{Dir}_{m_\nu}(\boldsymbol{p}_\nu | \tilde{a}_\nu v, \boldsymbol{j}^{(m_\nu)}) = \prod_{\nu=1}^N \Big\{\frac{\Gamma(\tilde{a}_\nu v + m_\nu)}{\Gamma(\tilde{a}_\nu v)} {p_{0,\nu}}^{\tilde{a}_\nu v - 1}\Big\},$$

where $v \in (0, \infty)$ and where $\tilde{a}_\nu \in (0, \infty)$ and $\boldsymbol{j}^{(m_\nu)} = (1, \ldots, 1)^\top \in \mathbb{R}^{m_\nu}$ for $\nu = 1, \ldots, N$. It corresponds to the Bayes estimator

$$\Big(\Big(\frac{X_{i,\nu}}{r_\nu + X_{\cdot,\nu} - 1 + \tilde{a}_\nu v + m_\nu}\Big)_{i=1}^{m_\nu}\Big)_{\nu=1,\ldots,N}$$

of $\boldsymbol{p}$. On the other hand, since the maximum likelihood estimator and the prior mean of $p_{0,\nu}$ is $r_\nu/(r_\nu + X_{\cdot,\nu})$ and $\tilde{a}_\nu v/(\tilde{a}_\nu v + m_\nu)$ for $\nu = 1, \ldots, N$, a reasonable estimator of v would be

$$\frac{1}{X_{\cdot,\cdot}} \sum_{\nu=1}^N \frac{m_\nu r_\nu}{\tilde{a}_\nu}.$$

Thus, we obtain the empirical Bayes estimator

$$\hat{\boldsymbol{p}}^{(\tilde{\boldsymbol{a}})} = \Big(\Big(\frac{X_{i,\nu}}{r_\nu + X_{\cdot,\nu} - 1 + \delta_\nu^{(\tilde{\boldsymbol{a}})}(X_{\cdot,\cdot})}\Big)_{i=1}^{m_\nu}\Big)_{\nu=1,\ldots,N}, \tag{4.2.9}$$

where $\tilde{\boldsymbol{a}} = (\tilde{a}_\nu)_{\nu=1}^N$ and where

$$\delta_\nu^{(\tilde{\boldsymbol{a}})}(X_{\cdot,\cdot}) = m_\nu + \frac{\tilde{a}_\nu}{X_{\cdot,\cdot}} \sum_{\nu'=1}^N \frac{m_{\nu'} r_{\nu'}}{\tilde{a}_{\nu'}}$$

if $X_{\cdot,\cdot} \geq 1$ while $\delta_\nu^{(\tilde{\boldsymbol{a}})}(0) \in (0,\infty)$ for $\nu = 1,\ldots,N$. This estimator was not considered by Hamura and Kubokawa (2020b). It is of the form (4.2.3) and clearly satisfies condition (4.2.4). Whether the other conditions hold or not depends on the choice of the hyperparameter $\tilde{\boldsymbol{a}}$. For example,

$$\rho = \begin{cases} \inf\limits_{x\in\mathbb{N}\setminus\{1\}} \dfrac{(\min_{1\leq\nu\leq n} m_\nu)\big(1+\sum_{\nu'=1}^N r_{\nu'}/x\big)}{(\max_{1\leq\nu\leq n} m_\nu)\big(1+\sum_{\nu'=1}^N r_{\nu'}/x\big)} = \dfrac{\min_{1\leq\nu\leq n} m_\nu}{\max_{1\leq\nu\leq n} m_\nu}, & \text{if } \tilde{\boldsymbol{a}} = (m_\nu)_{\nu=1}^N, \\ \inf\limits_{x\in\mathbb{N}\setminus\{1\}} \dfrac{\min_{1\leq\nu\leq n} m_\nu + \sum_{\nu'=1}^N m_{\nu'} r_{\nu'}/x}{\max_{1\leq\nu\leq n} m_\nu + \sum_{\nu'=1}^N m_{\nu'} r_{\nu'}/x} = \dfrac{\min_{1\leq\nu\leq n} m_\nu}{\max_{1\leq\nu\leq n} m_\nu}, & \text{if } \tilde{\boldsymbol{a}} = \boldsymbol{j}^{(N)}, \\ \inf\limits_{x\in\mathbb{N}\setminus\{1\}} \dfrac{\min_{1\leq\nu\leq n}\big(m_\nu + r_\nu \sum_{\nu'=1}^N m_{\nu'}/x\big)}{\max_{1\leq\nu\leq n}\big(m_\nu + r_\nu \sum_{\nu'=1}^N m_{\nu'}/x\big)}, & \text{if } \tilde{\boldsymbol{a}} = (r_\nu)_{\nu=1}^N, \end{cases}$$

where $\boldsymbol{j}^{(N)} = (1,\ldots,1)^\top \in \mathbb{R}^N$.

There are other empirical Bayes estimators. For example, since the prior mean of $E[\boldsymbol{X}_{\cdot,\cdot}] = \sum_{\nu=1}^N \sum_{i=1}^{m_\nu} r_\nu p_{i,\nu}/p_{0,\nu}$ is $\sum_{\nu=1}^N \sum_{i=1}^{m_\nu} r_\nu/(v-1) = \sum_{\nu=1}^N m_\nu r_\nu/(v-1)$ when $\tilde{a}_\nu = 1$ and $v > 1$ for all $\nu = 1,\ldots,N$, one estimator of v based on the method of moments would be

$$1 + \frac{1}{X_{\cdot,\cdot}} \sum_{\nu=1}^N m_\nu r_\nu.$$

We could also use $1+\big(\sum_{\nu=1}^N \sum_{i=1}^{m_\nu} r_\nu \tilde{c}_{i,\nu}\big)/\sum_{\nu=1}^N \sum_{i=1}^{m_\nu} \tilde{c}_{i,\nu} X_{i,\nu}$ for $((\tilde{c}_{i,\nu})_{i=1}^{m_\nu})_{\nu=1,\ldots,N} \in (0,\infty)^{m_1} \times \cdots \times (0,\infty)^{m_N}$. More generally, we consider the shrinkage estimator

$$\hat{\boldsymbol{p}}^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})} = ((\hat{p}_{i,\nu}^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})})_{i=1}^{m_\nu})_{\nu=1,\ldots,N} = \Big(\Big(\frac{X_{i,\nu}}{r_\nu + X_{\cdot,\nu} - 1 + \tilde{b}_\nu + 1/\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})}}\Big)_{i=1}^{m_\nu}\Big)_{\nu=1,\ldots,N}, \tag{4.2.10}$$

where $\tilde{\boldsymbol{b}} = (\tilde{b}_\nu)_{\nu=1}^N \in (0,\infty)^N$ and $\tilde{\boldsymbol{c}} = (\tilde{\boldsymbol{c}}^{(\nu)})_{\nu=1}^N = (((\tilde{c}_{i,\nu'}^{(\nu)})_{i=1}^{m_{\nu'}})_{\nu'=1,\ldots,N})_{\nu=1}^N \in ((0,\infty)^{m_1} \times \cdots \times (0,\infty)^{m_N})^N$ and where $\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})} = \sum_{\nu'=1}^N \sum_{i=1}^{m_{\nu'}} \tilde{c}_{i,\nu'}^{(\nu)} X_{i,\nu'}$ for $\nu = 1,\ldots,N$.

Theorem 4.2.2 *Under Assumption 4.6.1 given in the Appendix, the shrinkage estimator $\hat{\boldsymbol{p}}^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}$ given in (4.2.10) dominates the UMVU estimator $\hat{\boldsymbol{p}}^{\mathrm{U}}$ given by (4.2.2) under the standardized squared loss (4.2.1).*

When $\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(1)})} = \cdots = \widetilde{X}^{(\tilde{\boldsymbol{c}}^{(N)})} = \tilde{c} X_{\cdot,\cdot}$, where $\tilde{c} \in (0,\infty)$, we have the following result.

Corollary 4.2.1 *Assume that $\tilde{\boldsymbol{c}}^{(1)} = \cdots = \tilde{\boldsymbol{c}}^{(N)} = (\tilde{c}\boldsymbol{j}^{(m_1)},\ldots,\tilde{c}\boldsymbol{j}^{(m_N)})$. Then, under Assumption 4.6.2 given in the Appendix, $\hat{\boldsymbol{p}}^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}$ dominates $\hat{\boldsymbol{p}}^{\mathrm{U}}$ under the loss (4.2.1).*

In Corollary 4.2.1, it is not necessarily assumed as in Theorem 4.2.1 that $r_\nu \geq 5/2$ for all $\nu = 1,\ldots,n$ with $c_{\cdot,\nu} > 0$. Moreover, for the balanced case with $r_1 \geq 1$, another dominance condition can be obtained by modifying the proof of Theorem 4.2.2 given in the Appendix. See Remark 4.6.1 for details.

Finally, in order to estimate $\boldsymbol{p}$, we could also use the hierarchical shrinkage prior introduced by Hamura and Kubokawa (2020b) or its generalization. However, since they considered essentially the same hierarchical Bayes estimator and gave important methods of evaluating the risk function, we do not discuss the approach further. The usefulness of hierarchical Bayes procedures will be shown in the next section.

4.3 Hierarchical Bayes Predictive Density Estimation

In this section, we consider predictive density estimation for the multinomial distribution. Let $L \in \mathbb{N}$ and $d^{(1)}, \dots, d^{(L)} \in \{1, \dots, N\}$. For $\lambda = 1, \dots, L$, let $\nu_1^{(\lambda)}, \dots, \nu_{d^{(\lambda)}}^{(\lambda)} \in \mathbb{N}$ be such that $1 \leq \nu_1^{(\lambda)} < \cdots < \nu_{d^{(\lambda)}}^{(\lambda)} \leq N$ and let $I_0^{(\lambda)} = \{0, 1, \dots, m_{\nu_1^{(\lambda)}}\} \times \cdots \times \{0, 1, \dots, m_{\nu_{d^{(\lambda)}}^{(\lambda)}}\}$ and $\mathcal{W}^{(\lambda)} = \big\{(\mathring{w}_{\boldsymbol{i}})_{\boldsymbol{i} \in I_0^{(\lambda)}} \big| \mathring{w}_{\boldsymbol{i}} \in \mathbb{N}_0 \quad \text{for all } \boldsymbol{i} \in I_0^{(\lambda)} \quad \text{and} \quad \sum_{\boldsymbol{i} \in I_0^{(\lambda)}} \mathring{w}_{\boldsymbol{i}} = l^{(\lambda)}\big\}$. Now let $l^{(1)}, \dots, l^{(L)} \in \mathbb{N}$ and let $\boldsymbol{W}^{(1)}, \dots, \boldsymbol{W}^{(L)}$ be independent multinomial variables such that for $\lambda = 1, \dots, L$, the probability mass function of $\boldsymbol{W}^{(\lambda)}$ is given by

$$f_\lambda(\boldsymbol{w}^{(\lambda)}|\boldsymbol{p}) = \frac{l^{(\lambda)}!}{\prod_{\boldsymbol{i} \in I_0^{(\lambda)}} w_{\boldsymbol{i}}^{(\lambda)}!} \prod_{\boldsymbol{i}=(i_h)_{h=1}^{d^{(\lambda)}} \in I_0^{(\lambda)}} \Big\{ \prod_{h=1}^{d^{(\lambda)}} p_{i_h, \nu_h^{(\lambda)}} \Big\}^{w_{\boldsymbol{i}}^{(\lambda)}}$$

for $\boldsymbol{w}^{(\lambda)} = (w_{\boldsymbol{i}}^{(\lambda)})_{\boldsymbol{i} \in I_0^{(\lambda)}} \in \mathcal{W}^{(\lambda)}$. We consider the problem of estimating the joint probability mass of $\boldsymbol{W}^{(1)}, \dots, \boldsymbol{W}^{(L)}$, namely $f(\boldsymbol{w}|\boldsymbol{p}) = \prod_{\lambda=1}^L f_\lambda(\boldsymbol{w}^{(\lambda)}|\boldsymbol{p})$, $\boldsymbol{w} = (\boldsymbol{w}^{(\lambda)})_{\lambda=1,\dots,L} \in \mathcal{W} = \mathcal{W}^{(1)} \times \cdots \times \mathcal{W}^{(L)}$, on the basis of $\boldsymbol{X}$ given in the previous section under the Kullback-Leibler divergence. The risk function of a predictive mass $\hat{f}(\cdot; \boldsymbol{X})$ is given by

$$E\Big[\log \frac{f(\boldsymbol{W}|\boldsymbol{p})}{\hat{f}(\boldsymbol{W}; \boldsymbol{X})}\Big],$$

where $\boldsymbol{W} = (\boldsymbol{W}^{(\lambda)})_{\lambda=1,\dots,L} = ((W_{\boldsymbol{i}}^{(\lambda)})_{\boldsymbol{i} \in I_0^{(\lambda)}})_{\lambda=1,\dots,L}$.

As noted in Remark 2.2 of Hamura and Kubokawa (2019a), defining a natural plug-in predictive mass is not necessarily easy. Therefore, in this section, we seek a good Bayesian predictive mass. As shown by Aitchison (1975), the Bayesian predictive mass $\hat{f}^{(\pi)}(\cdot; \boldsymbol{X})$ associated with a prior $\boldsymbol{p} \sim \pi(\boldsymbol{p})$ is given by

$$\hat{f}^{(\pi)}(\boldsymbol{w}; \boldsymbol{X}) = E_\pi[f(\boldsymbol{w}|\boldsymbol{p})|\boldsymbol{X}]. \tag{4.3.1}$$

We first consider the natural conjugate Dirichlet distribution with density

$$\pi_{\boldsymbol{a}_0, \boldsymbol{a}}(\boldsymbol{p}) \propto \prod_{\nu=1}^N \Big({p_{0,\nu}}^{a_{0,\nu}-1} \prod_{i=1}^{m_\nu} {p_{i,\nu}}^{a_{i,\nu}-1} \Big), \tag{4.3.2}$$

where $\boldsymbol{a}_0 = (a_{0,\nu})_{\nu=1}^N \in \mathbb{R}^N$, $\boldsymbol{a} = (\boldsymbol{a}_\nu)_{\nu=1,\dots,N} = ((a_{i,\nu})_{i=1}^{m_\nu})_{\nu=1,\dots,N} \in (0,\infty)^{m_1} \times \cdots \times (0,\infty)^{m_N}$, and $a_{\cdot,\nu} = \sum_{i=1}^{m_\nu} a_{i,\nu}$ for $\nu = 1, \dots, N$. The Jeffreys prior is a special case of the Dirichlet prior.

Lemma 4.3.1 *The Dirichlet prior (4.3.2) with* $\boldsymbol{a}_0 = ((1-m_\nu)/2)_{\nu=1}^N$ *and* $\boldsymbol{a} = (\boldsymbol{j}^{(m_\nu)}/2)_{\nu=1,\dots,N}$ *is the Jeffreys prior.*

Next we consider the following conjugate shrinkage prior. Let

$$\pi_{\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a}}(\boldsymbol{p}) = \int_0^\infty u^{\alpha-1} e^{-\beta u} \Big\{ \prod_{\nu=1}^N \Big({p_{0,\nu}}^{\gamma_\nu u + a_{0,\nu} - 1} \prod_{i=1}^{m_\nu} {p_{i,\nu}}^{a_{i,\nu}-1} \Big) \Big\} du, \tag{4.3.3}$$

where $\alpha > 0$, $\beta > 0$, and $\boldsymbol{\gamma} = (\gamma_\nu)_{\nu=1}^N \in (0,\infty)^N$. This shrinkage prior is based on that of Section 3 of Hamura and Kubokawa (2020b) and is a slightly simplified version of the one mentioned in the discussion of their papar.

Under the prior (4.3.2), the posterior distribution of $\boldsymbol{p}$ given $\boldsymbol{X} = \boldsymbol{x}$ is proper for all $\boldsymbol{x} \in {\mathbb{N}_0}^{m_1} \times \cdots \times {\mathbb{N}_0}^{m_N}$ if and only if $r_\nu + a_{0,\nu} > 0$ for all $\nu = 1, \ldots, N$. Also, this condition implies that the posterior under (4.3.3) is proper, since we have assumed that $\beta \neq 0$ for simplicity.

In order to derive the Bayesian predictive mass with respect to (4.3.2) and that with respect to (4.3.3) in Proposition 4.3.1, we first rewrite $f(\boldsymbol{w}|\boldsymbol{p})$. Let $S(\lambda) = \{\nu_1^{(\lambda)}, \ldots, \nu_{d^{(\lambda)}}^{(\lambda)}\}$ for $\lambda = 1, \ldots, L$. For $\nu = 1, \ldots, N$, let $\Lambda(\nu) = \{\lambda \in \{1, \ldots, L\} | \nu \in S(\lambda)\}$ and, for $\lambda \in \Lambda(\nu)$, let $\{h_\nu^{(\lambda)}\} = \{h \in \{1, \ldots, d^{(\lambda)}\} | \nu = \nu_h^{(\lambda)}\}$ and let, for $i = 0, 1, \ldots, m_\nu$, $I_0^{(\lambda)}(i,\nu) = \{(i_h)_{h=1}^{d^{(\lambda)}} \in I_0^{(\lambda)} | i_{h_\nu^{(\lambda)}} = i\}$.

Lemma 4.3.2 *For any* $\boldsymbol{w} = ((w_{\boldsymbol{i}}^{(\lambda)})_{\boldsymbol{i} \in I_0^{(\lambda)}})_{\lambda=1,\ldots,L} \in \mathcal{W}$, *we have*

$$f(\boldsymbol{w}|\boldsymbol{p}) = \Big\{ \prod_{\lambda=1}^L \frac{l^{(\lambda)}!}{\prod_{\boldsymbol{i} \in I_0^{(\lambda)}} w_{\boldsymbol{i}}^{(\lambda)}!} \Big\} \prod_{\nu=1}^N \prod_{i=0}^{m_\nu} p_{i,\nu}^{\sum_{\lambda \in \Lambda(\nu)} \sum_{\boldsymbol{i} \in I_0^{(\lambda)}(i,\nu)} w_{\boldsymbol{i}}^{(\lambda)}}.$$

Let

$$C(\boldsymbol{w}) = \prod_{\lambda=1}^L \frac{l^{(\lambda)}!}{\prod_{\boldsymbol{i} \in I_0^{(\lambda)}} w_{\boldsymbol{i}}^{(\lambda)}!}$$

for $\boldsymbol{w} = ((w_{\boldsymbol{i}}^{(\lambda)})_{\boldsymbol{i} \in I_0^{(\lambda)}})_{\lambda=1,\ldots,L} \in \mathcal{W}$. For $(i,\nu) \in \mathbb{N}_0 \times \{1, \ldots, N\}$ with $i \leq m_\nu$, let

$$s_{i,\nu}(\boldsymbol{w}) = \sum_{\lambda \in \Lambda(\nu)} \sum_{\boldsymbol{i} \in I_0^{(\lambda)}(i,\nu)} w_{\boldsymbol{i}}^{(\lambda)}$$

for $\boldsymbol{w} = ((w_{\boldsymbol{i}}^{(\lambda)})_{\boldsymbol{i} \in I_0^{(\lambda)}})_{\lambda=1,\ldots,L} \in \mathcal{W}$. Using (4.3.1) and Lemma 4.3.2, the following expressions for $\hat{f}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot;\boldsymbol{X})$ and $\hat{f}^{(\pi_{\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot;\boldsymbol{X})$ are obtained.

Proposition 4.3.1 *Suppose that* $r_\nu + a_{0,\nu} > 0$ *for all* $\nu = 1, \ldots, N$.

(i) *The Bayesian predictive mass* $\hat{f}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot;\boldsymbol{X})$ *is given by*

$$\hat{f}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})}(\boldsymbol{w};\boldsymbol{X}) = C(\boldsymbol{w}) \frac{\displaystyle\prod_{\nu=1}^N \frac{\Gamma(s_{0,\nu}(\boldsymbol{w}) + r_\nu + a_{0,\nu}) \prod_{i=1}^{m_\nu} \Gamma(s_{i,\nu}(\boldsymbol{w}) + X_{i,\nu} + a_i)}{\Gamma\big(\sum_{\lambda \in \Lambda(\nu)} l^{(\lambda)} + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu}\big)}}{\displaystyle\prod_{\nu=1}^N \frac{\Gamma(r_\nu + a_{0,\nu}) \prod_{i=1}^{m_\nu} \Gamma(X_{i,\nu} + a_i)}{\Gamma(r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu})}}.$$

(ii) *The Bayesian predictive mass* $\hat{f}^{(\pi_{\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot;\boldsymbol{X})$ *is given by*

$$\begin{aligned}
&\hat{f}^{(\pi_{\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a}})}(\boldsymbol{w};\boldsymbol{X}) \\
&= C(\boldsymbol{w}) \frac{\displaystyle\int_0^\infty u^{\alpha-1} e^{-\beta u} \Big\{ \prod_{\nu=1}^N \frac{\Gamma(\gamma_\nu u + s_{0,\nu}(\boldsymbol{w}) + r_\nu + a_{0,\nu}) \prod_{i=1}^{m_\nu} \Gamma(s_{i,\nu}(\boldsymbol{w}) + X_{i,\nu} + a_i)}{\Gamma\big(\gamma_\nu u + \sum_{\lambda \in \Lambda(\nu)} l^{(\lambda)} + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu}\big)} \Big\} du}{\displaystyle\int_0^\infty u^{\alpha-1} e^{-\beta u} \Big\{ \prod_{\nu=1}^N \frac{\Gamma(\gamma_\nu u + r_\nu + a_{0,\nu}) \prod_{i=1}^{m_\nu} \Gamma(X_{i,\nu} + a_i)}{\Gamma(\gamma_\nu u + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu})} \Big\} du}.
\end{aligned}$$

We now compare the risk functions of $\hat{f}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot;\boldsymbol{X})$ and $\hat{f}^{(\pi_{\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot;\boldsymbol{X})$.

Theorem 4.3.1 *Assume that $r_\nu + a_{0,\nu} > 0$ for all $\nu = 1,\dots,N$. Assume that $r_\nu \geq 1$ for all $\nu = 1,\dots,N$. Suppose that*

$$\Big\{\frac{(\alpha+1)\gamma_\nu}{\beta+\gamma_\nu} - a_{\cdot,\nu}\Big\}(r_\nu - 1) \leq x_\nu\Big\{-\frac{(\alpha+1)\gamma_\nu}{\beta+\gamma_\nu} - \sum_{\lambda\in\Lambda(\nu)} l^{(\lambda)} - a_{0,\nu}\Big\} \tag{4.3.4}$$

for all $x_\nu \in \mathbb{N}$ for all $\nu = 1,\dots,N$. Then $\hat{f}^{(\pi_{\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot;\boldsymbol{X})$ dominates $\hat{f}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot;\boldsymbol{X})$.

Corollary 4.3.1 *If $1 \leq r_\nu > (m_\nu - 1)/2 > \sum_{\lambda\in\Lambda(\nu)} l^{(\lambda)}$ for all $\nu = 1,\dots,N$, then the Bayesian predictive mass with respect to the Jeffreys prior, namely $\hat{f}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot;\boldsymbol{X})$ with $\boldsymbol{a}_0 = ((1-m_\nu)/2)_{\nu=1}^N$ and $\boldsymbol{a} = (\boldsymbol{j}^{(m_\nu)}/2)_{\nu=1,\dots,N}$, is inadmissible and dominated by the Bayesian predictive mass $\hat{f}^{(\pi_{\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot;\boldsymbol{X})$ with $\boldsymbol{a}_0 = ((1-m_\nu)/2)_{\nu=1}^N$ and $\boldsymbol{a} = (\boldsymbol{j}^{(m_\nu)}/2)_{\nu=1,\dots,N}$ for some $\alpha > 0$, $\beta > 0$, and $\boldsymbol{\gamma} \in (0,\infty)^N$.*

4.4 Simulation Studies

4.4.1 Simulation study for the model in Section 4.2

In this section, we investigate through simulation the numerical performance of the risk functions of point estimators of $\boldsymbol{p}$ under the standardized squared error loss given by (4.2.1). Although there are a number of conceivable unbalanced settings, for the sake of simplicity, we only consider some of the most uncomplicated cases. In particular, we set $n = N = 2$, $m_1 = m_2 = 7$, and $\boldsymbol{c} = (\boldsymbol{j}^{(7)}, \boldsymbol{j}^{(7)})$ and focus on the effect of r_1, r_2, and $\boldsymbol{p}$. As in the Poisson case (see, for example, Hamura and Kubokawa (2019b, 2020c)), although the dominance conditions given in Section 4.2 tend to be restrictive and may not be satisfied especially when r_1 and r_2 are highly unbalanced, our proposed estimator turns out to perform well in such cases also.

We compare the UMVU estimator $\hat{\boldsymbol{p}}^{\mathrm{U}}$ given by (4.2.2) and the empirical Bayes estimator $\hat{\boldsymbol{p}}^{(\tilde{\boldsymbol{a}})}$ given in (4.2.9) with $\tilde{\boldsymbol{a}} = \boldsymbol{j}^{(N)}$, namely

$$\hat{\boldsymbol{p}}^{\mathrm{EB}} = \Big(\Big(\frac{X_{i,\nu}}{r_\nu + X_{\cdot,\nu} - 1 + 7 + 7\sum_{\nu'=1}^2 r_{\nu'}/X_{\cdot,\cdot}}\Big)_{i=1}^7\Big)_{\nu=1,2}.$$

Let $\boldsymbol{p}_0(0) = (1,1,1,1,1,1,1)^\top/8$, $\boldsymbol{p}_0(1) = (1,1,1,1,10,10,10)^\top/44$, and $\boldsymbol{p}_0(2) = (10,10,10,10,1,1,1)^\top/44$. We consider the following cases:

(i) Let $r_1 = r_2 = 12$ and let $\boldsymbol{p}_1 = \boldsymbol{p}_2 = (1-\omega)\boldsymbol{p}_0(0) + \omega\boldsymbol{p}_0(1)$ for $\omega = 0, 1/5, \dots, 4/5, 1$.

(ii) Let $r_1 = r_2 = 12$ and let $\boldsymbol{p}_1 = (1-\omega)\boldsymbol{p}_0(0) + \omega\boldsymbol{p}_0(1)$ and $\boldsymbol{p}_2 = (1-\omega)\boldsymbol{p}_0(0) + \omega\boldsymbol{p}_0(2)$ for $\omega = 0, 1/5, \dots, 4/5, 1$.

(iii) Let $r_1 = 8$ and $r_2 = 16$ and let $\boldsymbol{p}_1 = \boldsymbol{p}_2 = (1-\omega)\boldsymbol{p}_0(0) + \omega\boldsymbol{p}_0(1)$ for $\omega = 0, 1/5, \dots, 4/5, 1$.

(iv) Let $r_1 = 8$ and $r_2 = 16$ and let $\boldsymbol{p}_1 = (1-\omega)\boldsymbol{p}_0(0) + \omega\boldsymbol{p}_0(1)$ and $\boldsymbol{p}_2 = (1-\omega)\boldsymbol{p}_0(0) + \omega\boldsymbol{p}_0(2)$ for $\omega = 0, 1/5, \dots, 4/5, 1$.

In Cases (i) and (ii), r_1 and r_2 are balanced. On the other hand, they are highly unbalanced in Cases (iii) and (iv). The parameter vectors $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are identical for all $\omega = 0, 1/5, \ldots, 4/5, 1$ in Cases (i) and (iii) and distinct for $\omega = 1/5, \ldots, 4/5, 1$ in Cases (ii) and (iv). We obtain approximated values of the risk functions of $\hat{\boldsymbol{p}}^{\mathrm{U}}$ and $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ by simulation with $100,000$ replications.

The results are illustrated in Figure 4.1. It seems that $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ dominates $\hat{\boldsymbol{p}}^{\mathrm{U}}$ in every case. In Cases (i) and (iii), both $\hat{\boldsymbol{p}}^{\mathrm{U}}$ and $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ have large values of risks for large ω. In Case (ii), the risk values of $\hat{\boldsymbol{p}}^{\mathrm{U}}$ are almost the same while those of $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ are small for large ω. On the other hand, in Case (iv), where the amount of information from $\boldsymbol{X}_2$ is much larger than the amount of information from $\boldsymbol{X}_1$, the results are similar to those in Cases (i) and (iii). Overall, the risk values are smaller in Cases (i) and (ii) than in Cases (iii) and (iv) and larger in Cases (i) and (iii) than in Cases (ii) and (iv).

4.4.2 Simulation study for the model in Section 4.3

This section corresponds to Section 4.3. As in Section 4.4.1, we focus on simple cases and in particular consider low-dimensional settings for computational convenience. We set $N = 2$, $m_1 = m_2 = 3$, $L = 2$, $d^{(1)} = 1$, $d^{(2)} = 2$, $\nu_1^{(1)} = 1$, $\nu_1^{(2)} = 1$, $\nu_2^{(2)} = 2$, and $l^{(1)} = l^{(2)} = 1$. We note that $\boldsymbol{p}_1$ is related to both the vector $\boldsymbol{W}^{(1)}$ and the matrix $\boldsymbol{W}^{(2)}$. We investigate through simulation the numerical performance of the risk functions of $\hat{f}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot;\boldsymbol{X})$ given in part (i) of Proposition 4.3.1 and $\hat{f}^{(\pi_{\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot;\boldsymbol{X})$ given in part (ii) of Proposition 4.3.1; more specifically, we set $\boldsymbol{a}_0 = (-1,-1)^\top$, $\boldsymbol{a} = (\boldsymbol{j}^{(3)}/2, \boldsymbol{j}^{(3)}/2)$, $\alpha = 1$, $\beta = 1$, and $\gamma = (1,1)^\top$ and compare the Bayesian predictive mass with respect to the Jeffreys prior, namely $\hat{f}^{\mathrm{J}}(\cdot;\boldsymbol{X}) = \hat{f}^{(\pi_{(-1,-1)^\top,(\boldsymbol{j}^{(3)}/2,\boldsymbol{j}^{(3)}/2)})}(\cdot;\boldsymbol{X})$, and the Bayesian predictive mass $\hat{f}^{\mathrm{HB}}(\cdot;\boldsymbol{X}) = \hat{f}^{(\pi_{1,1,(1,1)^\top,(-1,-1)^\top,(\boldsymbol{j}^{(3)}/2,\boldsymbol{j}^{(3)}/2)})}(\cdot;\boldsymbol{X})$. Let $\boldsymbol{p}(0) = ((1,1,1)^\top/4, (1,1,1)^\top/4)$, $\boldsymbol{p}(1) = ((1,1,2)^\top/6, (1,1,2)^\top/6)$, and $\boldsymbol{p}(2) = ((1,1,2)^\top/6, (2,2,1)^\top/6)$. For each $\boldsymbol{p} = \boldsymbol{p}(0), \boldsymbol{p}(1), \boldsymbol{p}(2)$, we consider the following cases: (I) $r_1 = r_2 = 5$; (II) $r_1 = 4$ and $r_2 = 6$; (III) $r_1 = 6$ and $r_2 = 4$.

We obtain approximated values of the risk functions of $\hat{f}^{\mathrm{J}}(\cdot;\boldsymbol{X})$ and $\hat{f}^{\mathrm{HB}}(\cdot;\boldsymbol{X})$ by simulation with $1,000$ replications. The Bayesian predictive mass $\hat{f}^{\mathrm{J}}(\cdot;\boldsymbol{X})$ is computed by generating $2,000$ independent posterior samples while $\hat{f}^{\mathrm{HB}}(\cdot;\boldsymbol{X})$ is computed based on a Gibbs sampler by generating $20,000$ approximate posterior samples after discarding the first $10,000$ samples. The percentage relative improvement in average loss (PRIAL) of $\hat{f}^{\mathrm{HB}}(\cdot;\boldsymbol{X})$ over $\hat{f}^{\mathrm{J}}(\cdot;\boldsymbol{X})$ is defined by

$$\mathrm{PRIAL} = 100\Big\{E\Big[\log\frac{f(\boldsymbol{W}|\boldsymbol{p})}{\hat{f}^{\mathrm{J}}(\boldsymbol{W};\boldsymbol{X})}\Big] - E\Big[\log\frac{f(\boldsymbol{W}|\boldsymbol{p})}{\hat{f}^{\mathrm{HB}}(\boldsymbol{W};\boldsymbol{X})}\Big]\Big\}/E\Big[\log\frac{f(\boldsymbol{W}|\boldsymbol{p})}{\hat{f}^{\mathrm{J}}(\cdot;\boldsymbol{X})}\Big].$$

Table 4.1 reports values of the risks of $\hat{f}^{\mathrm{J}}(\cdot;\boldsymbol{X})$ and $\hat{f}^{\mathrm{HB}}(\cdot;\boldsymbol{X})$ with values of PRIAL given in parentheses. It can be seen from the values of PRIAL that $\hat{f}^{\mathrm{HB}}(\cdot;\boldsymbol{X})$ has smaller values of risks than $\hat{f}^{\mathrm{J}}(\cdot;\boldsymbol{X})$ in every case. When $\boldsymbol{p} = \boldsymbol{p}(0), \boldsymbol{p}(2)$, PRIAL is smallest in Case (II) and largest in Case (III). On the other hand, when $\boldsymbol{p} = \boldsymbol{p}(1)$, $\hat{f}^{\mathrm{HB}}(\cdot;\boldsymbol{X})$ has the largest and smallest values of PRIAL in Cases (II) and (III), respectively.

4.5 Discussion

In this chapter, we considered the problems of estimating negative multinomial parameter vectors and the joint predictive density of multinomial tables on the basis of observations of negative

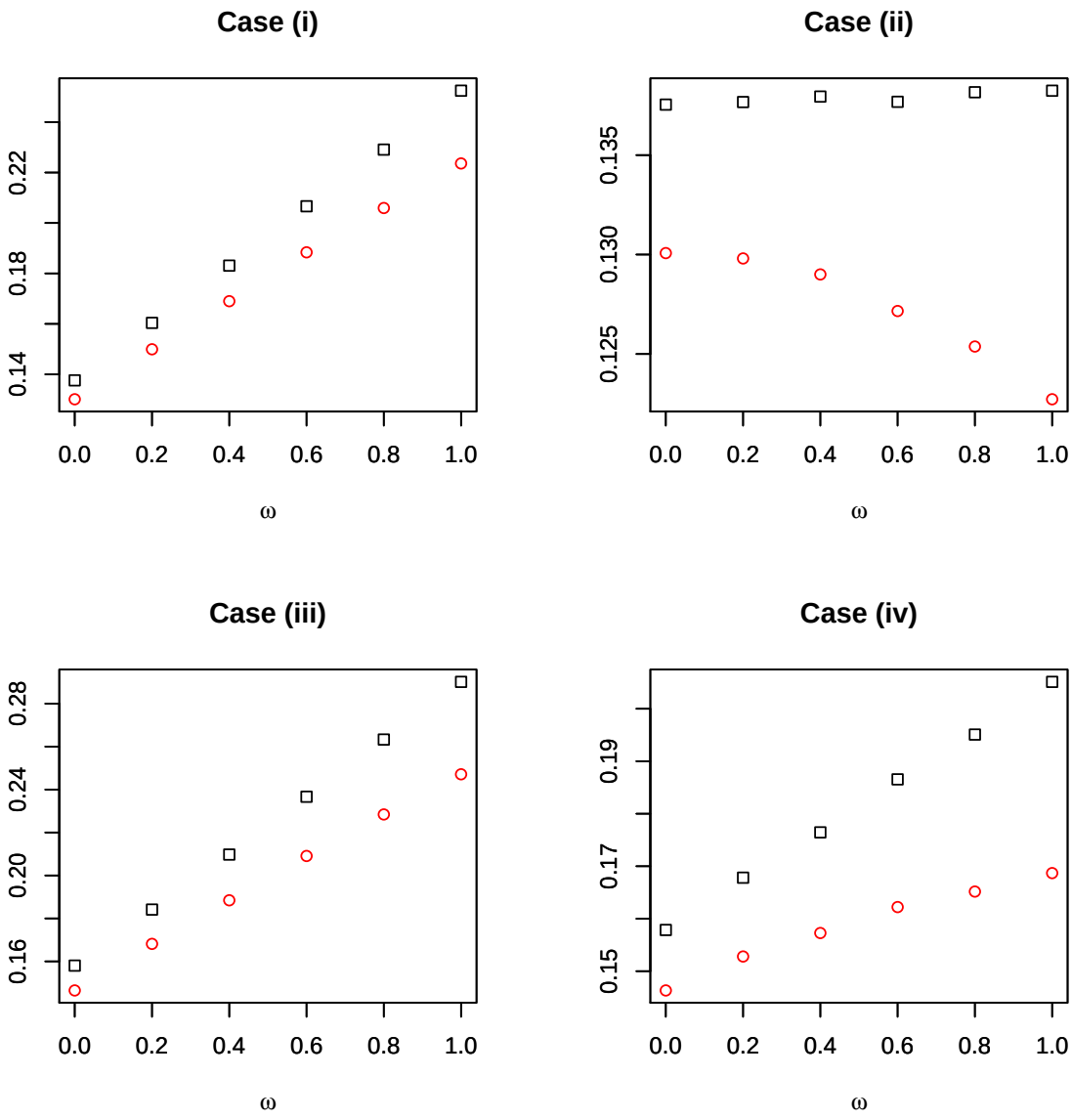

Figure 4.1: Risks of the estimators $\hat{\boldsymbol{p}}^{\mathrm{U}}$ and $\hat{\boldsymbol{p}}^{\mathrm{EB}}$ for $\omega = 0, 1/5, \ldots, 4/5, 1$ in Cases (i), (ii), (iii), and (iv). The black squares and red circles correspond to $\hat{\boldsymbol{p}}^{\mathrm{U}}$ and $\hat{\boldsymbol{p}}^{\mathrm{EB}}$, respectively.

Table 4.1: Risks of $\hat{f}^{\mathrm{J}}(\cdot;\boldsymbol{X})$ (J) and $\hat{f}^{\mathrm{HB}}(\cdot;\boldsymbol{X})$ (HB). Values of PRIAL of HB are given in parentheses.

Case	$\boldsymbol{p}$	J	HB
(I)	$\boldsymbol{p}(0)$	0.22	0.22 (1.13)
(I)	$\boldsymbol{p}(1)$	0.23	0.23 (1.08)
(I)	$\boldsymbol{p}(2)$	0.27	0.27 (1.40)
(II)	$\boldsymbol{p}(0)$	0.28	0.27 (1.00)
(II)	$\boldsymbol{p}(1)$	0.32	0.31 (2.78)
(II)	$\boldsymbol{p}(2)$	0.30	0.30 (1.35)
(III)	$\boldsymbol{p}(0)$	0.23	0.23 (1.34)
(III)	$\boldsymbol{p}(1)$	0.30	0.29 (0.52)
(III)	$\boldsymbol{p}(2)$	0.25	0.24 (2.02)

multinomial variables in unbalanced settings. A related problem of mathematical interest is that of estimating the joint predictive density of future negative multinomial variables on the basis of the current negative multinomial observations. Although no dominance result has been obtained, we here derive identities which relate prediction to estimation in the negative multinomial case.

Let $s_1,\dots,s_n>0$ and let $\boldsymbol{Y}_\nu=(Y_{i,\nu})_{i=1}^{m_\nu}$, $\nu=1,\dots,n$, be independent negative multinomial variables with mass functions

$$g_\nu(\boldsymbol{y}_\nu|\boldsymbol{p}_\nu)=\frac{\Gamma\bigl(s_\nu+\sum_{i=1}^{m_\nu}y_{i,\nu}\bigr)}{\Gamma(s_\nu)\prod_{i=1}^{m_\nu}y_{i,\nu}!}{p_{0,\nu}}^{s_\nu}\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{y_{i,\nu}}, \tag{4.5.1}$$

$\boldsymbol{y}_\nu=(y_{i,\nu})_{i=1}^{m_\nu}\in{\mathbb{N}_0}^{m_\nu}$, $\nu=1,\dots,n$, respectively. Consider the problem of estimating the predictive density $g(\boldsymbol{y}|\boldsymbol{p})=\prod_{\nu=1}^n g_\nu(\boldsymbol{y}_\nu|\boldsymbol{p}_\nu)$, $\boldsymbol{y}=(\boldsymbol{y}_\nu)_{\nu=1,\dots,n}\in{\mathbb{N}_0}^{m_1}\times\cdots\times{\mathbb{N}_0}^{m_n}$, on the basis of $\boldsymbol{X}$ given in Section 4.2 under the Kullback-Leibler divergence. As shown by Aitchison (1975), the Bayesian predictive mass $\hat{g}^{(\pi)}(\cdot;\boldsymbol{X})$ with respect to a prior $\boldsymbol{p}\sim\pi(\boldsymbol{p})$ is given by

$$\begin{aligned}\hat{g}^{(\pi)}(\boldsymbol{y};\boldsymbol{X})&=E_\pi[g(\boldsymbol{y}|\boldsymbol{p})|\boldsymbol{X}]\\&=\Bigl\{\prod_{\nu=1}^n\frac{\Gamma\bigl(s_\nu+\sum_{i=1}^{m_\nu}y_{i,\nu}\bigr)}{\Gamma(s_\nu)\prod_{i=1}^{m_\nu}y_{i,\nu}!}\Bigr\}\frac{\int_D\pi(\boldsymbol{p})\bigl\{\prod_{\nu=1}^N\bigl({p_{0,\nu}}^{s_\nu+r_\nu}\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{y_{i,\nu}+X_{i,\nu}}\bigr)\bigr\}d\boldsymbol{p}}{\int_D\pi(\boldsymbol{p})\bigl\{\prod_{\nu=1}^N\bigl({p_{0,\nu}}^{r_\nu}\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{X_{i,\nu}}\bigr)\bigr\}d\boldsymbol{p}},\end{aligned} \tag{4.5.2}$$

where $s_\nu=y_{1,\nu}=\cdots=y_{m_\nu,\nu}=0$ if $\nu\in\{1,\dots,N\}\cap[n+1,\infty)$, and has risk given by

$$R(\boldsymbol{p},\hat{g}^{(\pi)})=E\Bigl[\log\frac{g(\boldsymbol{Y}|\boldsymbol{p})}{\hat{g}^{(\pi)}(\boldsymbol{Y};\boldsymbol{X})}\Bigr]. \tag{4.5.3}$$

Let $t_1,\dots,t_N\colon[0,1]\to(0,\infty)$ be smooth, nondecreasing functions such that for all $\nu=1,\dots,N$,

$$t_\nu(0)=r_\nu\quad\text{and}\quad t_\nu(1)=\begin{cases}r_\nu+s_\nu, & \text{if } \nu\le n,\\ r_\nu, & \text{if } \nu\ge n+1.\end{cases} \tag{4.5.4}$$

For each $\tau \in [0,1]$, let $\boldsymbol{Z}_\nu(\tau) = (Z_{i,\nu}(\tau))_{i=1}^{m_\nu}$, $\nu = 1, \dots, N$, be independent negative multinomial variables with mass functions

$$\frac{\Gamma\big(t_\nu(\tau) + \sum_{i=1}^{m_\nu} z_{i,\nu}\big)}{\Gamma(t_\nu(\tau)) \prod_{i=1}^{m_\nu} z_{i,\nu}!} {p_{0,\nu}}^{t_\nu(\tau)} \prod_{i=1}^{m_\nu} {p_{i,\nu}}^{z_{i,\nu}},$$

$(z_{i,\nu})_{i=1}^{m_\nu} \in {\mathbb{N}_0}^{m_\nu}$, $\nu = 1, \dots, N$, respectively, and let $\boldsymbol{Z}(\tau) = (\boldsymbol{Z}_\nu(\tau))_{\nu=1,\dots,N}$. Let $\mathcal{W}_{\nu,k} = \big\{(\mathring{w}_i)_{i=1}^{m_\nu} \in {\mathbb{N}_0}^{m_\nu} \big| \sum_{i=1}^{m_\nu} \mathring{w}_i = k\big\}$ for $\nu = 1, \dots, N$ and $k \in \mathbb{N}_0$. Let

$$L^{\mathrm{KL}}(\tilde{d}, \theta) = \tilde{d} - \theta - \theta \log(\tilde{d}/\theta) \tag{4.5.5}$$

for $\tilde{d}, \theta \in (0, \infty)$. The following theorem shows that the risk function of an arbitrary Bayesian predictive mass can be expressed using the risk functions of the corresponding Bayes estimators of an infinite number of monomials of the unknown probabilities.

Theorem 4.5.1 *Let $\boldsymbol{p} \sim \pi(\boldsymbol{p})$ be a prior density. Then the risk of $\hat{g}^{(\pi)}(\cdot; \boldsymbol{X})$ is expressed as*

$$\begin{aligned} &R(\boldsymbol{p}, \hat{g}^{(\pi)}) \\ &= \int_0^1 \Big\{ \sum_{\nu=1}^{n} {t_\nu}'(\tau) \sum_{k=1}^{\infty} \frac{1}{k} \sum_{(w_i)_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,k}} \frac{k!}{\prod_{i=1}^{m_\nu} w_i!} E\Big[L^{\mathrm{KL}}\Big(E_\pi\Big[\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big| \boldsymbol{Z}(\tau) \Big], \prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big) \Big] \Big\} d\tau. \end{aligned}$$

Theorem 3 of Hamura and Kubokawa (2020b) is related to the monomials of degree 1 in the above expression. In the negative binomial case, the "intrinsic loss" derived by Robert (1996) is not given by (4.5.5); see Remark 2.2 of Hamura and Kubokawa (2019a) for details.

We also have the following somewhat simpler result. Let

$$\pi_{M,\tilde{\gamma},\boldsymbol{a}_0,\boldsymbol{a}}(\boldsymbol{p}) = \int_0^\infty \Big[\prod_{\nu=1}^{N} \Big\{ {p_{0,\nu}}^{\tilde{\gamma}_\nu(u) + a_{0,\nu} - 1} \prod_{i=1}^{m_\nu} {p_{i,\nu}}^{a_{i,\nu}-1} \Big\} \Big] dM(u), \tag{4.5.6}$$

where M is a measure on $(0, \infty)$ while $\widetilde{\gamma} = (\tilde{\gamma}_\nu)_{\nu=1}^N \colon (0,\infty) \to (0,\infty)^N$. Then Corollary 4.5.1 gives an expression for the risk difference between the Bayesian predictive mass with respect to the prior (4.5.6) and that with respect to the prior (4.3.2).

Corollary 4.5.1 *The risk difference between $\hat{g}^{(\pi_{M,\tilde{\gamma},\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot; \boldsymbol{X})$ and $\hat{g}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot; \boldsymbol{X})$ is expressed as*

$$\begin{aligned} &R(\boldsymbol{p}, \hat{g}^{(\pi_{M,\tilde{\gamma},\boldsymbol{a}_0,\boldsymbol{a}})}) - R(\boldsymbol{p}, \hat{g}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})}) \\ &= \int_0^1 \Big\{ \sum_{\nu=1}^{n} {t_\nu}'(\tau) \sum_{k=1}^{\infty} \frac{1}{k} E[L^{\mathrm{KL}}(E_{\pi_{M,\tilde{\gamma},\boldsymbol{a}_0,\boldsymbol{a}}}[{p_{\cdot,\nu}}^k | \boldsymbol{Z}(\tau)], {p_{\cdot,\nu}}^k) - L^{\mathrm{KL}}(E_{\pi_{\boldsymbol{a}_0,\boldsymbol{a}}}[{p_{\cdot,\nu}}^k | \boldsymbol{Z}(\tau)], {p_{\cdot,\nu}}^k)] \Big\} d\tau. \end{aligned}$$

Despite these identities, dominance conditions have not been obtained. It may be worth noting that $\log\{\hat{g}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})}(\boldsymbol{Y}; \boldsymbol{X}) / \hat{g}^{(\pi_{M,\tilde{\gamma},\boldsymbol{a}_0,\boldsymbol{a}})}(\boldsymbol{Y}; \boldsymbol{X})\}$, whose expectation is the risk difference, is a function only of $X_{\cdot,\nu}$, $\nu = 1, \dots, N$, and $Y_{\cdot,\nu} = \sum_{i=1}^{m_\nu} Y_{i,\nu}$, $\nu = 1, \dots, n$. Inadmissibility of $\hat{g}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})}(\cdot; \boldsymbol{X})$ could be studied in a future paper.

4.6 Appendix

4.6.1 Assumptions

Let $\bar{c}_\nu = \max_{1\le i\le m_\nu} c_{i,\nu}$ for $\nu = 1,\dots,N$. Let $\underline{\tilde{c}}_\nu^{(\nu)} = \min_{1\le i\le m_\nu} \tilde{c}_{i,\nu}^{(\nu)}$, $\overline{\tilde{c}}_\nu^{(\nu)} = \max_{1\le i\le m_\nu} \tilde{c}_{i,\nu}^{(\nu)}$, and $\widetilde{C}_\nu = (\overline{\tilde{c}}_\nu^{(\nu)}/\underline{\tilde{c}}_\nu^{(\nu)})/\{1+\tilde{b}_\nu(\underline{\tilde{c}}_\nu^{(\nu)}+\overline{\tilde{c}}_\nu^{(\nu)})\}$ for $\nu = 1,\dots,N$. Let $A = \max_{1\le\nu\le n}\bar{c}_\nu(\widetilde{C}_\nu+2)$, $\underline{\tilde{b}} = \min_{1\le\nu\le n}\tilde{b}_\nu$, $\overline{\tilde{b}} = \max_{1\le\nu\le n}\tilde{b}_\nu$, $\underline{\overline{\tilde{c}}} = \min_{1\le\nu\le n}\overline{\tilde{c}}_\nu^{(\nu)}$, and $\overline{\overline{\tilde{c}}} = \max_{1\le\nu\le n}\overline{\tilde{c}}_\nu^{(\nu)}$ and let $\tilde{c}_* = \min_{1\le\nu\le N}\min_{1\le\nu'\le N}\min_{1\le i\le m_{\nu'}}\tilde{c}_{i,\nu'}^{(\nu)}$ and $\tilde{c}^* = \max_{1\le\nu\le N}\max_{1\le\nu'\le N}\max_{1\le i\le m_{\nu'}}\tilde{c}_{i,\nu'}^{(\nu)}$. Let $A_1 = \max_{1\le\nu\le n}\{\bar{c}_\nu(3+4\tilde{b}_\nu\tilde{c})/(1+2\tilde{b}_\nu\tilde{c})\}$.

Assumption 4.6.1 and Assumption 4.6.2 correspond to Theorem 4.2.2 and Corollary 4.2.1, respectively.

Assumption 4.6.1

(a) $\overline{\overline{\tilde{c}}} > 0$.

(b) $r_\nu \ge \widetilde{C}_\nu + 1$ and $r_\nu + \tilde{b}_\nu \ge \widetilde{C}_\nu + 2$ for all $\nu = 1,\dots,n$ with $c_{\cdot,\nu} > 0$.

(c) $\underline{c}_\cdot - A \ge 0$.

(d) For all $x \in \mathbb{N}$, either

- $\overline{\overline{\tilde{c}}}\{\overline{\tilde{b}} + 1/(\tilde{c}_* x + \overline{\tilde{c}})\} - 2(\underline{r}/\overline{r})^2(\underline{c}_\cdot - A)\{\underline{\tilde{b}}\tilde{c}_*\underline{\overline{\tilde{c}}}/(\overline{\tilde{b}}\tilde{c}^*\overline{\overline{\tilde{c}}})\} \le 0$ implies

$$\underline{c}_\cdot\{\overline{\tilde{b}} + 1/(\tilde{c}_* x + \overline{\tilde{c}})\} - 2\Big(\frac{\underline{r}}{\overline{r}}\Big)^2(\underline{c}_\cdot - A)\frac{\underline{\tilde{b}}\tilde{c}_*\underline{\overline{\tilde{c}}}}{\overline{\overline{\tilde{b}}}\tilde{c}^*\overline{\overline{\tilde{c}}}}\{\underline{r} + \overline{\tilde{b}} + 1/(\tilde{c}_* x + \overline{\tilde{c}})\} \le 0 \quad \text{and}$$

- $\overline{\overline{\tilde{c}}}\{\overline{\tilde{b}} + 1/(\tilde{c}_* x + \overline{\tilde{c}})\} - 2(\underline{r}/\overline{r})^2(\underline{c}_\cdot - A)\{\underline{\tilde{b}}\tilde{c}_*\underline{\overline{\tilde{c}}}/(\overline{\tilde{b}}\tilde{c}^*\overline{\overline{\tilde{c}}})\} > 0$ implies

$$x\Big[\overline{\overline{\tilde{c}}}\{\overline{\tilde{b}} + 1/(\tilde{c}_* x + \overline{\tilde{c}})\} - 2\Big(\frac{\underline{r}}{\overline{r}}\Big)^2(\underline{c}_\cdot - A)\frac{\underline{\tilde{b}}\tilde{c}_*\underline{\overline{\tilde{c}}}}{\overline{\overline{\tilde{b}}}\tilde{c}^*\overline{\overline{\tilde{c}}}}\Big]$$
$$+ n\underline{c}_\cdot\{\overline{\tilde{b}} + 1/(\tilde{c}_* x + \overline{\tilde{c}})\} - 2n\Big(\frac{\underline{r}}{\overline{r}}\Big)^2(\underline{c}_\cdot - A)\frac{\underline{\tilde{b}}\tilde{c}_*\underline{\overline{\tilde{c}}}}{\overline{\overline{\tilde{b}}}\tilde{c}^*\overline{\overline{\tilde{c}}}}\{\underline{r} + \overline{\tilde{b}} + 1/(\tilde{c}_* x + \overline{\tilde{c}})\} \le 0$$

or

- $\overline{\overline{\tilde{c}}}\{\overline{\tilde{b}} + 1/(\tilde{c}_* x + \overline{\tilde{c}})\} - 2(\underline{c}_\cdot - A)\{\underline{\tilde{b}}\tilde{c}_*\underline{\overline{\tilde{c}}}/(\overline{\tilde{b}}\tilde{c}^*\overline{\overline{\tilde{c}}})\} \le 0$ implies

$$(\underline{c}_\cdot - \overline{\overline{\tilde{c}}}\underline{r}) - 2(\underline{c}_\cdot - A)\frac{\underline{\tilde{b}}\tilde{c}_*\underline{\overline{\tilde{c}}}}{\overline{\overline{\tilde{b}}}\tilde{c}^*\overline{\overline{\tilde{c}}}} \le 0 \quad \text{and}$$

- $\overline{\overline{\tilde{c}}}\{\overline{\tilde{b}} + 1/(\tilde{c}_* x + \overline{\tilde{c}})\} - 2(\underline{c}_\cdot - A)\{\underline{\tilde{b}}\tilde{c}_*\underline{\overline{\tilde{c}}}/(\overline{\tilde{b}}\tilde{c}^*\overline{\overline{\tilde{c}}})\} > 0$ implies

$$\Big(\sum_{\nu=1}^n r_\nu + x\Big)\Big[\overline{\overline{\tilde{c}}}\{\overline{\tilde{b}} + 1/(\tilde{c}_* x + \overline{\tilde{c}})\} - 2(\underline{c}_\cdot - A)\frac{\underline{\tilde{b}}\tilde{c}_*\underline{\overline{\tilde{c}}}}{\overline{\overline{\tilde{b}}}\tilde{c}^*\overline{\overline{\tilde{c}}}}\Big]$$
$$+ n(\underline{c}_\cdot - \overline{\overline{\tilde{c}}}\underline{r})\{\overline{\tilde{b}} + 1/(\tilde{c}_* x + \overline{\tilde{c}})\} - 2n(\underline{c}_\cdot - A)\frac{\underline{\tilde{b}}\tilde{c}_*\underline{\overline{\tilde{c}}}}{\overline{\overline{\tilde{b}}}\tilde{c}^*\overline{\overline{\tilde{c}}}}\{\overline{\tilde{b}} + 1/(\tilde{c}_* x + \overline{\tilde{c}})\} \le 0.$$

Assumption 4.6.2

(a) $\overline{\tilde{c}} > 0$.

(b) $r_\nu \geq 1/(1 + 2\tilde{b}_\nu \tilde{c}) + 1$ and $r_\nu + \tilde{b}_\nu \geq 1/(1 + 2\tilde{b}_\nu \tilde{c}) + 2$ for all $\nu = 1, \ldots, n$ with $c_{\cdot,\nu} > 0$.

(c) $\underline{c}_\cdot - A_1 \geq 0$.

(d) For all $x \in \mathbb{N}$, either

- $\overline{\tilde{c}}[\overline{\tilde{b}} + 1/\{\tilde{c}(x+1)\}] - 2(\underline{r}/\overline{r})^2(\underline{c}_\cdot - A_1)\underline{\tilde{b}}/\overline{\tilde{b}} \leq 0$ implies

$$\underline{c}_\cdot[\overline{\tilde{b}} + 1/\{\tilde{c}(x+1)\}] - 2\Big(\frac{\underline{r}}{\overline{r}}\Big)^2(\underline{c}_\cdot - A_1)\frac{\underline{\tilde{b}}}{\overline{\tilde{b}}}[\underline{r} + \overline{\tilde{b}} + 1/\{\tilde{c}(x+1)\}] \leq 0 \quad \text{and}$$

- $\overline{\tilde{c}}[\overline{\tilde{b}} + 1/\{\tilde{c}(x+1)\}] - 2(\underline{r}/\overline{r})^2(\underline{c}_\cdot - A_1)\underline{\tilde{b}}/\overline{\tilde{b}} > 0$ implies

$$\begin{aligned} &x\Big(\overline{\tilde{c}}[\overline{\tilde{b}} + 1/\{\tilde{c}(x+1)\}] - 2\Big(\frac{\underline{r}}{\overline{r}}\Big)^2(\underline{c}_\cdot - A_1)\frac{\underline{\tilde{b}}}{\overline{\tilde{b}}}\Big) \\ &+ n\underline{c}_\cdot[\overline{\tilde{b}} + 1/\{\tilde{c}(x+1)\}] - 2n\Big(\frac{\underline{r}}{\overline{r}}\Big)^2(\underline{c}_\cdot - A_1)\frac{\underline{\tilde{b}}}{\overline{\tilde{b}}}[\underline{r} + \overline{\tilde{b}} + 1/\{\tilde{c}(x+1)\}] \leq 0 \end{aligned}$$

or

- $\overline{\tilde{c}}[\overline{\tilde{b}} + 1/\{\tilde{c}(x+1)\}] - 2(\underline{c}_\cdot - A_1)\underline{\tilde{b}}/\overline{\tilde{b}} \leq 0$ implies

$$(\underline{c}_\cdot - \overline{\tilde{c}}\underline{r}) - 2(\underline{c}_\cdot - A_1)\frac{\underline{\tilde{b}}}{\overline{\tilde{b}}} \leq 0 \quad \text{and}$$

- $\overline{\tilde{c}}[\overline{\tilde{b}} + 1/\{\tilde{c}(x+1)\}] - 2(\underline{c}_\cdot - A_1)\underline{\tilde{b}}/\overline{\tilde{b}} > 0$ implies

$$\begin{aligned} &\Big(\sum_{\nu=1}^n r_\nu + x\Big)\Big(\overline{\tilde{c}}[\overline{\tilde{b}} + 1/\{\tilde{c}(x+1)\}] - 2(\underline{c}_\cdot - A_1)\frac{\underline{\tilde{b}}}{\overline{\tilde{b}}}\Big) \\ &+ n(\underline{c}_\cdot - \overline{\tilde{c}}\underline{r})[\overline{\tilde{b}} + 1/\{\tilde{c}(x+1)\}] - 2n(\underline{c}_\cdot - A_1)\frac{\underline{\tilde{b}}}{\overline{\tilde{b}}}[\overline{\tilde{b}} + 1/\{\tilde{c}(x+1)\}] \leq 0. \end{aligned}$$

4.6.2 Proofs

Here we prove Theorems 4.2.1, 4.2.2, 4.3.1, and 4.5.1, Lemma 4.3.2, and Corollary 4.5.1. We use Lemma 4.6.1, which is due to Hudson (1978).

For $(i,\nu), (i',\nu') \in \mathbb{N} \times \{1, \ldots, N\}$ with $i \leq m_\nu$ and $i' \leq m_{\nu'}$, let $\delta_{i,i',\nu,\nu'} = 1$ if $i = i'$ and $\nu = \nu'$ and $= 0$ otherwise. Let $\boldsymbol{X}_\cdot = (X_{\cdot,\nu})_{\nu=1}^N$. For $\nu = 1, \ldots, N$, let $\boldsymbol{e}_\nu^{(N)}$ be the νth unit vector in $\mathbb{R}^N$, namely the νth column of the $N \times N$ identity matrix. For $\nu = 1, \ldots, N$, let $\mathbf{0}^{(m_\nu)} = (0, \ldots, 0)^\top \in \mathbb{R}^{m_\nu}$. For $\nu, \nu' = 1, \ldots, N$, let $\delta_{\nu,\nu'}^{(N)} = \boldsymbol{e}_\nu^{(N)\top}\boldsymbol{e}_{\nu'}^{(N)}$.

Lemma 4.6.1 *Let* $\varphi\colon {\mathbb{N}_0}^{m_1} \times \cdots \times {\mathbb{N}_0}^{m_N} \to \mathbb{R}$ *and suppose that either* $\varphi(\boldsymbol{x}) \geq 0$ *for all* $\boldsymbol{x} \in {\mathbb{N}_0}^{m_1} \times \cdots \times {\mathbb{N}_0}^{m_N}$ *or* $E[|\varphi(\boldsymbol{X})|] < \infty$. *Then for all* $(i,\nu) \in \mathbb{N} \times \{1,\ldots,N\}$ *with* $i \leq m_\nu$, *if* $\varphi(\boldsymbol{x}) = 0$ *for all* $\boldsymbol{x} = ((x_{i',\nu'})_{i'=1}^{m_{\nu'}})_{\nu'=1,\ldots,N} \in {\mathbb{N}_0}^{m_1} \times \cdots \times {\mathbb{N}_0}^{m_N}$ *such that* $x_{i,\nu} = 0$, *we have*

$$E\Big[\frac{\varphi(\boldsymbol{X})}{p_{i,\nu}}\Big] = E\Big[\frac{r_\nu + X_{\cdot,\nu}}{X_{i,\nu}+1}\varphi(\boldsymbol{X}+\boldsymbol{e}_{i,\nu})\Big],$$

where $\boldsymbol{X}+\boldsymbol{e}_{i,\nu} = ((X_{i',\nu'} + \delta_{i,i',\nu,\nu'})_{i'=1}^{m_{\nu'}})_{\nu'=1,\ldots,N}$.

Proof of Theorem 4.2.1. Let $\Delta_{\boldsymbol{c}}^{(\boldsymbol{\delta})} = E[L_{\boldsymbol{c}}(\hat{\boldsymbol{p}}^{(\boldsymbol{\delta})}, \boldsymbol{p})] - E[L_{\boldsymbol{c}}(\hat{\boldsymbol{p}}^{\mathrm{U}}, \boldsymbol{p})]$. For $\nu = 1,\ldots,N$, let

$$\phi_\nu^{(\boldsymbol{\delta})}(\boldsymbol{X}_\cdot) = \begin{cases} \dfrac{\delta_\nu(X_{\cdot,\cdot})}{r_\nu + X_{\cdot,\nu} - 1 + \delta_\nu(X_{\cdot,\cdot})}, & \text{if } X_{\cdot,\nu} \geq 1, \\ 0, & \text{if } X_{\cdot,\nu} = 0, \end{cases}$$

so that $\hat{p}_{i,\nu}^{(\boldsymbol{\delta})} = \hat{p}_{i,\nu}^{\mathrm{U}} - \hat{p}_{i,\nu}^{\mathrm{U}}\phi_\nu^{(\boldsymbol{\delta})}(\boldsymbol{X}_\cdot)$ for all $i = 1,\ldots,m_\nu$. Then, by Lemma 4.6.1,

$$\begin{aligned}
\Delta_{\boldsymbol{c}}^{(\boldsymbol{\delta})} &= E\Big[\sum_{\nu=1}^{n}\sum_{i=1}^{m_\nu}\Big[c_{i,\nu}\frac{(\hat{p}_{i,\nu}^{\mathrm{U}})^2\{\phi_\nu^{(\boldsymbol{\delta})}(\boldsymbol{X}_\cdot)\}^2 - 2(\hat{p}_{i,\nu}^{\mathrm{U}})^2\phi_\nu^{(\boldsymbol{\delta})}(\boldsymbol{X}_\cdot)}{p_{i,\nu}} + 2c_{i,\nu}\hat{p}_{i,\nu}^{\mathrm{U}}\phi_\nu^{(\boldsymbol{\delta})}(\boldsymbol{X}_\cdot)\Big]\Big] \\
&= E\Big[\sum_{\nu=1}^{n}\sum_{i=1}^{m_\nu}\Big(c_{i,\nu}\frac{X_{i,\nu}+1}{r_\nu + X_{\cdot,\nu}}[\{\phi_\nu^{(\boldsymbol{\delta})}(\boldsymbol{X}_\cdot + \boldsymbol{e}_\nu^{(N)})\}^2 - 2\phi_\nu^{(\boldsymbol{\delta})}(\boldsymbol{X}_\cdot + \boldsymbol{e}_\nu^{(N)})] \\
&\quad + 2c_{i,\nu}\frac{X_{i,\nu}}{r_\nu + X_{\cdot,\nu} - 1}\phi_\nu^{(\boldsymbol{\delta})}(\boldsymbol{X}_\cdot)\Big)\Big] \\
&= E\Big[\sum_{\nu=1}^{n}\{I_{1,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{X}) - 2I_{2,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{X}) + 2I_{3,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{X})\}\Big],
\end{aligned}$$

where

$$\begin{aligned}
I_{1,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) &= \frac{\sum_{i=1}^{m_\nu} c_{i,\nu}x_{i,\nu} + c_{\cdot,\nu}}{r_\nu + \sum_{i=1}^{m_\nu} x_{i,\nu}}\Big\{\frac{\delta_\nu\big(\sum_{\nu=1}^{N}\sum_{i=1}^{m_\nu} x_{i,\nu} + 1\big)}{r_\nu + \sum_{i=1}^{m_\nu} x_{i,\nu} + \delta_\nu\big(\sum_{\nu=1}^{N}\sum_{i=1}^{m_\nu} x_{i,\nu} + 1\big)}\Big\}^2, \\
I_{2,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) &= \frac{\sum_{i=1}^{m_\nu} c_{i,\nu}x_{i,\nu} + c_{\cdot,\nu}}{r_\nu + \sum_{i=1}^{m_\nu} x_{i,\nu}}\frac{\delta_\nu\big(\sum_{\nu=1}^{N}\sum_{i=1}^{m_\nu} x_{i,\nu} + 1\big)}{r_\nu + \sum_{i=1}^{m_\nu} x_{i,\nu} + \delta_\nu\big(\sum_{\nu=1}^{N}\sum_{i=1}^{m_\nu} x_{i,\nu} + 1\big)}, \\
I_{3,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) &= \frac{\big(\sum_{i=1}^{m_\nu} c_{i,\nu}x_{i,\nu}\big)\delta_\nu\big(\sum_{\nu=1}^{N}\sum_{i=1}^{m_\nu} x_{i,\nu}\big)}{\big(r_\nu + \sum_{i=1}^{m_\nu} x_{i,\nu} - 1\big)\big\{r_\nu + \sum_{i=1}^{m_\nu} x_{i,\nu} - 1 + \delta_\nu\big(\sum_{\nu=1}^{N}\sum_{i=1}^{m_\nu} x_{i,\nu}\big)\big\}},
\end{aligned}$$

for $\boldsymbol{x} = ((x_{i,\nu'})_{i=1}^{m_{\nu'}})_{\nu'=1,\ldots,N} \in {\mathbb{N}_0}^{m_1} \times \cdots \times {\mathbb{N}_0}^{m_N}$ for each $\nu = 1,\ldots,N$. Since $\bar{\bar{c}} > 0$, it follows that $\sum_{\nu=1}^{n}\{I_{1,\nu}^{(\boldsymbol{\delta})}((\boldsymbol{0}^{(m_\nu)})_{\nu=1,\ldots,N}) - 2I_{2,\nu}^{(\boldsymbol{\delta})}((\boldsymbol{0}^{(m_\nu)})_{\nu=1,\ldots,N}) + 2I_{3,\nu}^{(\boldsymbol{\delta})}((\boldsymbol{0}^{(m_\nu)})_{\nu=1,\ldots,N})\} < 0$.

Fix $\boldsymbol{x} = ((x_{i,\nu})_{i=1}^{m_\nu})_{\nu=1,\ldots,N} \in ({\mathbb{N}_0}^{m_1} \times \cdots \times {\mathbb{N}_0}^{m_N}) \setminus \{(\boldsymbol{0}^{(m_\nu)})_{\nu=1,\ldots,N}\}$. It is sufficient to show that $\sum_{\nu=1}^{n}\{I_{1,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) - 2I_{2,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) + 2I_{3,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x})\} \leq 0$. Let $x_{\cdot,\nu} = \sum_{i=1}^{m_\nu} x_{i,\nu}$ for $\nu = 1,\ldots,N$ and let $x_{\cdot,\cdot} = \sum_{\nu=1}^{N} x_{\cdot,\nu}$. Let $\bar{c}_\nu = \max_{1\leq i\leq m_\nu} c_{i,\nu}$ for $\nu = 1,\ldots,N$. Then for all $\nu = 1,\ldots,n$ such that $\sum_{i=1}^{m_\nu} c_{i,\nu}x_{i,\nu} > 0$, since, by (4.2.4), $\delta_\nu(x_{\cdot,\cdot}) \leq \{(x_{\cdot,\cdot}+1)/x_{\cdot,\cdot}\}\delta_\nu(x_{\cdot,\cdot}+1) \leq \{(x_{\cdot,\nu}+1)/x_{\cdot,\nu}\}\delta_\nu(x_{\cdot,\cdot}+1)$, we have that

$$I_{3,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) \leq \frac{\sum_{i=1}^{m_\nu} c_{i,\nu}x_{i,\nu}}{r_\nu + x_{\cdot,\nu} - 1}\frac{\delta_\nu(x_{\cdot,\cdot}+1)}{\{x_{\cdot,\nu}/(x_{\cdot,\nu}+1)\}(r_\nu + x_{\cdot,\nu} - 1) + \delta_\nu(x_{\cdot,\cdot}+1)}$$

and hence that

$$
\begin{aligned}
-I_{2,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) + I_{3,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) &\le -\frac{c_{\cdot,\nu}}{r_\nu + x_{\cdot,\nu}} \frac{\delta_\nu(x_{\cdot,\cdot}+1)}{r_\nu + x_{\cdot,\nu} + \delta_\nu(x_{\cdot,\cdot}+1)} \\
&\quad + \Big(\sum_{i=1}^{m_\nu} c_{i,\nu} x_{i,\nu}\Big)\delta_\nu(x_{\cdot,\cdot}+1)\Big[-\frac{1}{r_\nu + x_{\cdot,\nu}} \frac{1}{r_\nu + x_{\cdot,\nu} + \delta_\nu(x_{\cdot,\cdot}+1)} \\
&\quad + \frac{1}{r_\nu + x_{\cdot,\nu} - 1} \frac{1}{\{x_{\cdot,\nu}/(x_{\cdot,\nu}+1)\}(r_\nu + x_{\cdot,\nu} - 1) + \delta_\nu(x_{\cdot,\cdot}+1)}\Big] \\
&\le -\frac{c_{\cdot,\nu}}{r_\nu + x_{\cdot,\nu}} \frac{\delta_\nu(x_{\cdot,\cdot}+1)}{r_\nu + x_{\cdot,\nu} + \delta_\nu(x_{\cdot,\cdot}+1)} \\
&\quad + \overline{c}_\nu x_{\cdot,\nu}\delta_\nu(x_{\cdot,\cdot}+1)\Big[-\frac{1}{r_\nu + x_{\cdot,\nu}} \frac{1}{r_\nu + x_{\cdot,\nu} + \delta_\nu(x_{\cdot,\cdot}+1)} \\
&\quad + \frac{1}{r_\nu + x_{\cdot,\nu} - 1} \frac{1}{\{x_{\cdot,\nu}/(x_{\cdot,\nu}+1)\}(r_\nu + x_{\cdot,\nu} - 1) + \delta_\nu(x_{\cdot,\cdot}+1)}\Big],
\end{aligned}
$$

where

$$
\frac{1}{r_\nu + x_{\cdot,\nu} - 1} \frac{1}{\{x_{\cdot,\nu}/(x_{\cdot,\nu}+1)\}(r_\nu + x_{\cdot,\nu} - 1) + \delta_\nu(x_{\cdot,\cdot}+1)} \le \frac{x_{\cdot,\nu}+3}{r_\nu + x_{\cdot,\nu}} \frac{1/x_{\cdot,\nu}}{r_\nu + x_{\cdot,\nu} + \delta_\nu(x_{\cdot,\cdot}+1)}
$$

by the assumption that $r_\nu \ge 5/2$ for all $\nu = 1, \ldots, n$ with $c_{\cdot,\nu} > 0$. Thus, for any $\nu = 1, \ldots, n$,

$$
\begin{aligned}
&I_{1,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) - 2I_{2,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) + 2I_{3,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) \\
&\le \frac{\overline{c}_\nu x_{\cdot,\nu} + c_{\cdot,\nu}}{r_\nu + x_{\cdot,\nu}} \Big\{ \frac{\delta_\nu(x_{\cdot,\cdot}+1)}{r_\nu + x_{\cdot,\nu} + \delta_\nu(x_{\cdot,\cdot}+1)} \Big\}^2 + 2\frac{3\overline{c}_\nu - c_{\cdot,\nu}}{r_\nu + x_{\cdot,\nu}} \frac{\delta_\nu(x_{\cdot,\cdot}+1)}{r_\nu + x_{\cdot,\nu} + \delta_\nu(x_{\cdot,\cdot}+1)} \\
&= \frac{\delta_\nu(x_{\cdot,\cdot}+1)[(\overline{c}_\nu x_{\cdot,\nu} + c_{\cdot,\nu})\delta_\nu(x_{\cdot,\cdot}+1) - 2(c_{\cdot,\nu} - 3\overline{c}_\nu)\{r_\nu + x_{\cdot,\nu} + \delta_\nu(x_{\cdot,\cdot}+1)\}]}{(r_\nu + x_{\cdot,\nu})\{r_\nu + x_{\cdot,\nu} + \delta_\nu(x_{\cdot,\cdot}+1)\}^2} \\
&\le \frac{\delta_\nu(x_{\cdot,\cdot}+1)[(\overline{\overline{c}} x_{\cdot,\nu} + \underline{c}_\cdot)\delta_\nu(x_{\cdot,\cdot}+1) - 2(\underline{c}_\cdot - 3\overline{\overline{c}})\{r_\nu + x_{\cdot,\nu} + \delta_\nu(x_{\cdot,\cdot}+1)\}]}{(r_\nu + x_{\cdot,\nu})\{r_\nu + x_{\cdot,\nu} + \delta_\nu(x_{\cdot,\cdot}+1)\}^2} \\
&\le \frac{\overline{\overline{c}} x_{\cdot,\nu} + \underline{c}_\cdot}{r_\nu + x_{\cdot,\nu}} \Big\{ \frac{\overline{\delta}(x_{\cdot,\cdot}+1)}{r_\nu + x_{\cdot,\nu} + \overline{\delta}(x_{\cdot,\cdot}+1)} \Big\}^2 - 2\frac{\underline{c}_\cdot - 3\overline{\overline{c}}}{r_\nu + x_{\cdot,\nu}} \frac{\underline{\delta}(x_{\cdot,\cdot}+1)}{r_\nu + x_{\cdot,\nu} + \underline{\delta}(x_{\cdot,\cdot}+1)}
\end{aligned} \tag{4.6.1}
$$

by the assumption that $3\overline{\overline{c}} \le \underline{c}_\cdot$.

For part (i), we have by (4.6.1) that for any $\nu = 1, \ldots, n$,

$$
\begin{aligned}
&I_{1,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) - 2I_{2,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) + 2I_{3,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) \\
&\le \frac{\overline{\overline{c}} x_{\cdot,\nu} + \underline{c}_\cdot}{\underline{r} + x_{\cdot,\nu}} \Big\{ \frac{\overline{\delta}(x_{\cdot,\cdot}+1)}{\underline{r} + x_{\cdot,\nu} + \overline{\delta}(x_{\cdot,\cdot}+1)} \Big\}^2 - 2\frac{\underline{c}_\cdot - 3\overline{\overline{c}}}{\overline{r} + x_{\cdot,\nu}} \frac{\underline{\delta}(x_{\cdot,\cdot}+1)}{\overline{r} + x_{\cdot,\nu} + \underline{\delta}(x_{\cdot,\cdot}+1)} \\
&\le \frac{1}{\underline{r} + x_{\cdot,\nu}} \frac{\overline{\delta}(x_{\cdot,\cdot}+1)}{\{\underline{r} + x_{\cdot,\nu} + \overline{\delta}(x_{\cdot,\cdot}+1)\}^2} \\
&\quad \times [x_{\cdot,\nu}\{\overline{\overline{c}}\overline{\delta}(x_{\cdot,\cdot}+1) - 2(\underline{r}/\overline{r})^2(\underline{c}_\cdot - 3\overline{\overline{c}})\rho\} + \underline{c}_\cdot\overline{\delta}(x_{\cdot,\cdot}+1) - 2(\underline{r}/\overline{r})^2(\underline{c}_\cdot - 3\overline{\overline{c}})\rho\{\underline{r} + \overline{\delta}(x_{\cdot,\cdot}+1)\}],
\end{aligned}
$$

which is nonpositive by (4.2.5) if $\overline{\overline{c}}\overline{\delta}(x_{\cdot,\cdot}+1) - 2(\underline{r}/\overline{r})^2(\underline{c}_\cdot - 3\overline{\overline{c}})\rho \le 0$. On the other hand, if

$\overline{\underline{c}}\overline{\delta}(x_{\cdot,\cdot}+1) - 2(\underline{r}/\overline{r})^2(\underline{c}_{\cdot} - 3\overline{\underline{c}})\rho > 0$, then, by the covariance inequality,

$$\begin{aligned}
&\sum_{\nu=1}^{n}\{I_{1,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) - 2I_{2,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) + 2I_{3,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x})\} \\
&\le \frac{1}{n}\Big[\sum_{\nu=1}^{n}\frac{1}{\underline{r}+x_{\cdot,\nu}}\frac{\overline{\delta}(x_{\cdot,\cdot}+1)}{\{\underline{r}+x_{\cdot,\nu}+\overline{\delta}(x_{\cdot,\cdot}+1)\}^2}\Big] \\
&\quad\times[x_{\cdot,\cdot}\{\overline{\underline{c}}\overline{\delta}(x_{\cdot,\cdot}+1) - 2(\underline{r}/\overline{r})^2(\underline{c}_{\cdot}-3\overline{\underline{c}})\rho\} + n\underline{c}_{\cdot}\overline{\delta}(x_{\cdot,\cdot}+1) - 2n(\underline{r}/\overline{r})^2(\underline{c}_{\cdot}-3\overline{\underline{c}})\rho\{\underline{r}+\overline{\delta}(x_{\cdot,\cdot}+1)\}],
\end{aligned}$$

which is nonpositive by (4.2.6). This proves part (i).

For part (ii), it follows from (4.6.1) that for all $\nu = 1,\dots,n$,

$$\begin{aligned}
&I_{1,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) - 2I_{2,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) + 2I_{3,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) \\
&\le \frac{1}{r_\nu+x_{\cdot,\nu}}\frac{\overline{\delta}(x_{\cdot,\cdot}+1)}{\{r_\nu+x_{\cdot,\nu}+\overline{\delta}(x_{\cdot,\cdot}+1)\}^2}[(\overline{\underline{c}}x_{\cdot,\nu}+\underline{c}_{\cdot})\overline{\delta}(x_{\cdot,\cdot}+1) - 2(\underline{c}_{\cdot}-3\overline{\underline{c}})\rho\{r_\nu+x_{\cdot,\nu}+\overline{\delta}(x_{\cdot,\cdot}+1)\}] \\
&\le \frac{1}{r_\nu+x_{\cdot,\nu}}\frac{\overline{\delta}(x_{\cdot,\cdot}+1)}{\{r_\nu+x_{\cdot,\nu}+\overline{\delta}(x_{\cdot,\cdot}+1)\}^2} \\
&\quad\times[(r_\nu+x_{\cdot,\nu})\{\overline{\underline{c}}\overline{\delta}(x_{\cdot,\cdot}+1) - 2(\underline{c}_{\cdot}-3\overline{\underline{c}})\rho\} + \{\underline{c}_{\cdot} - \underline{r}\overline{\underline{c}} - 2(\underline{c}_{\cdot}-3\overline{\underline{c}})\rho\}\overline{\delta}(x_{\cdot,\cdot}+1)],
\end{aligned}$$

which is nonpositive by (4.2.7) if $\overline{\underline{c}}\overline{\delta}(x_{\cdot,\cdot}+1) - 2(\underline{c}_{\cdot}-3\overline{\underline{c}})\rho \le 0$. If $\overline{\underline{c}}\overline{\delta}(x_{\cdot,\cdot}+1) - 2(\underline{c}_{\cdot}-3\overline{\underline{c}})\rho > 0$, then, by the covariance inequality,

$$\begin{aligned}
&\sum_{\nu=1}^{n}\{I_{1,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) - 2I_{2,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x}) + 2I_{3,\nu}^{(\boldsymbol{\delta})}(\boldsymbol{x})\} \\
&\le \frac{1}{n}\Big[\sum_{\nu=1}^{n}\frac{1}{r_\nu+x_{\cdot,\nu}}\frac{\overline{\delta}(x_{\cdot,\cdot}+1)}{\{r_\nu+x_{\cdot,\nu}+\overline{\delta}(x_{\cdot,\cdot}+1)\}^2}\Big] \\
&\quad\times\Big[\Big(\sum_{\nu=1}^{n}r_\nu+x_{\cdot,\cdot}\Big)\{\overline{\underline{c}}\overline{\delta}(x_{\cdot,\cdot}+1) - 2(\underline{c}_{\cdot}-3\overline{\underline{c}})\rho\} + n\{\underline{c}_{\cdot} - \underline{r}\overline{\underline{c}} - 2(\underline{c}_{\cdot}-3\overline{\underline{c}})\rho\}\overline{\delta}(x_{\cdot,\cdot}+1)\Big],
\end{aligned}$$

which is nonpositive by (4.2.8). This proves part (ii). □

Proof of Theorem 4.2.2. Let $\Delta_{\boldsymbol{c}}^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})} = E[L_{\boldsymbol{c}}(\hat{\boldsymbol{p}}^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})},\boldsymbol{p})] - E[L_{\boldsymbol{c}}(\hat{\boldsymbol{p}}^{\mathrm{U}},\boldsymbol{p})]$. For $\nu = 1,\dots,N$, let

$$\tilde{\delta}_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})}) = \begin{cases}\tilde{b}_\nu + 1/\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})}, & \text{if } \widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})} > 0,\\ 0, & \text{if } \widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})} = 0,\end{cases}$$

so that

$$\hat{p}_{i,\nu}^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})} = \hat{p}_{i,\nu}^{\mathrm{U}} - \frac{\hat{p}_{i,\nu}^{\mathrm{U}}\tilde{\delta}_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}{r_\nu + X_{\cdot,\nu} - 1 + \tilde{\delta}_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}$$

for all $i = 1, \ldots, m_\nu$. By Lemma 4.6.1, we have

$$
\begin{aligned}
\Delta_{\boldsymbol{c}}^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})} &= E\Big[\sum_{\nu=1}^{n}\sum_{i=1}^{m_\nu}\Big(\frac{c_{i,\nu}}{p_{i,\nu}}\Big[\frac{(\hat{p}_{i,\nu}^{\mathrm{U}})^2\{\tilde{\delta}_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})})\}^2}{\{r_\nu + X_{\cdot,\nu} - 1 + \tilde{\delta}_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})})\}^2} \\
&\quad - 2\frac{(\hat{p}_{i,\nu}^{\mathrm{U}})^2\tilde{\delta}_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}{r_\nu + X_{\cdot,\nu} - 1 + \tilde{\delta}_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}\Big] + 2c_{i,\nu}\frac{\hat{p}_{i,\nu}^{\mathrm{U}}\tilde{\delta}_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}{r_\nu + X_{\cdot,\nu} - 1 + \tilde{\delta}_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}\Big)\Big] \\
&= E\Big[\sum_{\nu=1}^{n}\sum_{i=1}^{m_\nu}\Big(c_{i,\nu}\frac{X_{i,\nu}+1}{r_\nu + X_{\cdot,\nu}}\Big[\frac{\{\tilde{b}_\nu + 1/(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \tilde{c}_{i,\nu}^{(\nu)})\}^2}{\{r_\nu + X_{\cdot,\nu} + \tilde{b}_\nu + 1/(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \tilde{c}_{i,\nu}^{(\nu)})\}^2} \\
&\quad - 2\frac{\tilde{b}_\nu + 1/(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \tilde{c}_{i,\nu}^{(\nu)})}{r_\nu + X_{\cdot,\nu} + \tilde{b}_\nu + 1/(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \tilde{c}_{i,\nu}^{(\nu)})}\Big] + 2c_{i,\nu}\frac{\hat{p}_{i,\nu}^{\mathrm{U}}\tilde{\delta}_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}{r_\nu + X_{\cdot,\nu} - 1 + \tilde{\delta}_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}\Big)\Big] \\
&\le E\Big[\sum_{\nu=1}^{n}\sum_{i=1}^{m_\nu}\Big(c_{i,\nu}\frac{X_{i,\nu}+1}{r_\nu + X_{\cdot,\nu}}\Big[\frac{\{\tilde{b}_\nu + 1/(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\}^2}{\{r_\nu + X_{\cdot,\nu} + \tilde{b}_\nu + 1/(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\}^2} \\
&\quad - 2\frac{\tilde{b}_\nu + 1/(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})}{r_\nu + X_{\cdot,\nu} + \tilde{b}_\nu + 1/(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})}\Big] + 2c_{i,\nu}\frac{\hat{p}_{i,\nu}^{\mathrm{U}}\tilde{\delta}_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}{r_\nu + X_{\cdot,\nu} - 1 + \tilde{\delta}_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}(\widetilde{X}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}\Big)\Big].
\end{aligned}
$$

Fix $((x_{i,\nu})_{i=1}^{m_\nu})_{\nu=1,\ldots,N} \in ({\mathbb{N}_0}^{m_1} \times \cdots \times {\mathbb{N}_0}^{m_N}) \setminus \{(\mathbf{0}^{(m_\nu)})_{\nu=1,\ldots,N}\}$ and let $x_{\cdot,\nu} = \sum_{i=1}^{m_\nu} x_{i,\nu}$ and $\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} = \sum_{\nu'=1}^{N}\sum_{i=1}^{m_{\nu'}} \tilde{c}_{i,\nu'}^{(\nu)} x_{i,\nu'}$ for $\nu = 1, \ldots, N$. As in the proof of Theorem 4.2.1, it is sufficient to show that $\sum_{\nu=1}^{n} I_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})} \le 0$, where

$$
\begin{aligned}
I_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})} &= \sum_{i=1}^{m_\nu}\Big(c_{i,\nu}\frac{x_{i,\nu}+1}{r_\nu + x_{\cdot,\nu}}\Big[\frac{\{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\}^2}{\{r_\nu + x_{\cdot,\nu} + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\}^2} \\
&\quad - 2\frac{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})}{r_\nu + x_{\cdot,\nu} + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})}\Big] + \frac{2c_{i,\nu}x_{i,\nu}(\tilde{b}_\nu + 1/\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}{(r_\nu + x_{\cdot,\nu} - 1)(r_\nu + x_{\cdot,\nu} - 1 + \tilde{b}_\nu + 1/\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}\Big)
\end{aligned}
$$

for $\nu = 1, \ldots, n$. It can be verified that for all $\nu = 1, \ldots, n$,

$$
\begin{aligned}
&I_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})} \\
&\le \frac{\overline{c}_\nu x_{\cdot,\nu} + c_{\cdot,\nu}}{r_\nu + x_{\cdot,\nu}}\frac{\{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\}^2}{\{r_\nu + x_{\cdot,\nu} + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\}^2} - 2\frac{c_{\cdot,\nu}}{r_\nu + x_{\cdot,\nu}}\frac{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})}{r_\nu + x_{\cdot,\nu} + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})} \\
&\quad - 2\frac{\overline{c}_\nu x_{\cdot,\nu}}{r_\nu + x_{\cdot,\nu}}\frac{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})}{r_\nu + x_{\cdot,\nu} + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})} + \frac{2\overline{c}_\nu x_{\cdot,\nu}(\tilde{b}_\nu + 1/\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}{(r_\nu + x_{\cdot,\nu} - 1)(r_\nu + x_{\cdot,\nu} - 1 + \tilde{b}_\nu + 1/\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})})}.
\end{aligned}
$$

Now for all $\nu = 1, \ldots, n$ such that $\overline{c}_\nu x_{\cdot,\nu} > 0$, since

$$
\begin{aligned}
&x_{\cdot,\nu}(\tilde{b}_\nu + 1/\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})}) - (x_{\cdot,\nu} + \widetilde{C}_\nu)\{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\} \\
&= \overline{\tilde{c}}_\nu^{(\nu)} x_{\cdot,\nu}/\{\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})}(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\} - \widetilde{C}_\nu\{\tilde{b}_\nu(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)}) + 1\}/(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)}) \\
&\le \overline{\tilde{c}}_\nu^{(\nu)} x_{\cdot,\nu}/\{\underline{\tilde{c}}_\nu^{(\nu)} x_{\cdot,\nu}(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\} - \widetilde{C}_\nu\{\tilde{b}_\nu(\underline{\tilde{c}}_\nu^{(\nu)} + \overline{\tilde{c}}_\nu^{(\nu)}) + 1\}/(\tilde{x}^{(\tilde{\boldsymbol{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)}) = 0,
\end{aligned}
$$

it follows that

$$\begin{aligned}
&\frac{2\overline{c}_\nu x_{\cdot,\nu}(\tilde{b}_\nu + 1/\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})})}{(r_\nu + x_{\cdot,\nu} - 1)(r_\nu + x_{\cdot,\nu} - 1 + \tilde{b}_\nu + 1/\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})})} \\
&\le \frac{2\overline{c}_\nu(x_{\cdot,\nu} + \widetilde{C}_\nu)}{r_\nu + x_{\cdot,\nu} - 1} \frac{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})}{r_\nu + x_{\cdot,\nu} - 1 + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})} \\
&\le \frac{2\overline{c}_\nu(x_{\cdot,\nu} + \widetilde{C}_\nu + 1)}{r_\nu + x_{\cdot,\nu}} \frac{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})}{r_\nu + x_{\cdot,\nu} - 1 + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})} \\
&\le \frac{2\overline{c}_\nu(x_{\cdot,\nu} + \widetilde{C}_\nu + 2)}{r_\nu + x_{\cdot,\nu}} \frac{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})}{r_\nu + x_{\cdot,\nu} + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})} \qquad (4.6.2)
\end{aligned}$$

by assumption. Therefore, letting $x_{\cdot,\cdot} = \sum_{\nu=1}^{N} x_{\cdot,\nu}$ and noting that $\underline{c}_\cdot - A \ge 0$, we have for all $\nu = 1, \ldots, n$,

$$\begin{aligned}
I_\nu^{(\tilde{\mathbf{b}},\tilde{\mathbf{c}})} &\le \frac{\overline{c}_\nu x_{\cdot,\nu} + c_{\cdot,\nu}}{r_\nu + x_{\cdot,\nu}} \frac{\{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\}^2}{\{r_\nu + x_{\cdot,\nu} + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\}^2} \\
&\quad + 2\frac{\overline{c}_\nu(\widetilde{C}_\nu + 2) - c_{\cdot,\nu}}{r_\nu + x_{\cdot,\nu}} \frac{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})}{r_\nu + x_{\cdot,\nu} + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})} \\
&= \frac{1}{r_\nu + x_{\cdot,\nu}} \frac{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})}{\{r_\nu + x_{\cdot,\nu} + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\}^2} \\
&\quad \times [(\overline{c}_\nu x_{\cdot,\nu} + c_{\cdot,\nu})\{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\} \\
&\quad - 2\{c_{\cdot,\nu} - \overline{c}_\nu(\widetilde{C}_\nu + 2)\}\{r_\nu + x_{\cdot,\nu} + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\}] \\
&\le \frac{1}{r_\nu + x_{\cdot,\nu}} \frac{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})}{\{r_\nu + x_{\cdot,\nu} + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\}^2} \\
&\quad \times [(\overline{\overline{c}} x_{\cdot,\nu} + \underline{c}_\cdot)\{\tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\} - 2(\underline{c}_\cdot - A)\{r_\nu + x_{\cdot,\nu} + \tilde{b}_\nu + 1/(\tilde{x}^{(\tilde{\mathbf{c}}^{(\nu)})} + \overline{\tilde{c}}_\nu^{(\nu)})\}] \\
&\le \frac{\overline{\overline{c}} x_{\cdot,\nu} + \underline{c}_\cdot}{r_\nu + x_{\cdot,\nu}} \frac{\{\overline{\tilde{b}} + 1/(\tilde{c}_* x_{\cdot,\cdot} + \overline{\tilde{c}})\}^2}{\{r_\nu + x_{\cdot,\nu} + \overline{\tilde{b}} + 1/(\tilde{c}_* x_{\cdot,\cdot} + \overline{\tilde{c}})\}^2} - 2\frac{\underline{c}_\cdot - A}{r_\nu + x_{\cdot,\nu}} \frac{\underline{\tilde{b}} + 1/(\tilde{c}^* x_{\cdot,\cdot} + \overline{\overline{\tilde{c}}})}{r_\nu + x_{\cdot,\nu} + \underline{\tilde{b}} + 1/(\tilde{c}^* x_{\cdot,\cdot} + \overline{\overline{\tilde{c}}})},
\end{aligned}$$

which implies that

$$
\begin{aligned}
I_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})} &\le \frac{\overline{\tilde{c}}x_{\cdot,\nu}+\underline{c}_\cdot}{\underline{r}+x_{\cdot,\nu}}\frac{\{\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}^2}{\{\underline{r}+x_{\cdot,\nu}+\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}^2}-2\frac{\underline{c}_\cdot-A}{\overline{r}+x_{\cdot,\nu}}\frac{\underline{\tilde{b}}+1/(\tilde{c}^*x_{\cdot,\cdot}+\underline{\tilde{c}})}{\overline{r}+x_{\cdot,\nu}+\underline{\tilde{b}}+1/(\tilde{c}^*x_{\cdot,\cdot}+\underline{\tilde{c}})}\\
&\le \frac{1}{\underline{r}+x_{\cdot,\nu}}\frac{\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})}{\{\underline{r}+x_{\cdot,\nu}+\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}^2}\\
&\quad\times\Big[(\overline{\tilde{c}}x_{\cdot,\nu}+\underline{c}_\cdot)\{\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}-2\Big(\frac{\underline{r}}{\overline{r}}\Big)^2(\underline{c}_\cdot-A)\frac{\underline{\tilde{b}}\tilde{c}_*\overline{\tilde{c}}}{\overline{\tilde{b}}\tilde{c}^*\underline{\tilde{c}}}\{\underline{r}+x_{\cdot,\nu}+\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}\Big]\\
&= \frac{1}{\underline{r}+x_{\cdot,\nu}}\frac{\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})}{\{\underline{r}+x_{\cdot,\nu}+\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}^2}\\
&\quad\times\Big(x_{\cdot,\nu}\Big[\overline{\tilde{c}}\{\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}-2\Big(\frac{\underline{r}}{\overline{r}}\Big)^2(\underline{c}_\cdot-A)\frac{\underline{\tilde{b}}\tilde{c}_*\overline{\tilde{c}}}{\overline{\tilde{b}}\tilde{c}^*\underline{\tilde{c}}}\Big]\\
&\quad+\underline{c}_\cdot\{\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}-2\Big(\frac{\underline{r}}{\overline{r}}\Big)^2(\underline{c}_\cdot-A)\frac{\underline{\tilde{b}}\tilde{c}_*\overline{\tilde{c}}}{\overline{\tilde{b}}\tilde{c}^*\underline{\tilde{c}}}\{\underline{r}+\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}\Big)
\end{aligned}
\tag{4.6.3}
$$

and that

$$
\begin{aligned}
I_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})} &\le \frac{\overline{\tilde{c}}(r_\nu+x_{\cdot,\nu})+\underline{c}_\cdot-\overline{\tilde{c}}\underline{r}}{r_\nu+x_{\cdot,\nu}}\frac{\{\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}^2}{\{r_\nu+x_{\cdot,\nu}+\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}^2}-2\frac{\underline{c}_\cdot-A}{r_\nu+x_{\cdot,\nu}}\frac{\underline{\tilde{b}}+1/(\tilde{c}^*x_{\cdot,\cdot}+\underline{\tilde{c}})}{r_\nu+x_{\cdot,\nu}+\underline{\tilde{b}}+1/(\tilde{c}^*x_{\cdot,\cdot}+\underline{\tilde{c}})}\\
&\le \frac{1}{r_\nu+x_{\cdot,\nu}}\frac{\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})}{\{r_\nu+x_{\cdot,\nu}+\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}^2}\\
&\quad\times\Big[\{\overline{\tilde{c}}(r_\nu+x_{\cdot,\nu})+\underline{c}_\cdot-\overline{\tilde{c}}\underline{r}\}\{\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}-2(\underline{c}_\cdot-A)\frac{\underline{\tilde{b}}\tilde{c}_*\overline{\tilde{c}}}{\overline{\tilde{b}}\tilde{c}^*\underline{\tilde{c}}}\{r_\nu+x_{\cdot,\nu}+\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}\Big]\\
&= \frac{1}{r_\nu+x_{\cdot,\nu}}\frac{\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})}{\{r_\nu+x_{\cdot,\nu}+\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}^2}\\
&\quad\times\Big((r_\nu+x_{\cdot,\nu})\Big[\overline{\tilde{c}}\{\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}-2(\underline{c}_\cdot-A)\frac{\underline{\tilde{b}}\tilde{c}_*\overline{\tilde{c}}}{\overline{\tilde{b}}\tilde{c}^*\underline{\tilde{c}}}\Big]\\
&\quad+(\underline{c}_\cdot-\overline{\tilde{c}}\underline{r})\{\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}-2(\underline{c}_\cdot-A)\frac{\underline{\tilde{b}}\tilde{c}_*\overline{\tilde{c}}}{\overline{\tilde{b}}\tilde{c}^*\underline{\tilde{c}}}\{\overline{\tilde{b}}+1/(\tilde{c}_*x_{\cdot,\cdot}+\overline{\tilde{c}})\}\Big).
\end{aligned}
\tag{4.6.4}
$$

By (4.6.3) and (4.6.4) and by the covariance inequality, we conclude as in the proof of Theorem 4.2.1 that $\sum_{\nu=1}^n I_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})}\le 0$. □

Remark 4.6.1 Suppose that $m_1=\cdots=m_N$, that $r_1=\cdots=r_N$, and that $\boldsymbol{c}=(\boldsymbol{j}^{(m_\nu)})_{\nu=1,\ldots,N}$. Then, by modifying the above proof, we can show that if $r_1\ge 1$, the UMVU estimator is dominated by an empirical Bayes estimator for sufficiently large m_1, which is related to the problem of Section 5.1 of Hamura and Kubokawa (2020b). For example, the empirical Bayes estimator (4.2.9) with $\mathring{\boldsymbol{a}}=\boldsymbol{j}^{(N)}$ corresponds to $\tilde{\boldsymbol{b}}=m_1\boldsymbol{j}^{(N)}$ and $\tilde{\boldsymbol{c}}=(((1/(Nm_1r_1))_{i=1}^{m_{\nu'}})_{\nu'=1,\ldots,N})_{\nu=1}^N$. In

this case,

$$I_\nu^{(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}})} = \frac{x_{\cdot,\nu}+m_1}{r_1+x_{\cdot,\nu}}\Big[\Big\{\frac{m_1+Nm_1r_1/(x_{\cdot,\cdot}+1)}{r_1+x_{\cdot,\nu}+m_1+Nm_1r_1/(x_{\cdot,\cdot}+1)}\Big\}^2 - 2\frac{m_1+Nm_1r_1/(x_{\cdot,\cdot}+1)}{r_1+x_{\cdot,\nu}+m_1+Nm_1r_1/(x_{\cdot,\cdot}+1)}\Big] + \frac{2x_{\cdot,\nu}(m_1+Nm_1r_1/x_{\cdot,\cdot})}{(r_1+x_{\cdot,\nu}-1)(r_1+x_{\cdot,\nu}-1+m_1+Nm_1r_1/x_{\cdot,\cdot})}$$

for $\nu=1,\ldots,n$. Now suppose that $r_1\geq 1$ and that $r_1+m_1\geq 4$. Then for all $\nu=1,\ldots,n$ such that $x_{\cdot,\nu}\geq 1$, (4.6.2) can be replaced by

$$\begin{aligned}
&\frac{2x_{\cdot,\nu}(m_1+Nm_1r_1/x_{\cdot,\cdot})}{(r_1+x_{\cdot,\nu}-1)(r_1+x_{\cdot,\nu}-1+m_1+Nm_1r_1/x_{\cdot,\cdot})}\\
&\leq \frac{2(x_{\cdot,\nu}+1)}{r_1+x_{\cdot,\nu}}\frac{m_1+Nm_1r_1/x_{\cdot,\cdot}}{r_1+x_{\cdot,\nu}-1+m_1+Nm_1r_1/x_{\cdot,\cdot}}\\
&\leq \frac{2(x_{\cdot,\nu}+3)}{r_1+x_{\cdot,\nu}}\frac{m_1+Nm_1r_1/(x_{\cdot,\cdot}+1)}{r_1+x_{\cdot,\nu}-1+m_1+Nm_1r_1/x_{\cdot,\cdot}}\\
&\leq \frac{2(x_{\cdot,\nu}+4)}{r_1+x_{\cdot,\nu}}\frac{m_1+Nm_1r_1/(x_{\cdot,\cdot}+1)}{r_1+x_{\cdot,\nu}+m_1+Nm_1r_1/x_{\cdot,\cdot}},
\end{aligned}$$

where the second inequality holds even if $x_{\cdot,\nu}=x_{\cdot,\cdot}$ since $x_{\cdot,\cdot}\geq 1$. This leads to a dominance condition which is satisfied when m_1 is sufficiently large.

Proof of Lemma 4.3.2. We have

$$\begin{aligned}
\frac{f(\boldsymbol{w}|\boldsymbol{p})}{C(\boldsymbol{w})} &= \prod_{\lambda=1}^{L}\prod_{\boldsymbol{i}=(i_h)_{h=1}^{d(\lambda)}\in I_0^{(\lambda)}}\Big\{\prod_{h=1}^{d(\lambda)}p_{i_h,\nu_h^{(\lambda)}}\Big\}^{w_{\boldsymbol{i}}^{(\lambda)}} = \prod_{\lambda=1}^{L}\prod_{h=1}^{d(\lambda)}\prod_{\boldsymbol{i}=(i_h)_{h=1}^{d(\lambda)}\in I_0^{(\lambda)}}p_{i_h,\nu_h^{(\lambda)}}{}^{w_{\boldsymbol{i}}^{(\lambda)}}\\
&= \prod_{\nu=1}^{N}\prod_{i=0}^{m_\nu}\prod_{\lambda\in\Lambda(\nu)}\prod_{\boldsymbol{i}\in I_0^{(\lambda)}(i,\nu)}p_{i,\nu}{}^{w_{\boldsymbol{i}}^{(\lambda)}} = \prod_{\nu=1}^{N}\prod_{i=0}^{m_\nu}p_{i,\nu}{}^{\sum_{\lambda\in\Lambda(\nu)}\sum_{\boldsymbol{i}\in I_0^{(\lambda)}(i,\nu)}w_{\boldsymbol{i}}^{(\lambda)}},
\end{aligned}$$

which is the desired result. □

Proof of Theorem 4.3.1. In this proof, if φ is a continuous function from $(0,\infty)$ to $[0,\infty)$, we write

$$\int_0^\infty d\mu(u) = \int_0^\infty u^{\alpha-1}e^{-\beta u}\Big\{\prod_{\nu=1}^{N}\frac{\Gamma(\gamma_\nu u+r_\nu+a_{0,\nu})\Gamma(r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})}{\Gamma(\gamma_\nu u+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})\Gamma(r_\nu+a_{0,\nu})}\Big\}du,$$

$$\int_0^\infty \varphi(u)d\mu(u) = \int_0^\infty \varphi(u)u^{\alpha-1}e^{-\beta u}\Big\{\prod_{\nu=1}^{N}\frac{\Gamma(\gamma_\nu u+r_\nu+a_{0,\nu})\Gamma(r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})}{\Gamma(\gamma_\nu u+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})\Gamma(r_\nu+a_{0,\nu})}\Big\}du, \quad \text{and}$$

$$E^U[\varphi(U)] = \int_0^\infty \varphi(u)d\mu(u)/\int_0^\infty d\mu(u).$$

Let $\Delta^{(\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a})} = E[\log\{f(\boldsymbol{W}|\boldsymbol{p})/\hat{f}^{(\pi_{\alpha,\beta,\gamma,\boldsymbol{a}_0,\boldsymbol{a}})}(\boldsymbol{W};\boldsymbol{X})\}] - E[\log\{f(\boldsymbol{W}|\boldsymbol{p})/\hat{f}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})}(\boldsymbol{W};\boldsymbol{X})\}]$.

Then, by Proposition 4.3.1,

$$\begin{aligned}
\Delta^{(\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a})} &= E\Big[-\log \frac{\hat{f}^{(\pi_{\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a}})}(\boldsymbol{W};\boldsymbol{X})}{\hat{f}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})}(\boldsymbol{W};\boldsymbol{X})}\Big] \\
&= E\Big[-\log E^U\Big[\prod_{\nu=1}^{N} \Big\{ \frac{\Gamma(\gamma_\nu U + s_{0,\nu}(\boldsymbol{W}) + r_\nu + a_{0,\nu})\Gamma\big(\sum_{\lambda\in\Lambda(\nu)} l^{(\lambda)} + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu}\big)}{\Gamma\big(\gamma_\nu U + \sum_{\lambda\in\Lambda(\nu)} l^{(\lambda)} + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu}\big)\Gamma(s_{0,\nu}(\boldsymbol{W}) + r_\nu + a_{0,\nu})} \\
&\quad \times \frac{\Gamma(\gamma_\nu U + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu})\Gamma(r_\nu + a_{0,\nu})}{\Gamma(\gamma_\nu U + r_\nu + a_{0,\nu})\Gamma(r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu})} \Big\}\Big]\Big]. \qquad (4.6.5)
\end{aligned}$$

For $\nu = 1,\dots,N$, let $\tilde{p}_{0,\nu} = p_{0,\nu}$ and $\tilde{p}_{1,\nu} = p_{\cdot,\nu} = \sum_{i=1}^{m_\nu} p_{i,\nu}$ for notational convenience. For $\lambda = 1,\dots,L$, let $\widetilde{\mathcal{W}}^{(\lambda)} = \big\{(\mathring{w}_{\tilde{\boldsymbol{i}}})_{\tilde{\boldsymbol{i}}\in\{0,1\}^{d(\lambda)}} \big| \mathring{w}_{\tilde{\boldsymbol{i}}} \in \mathbb{N}_0 \quad \text{for all } \tilde{\boldsymbol{i}} \in \{0,1\}^{d^{(\lambda)}} \quad \text{and} \quad \sum_{\tilde{\boldsymbol{i}}\in\{0,1\}^{d(\lambda)}} \mathring{w}_{\tilde{\boldsymbol{i}}} = 1\big\}$. Let $\widetilde{\boldsymbol{W}}^{(\lambda)}(j) = (\tilde{W}_{\tilde{\boldsymbol{i}}}^{(\lambda)}(j))_{\tilde{\boldsymbol{i}}\in\{0,1\}^{d(\lambda)}}$, $j = 1,\dots,l^{(\lambda)}$, $\lambda = 1,\dots,L$, be independent multinomial random variables with mass functions

$$\prod_{\tilde{\boldsymbol{i}}=(\tilde{i}_h)_{h=1}^{d(\lambda)}\in\{0,1\}^{d(\lambda)}} \Big\{ \prod_{h=1}^{d^{(\lambda)}} \tilde{p}_{\tilde{i}_h,\nu_h^{(\lambda)}} \Big\}^{\tilde{w}_{\tilde{\boldsymbol{i}}}^{(\lambda)}(j)},$$

$(\tilde{w}_{\tilde{\boldsymbol{i}}}^{(\lambda)}(j))_{\tilde{\boldsymbol{i}}\in\{0,1\}^{d(\lambda)}} \in \widetilde{\mathcal{W}}^{(\lambda)}$, $j = 1,\dots,l^{(\lambda)}$, $\lambda = 1,\dots,L$, respectively. For $\nu = 1,\dots,N$, let $\widetilde{I}_0^{(\lambda)}(\nu) = \widetilde{I}_0^{(\lambda)}(0,\nu) = \{(\tilde{i}_h)_{h=1}^{d(\lambda)} \in \{0,1\}^{d^{(\lambda)}} | \tilde{i}_{h_\nu^{(\lambda)}} = 0\}$ for $\lambda \in \Lambda(\nu)$. Notice that

$$\Big(\Big(\sum_{\boldsymbol{i}\in I_0^{(\lambda)}(0,\nu)} W_{\boldsymbol{i}}^{(\lambda)}\Big)_{\lambda\in\Lambda(\nu)}\Big)_{\nu=1,\dots,N} \overset{\mathrm{d}}{=} \Big(\Big(\sum_{\tilde{\boldsymbol{i}}\in\widetilde{I}_0^{(\lambda)}(\nu)} \sum_{j=1}^{l^{(\lambda)}} \tilde{W}_{\tilde{\boldsymbol{i}}}^{(\lambda)}(j)\Big)_{\lambda\in\Lambda(\nu)}\Big)_{\nu=1,\dots,N}. \qquad (4.6.6)$$

Then it follows from (4.6.5) and (4.6.6) that

$$\begin{aligned}
\Delta^{(\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a})} &= E\Big[-\log E^U\Big[\prod_{\nu=1}^{N} \Big\{ \frac{\Gamma\big(\gamma_\nu U + \sum_{\lambda\in\Lambda(\nu)} \sum_{\boldsymbol{i}\in I_0^{(\lambda)}(0,\nu)} W_{\boldsymbol{i}}^{(\lambda)} + r_\nu + a_{0,\nu}\big)}{\Gamma\big(\gamma_\nu U + \sum_{\lambda\in\Lambda(\nu)} l^{(\lambda)} + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu}\big)} \\
&\quad \times \frac{\Gamma\big(\sum_{\lambda\in\Lambda(\nu)} l^{(\lambda)} + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu}\big)}{\Gamma\big(\sum_{\lambda\in\Lambda(\nu)} \sum_{\boldsymbol{i}\in I_0^{(\lambda)}(0,\nu)} W_{\boldsymbol{i}}^{(\lambda)} + r_\nu + a_{0,\nu}\big)} \\
&\quad \times \frac{\Gamma(\gamma_\nu U + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu})\Gamma(r_\nu + a_{0,\nu})}{\Gamma(\gamma_\nu U + r_\nu + a_{0,\nu})\Gamma(r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu})} \Big\}\Big]\Big] \\
&= E\Big[-\log E^U\Big[\prod_{\nu=1}^{N} \Big\{ \frac{\Gamma\big(\gamma_\nu U + \sum_{\lambda\in\Lambda(\nu)} \sum_{\tilde{\boldsymbol{i}}\in\widetilde{I}_0^{(\lambda)}(\nu)} \sum_{j=1}^{l^{(\lambda)}} \tilde{W}_{\tilde{\boldsymbol{i}}}^{(\lambda)}(j) + r_\nu + a_{0,\nu}\big)}{\Gamma\big(\gamma_\nu U + \sum_{\lambda\in\Lambda(\nu)} l^{(\lambda)} + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu}\big)} \\
&\quad \times \frac{\Gamma\big(\sum_{\lambda\in\Lambda(\nu)} l^{(\lambda)} + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu}\big)}{\Gamma\big(\sum_{\lambda\in\Lambda(\nu)} \sum_{\tilde{\boldsymbol{i}}\in\widetilde{I}_0^{(\lambda)}(\nu)} \sum_{j=1}^{l^{(\lambda)}} \tilde{W}_{\tilde{\boldsymbol{i}}}^{(\lambda)}(j) + r_\nu + a_{0,\nu}\big)} \\
&\quad \times \frac{\Gamma(\gamma_\nu U + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu})\Gamma(r_\nu + a_{0,\nu})}{\Gamma(\gamma_\nu U + r_\nu + a_{0,\nu})\Gamma(r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu})} \Big\}\Big]\Big].
\end{aligned}$$

Therefore,

$$
\begin{aligned}
\Delta^{(\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a})} = & \sum_{(((\tilde{w}^{(\lambda)}_{\tilde{\boldsymbol{i}}}(j))_{\tilde{\boldsymbol{i}}\in\{0,1\}^{d^{(\lambda)}}})_{j=1,\ldots,l^{(\lambda)}})_{\lambda=1,\ldots,L}\in(\widetilde{\mathcal{W}}^{(1)}\times\cdots\times\widetilde{\mathcal{W}}^{(1)})\times\cdots\times(\widetilde{\mathcal{W}}^{(L)}\times\cdots\times\widetilde{\mathcal{W}}^{(L)})} \Big(\\
& \Big[\prod_{\lambda=1}^{L}\prod_{j=1}^{l^{(\lambda)}}\prod_{\tilde{\boldsymbol{i}}=(\tilde{i}_h)_{h=1}^{d^{(\lambda)}}\in\{0,1\}^{d^{(\lambda)}}}\Big\{\prod_{h=1}^{d^{(\lambda)}}\tilde{p}_{\tilde{i}_h,\nu_h^{(\lambda)}}\Big\}^{\tilde{w}^{(\lambda)}_{\tilde{\boldsymbol{i}}}(j)}\Big] \\
& \times E\Big[-\log E^U\Big[\prod_{\nu=1}^{N}\Big\{\frac{\Gamma\big(\gamma_\nu U+\sum_{\lambda\in\Lambda(\nu)}\sum_{\tilde{\boldsymbol{i}}\in\tilde{I}_0^{(\lambda)}(\nu)}\sum_{j=1}^{l^{(\lambda)}}\tilde{w}^{(\lambda)}_{\tilde{\boldsymbol{i}}}(j)+r_\nu+a_{0,\nu}\big)}{\Gamma\big(\gamma_\nu U+\sum_{\lambda\in\Lambda(\nu)}l^{(\lambda)}+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu}\big)} \\
& \times\frac{\Gamma\big(\sum_{\lambda\in\Lambda(\nu)}l^{(\lambda)}+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu}\big)}{\Gamma\big(\sum_{\lambda\in\Lambda(\nu)}\sum_{\tilde{\boldsymbol{i}}\in\tilde{I}_0^{(\lambda)}(\nu)}\sum_{j=1}^{l^{(\lambda)}}\tilde{w}^{(\lambda)}_{\tilde{\boldsymbol{i}}}(j)+r_\nu+a_{0,\nu}\big)} \\
& \times\frac{\Gamma(\gamma_\nu U+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})\Gamma(r_\nu+a_{0,\nu})}{\Gamma(\gamma_\nu U+r_\nu+a_{0,\nu})\Gamma(r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})}\Big\}\Big]\Big]\Big) \\
= & \sum_{\tilde{i}_1^{(1)}(1)=0}^{1}\tilde{p}_{\tilde{i}_1^{(1)}(1),\nu_1^{(1)}}\cdots\sum_{\tilde{i}_{d^{(1)}}^{(1)}(1)=0}^{1}\tilde{p}_{\tilde{i}_{d^{(1)}}^{(1)}(1),\nu_{d^{(1)}}^{(1)}} \\
& \cdots\sum_{\tilde{i}_1^{(1)}(l^{(1)})=0}^{1}\tilde{p}_{\tilde{i}_1^{(1)}(l^{(1)}),\nu_1^{(1)}}\cdots\sum_{\tilde{i}_{d^{(1)}}^{(1)}(l^{(1)})=0}^{1}\tilde{p}_{\tilde{i}_{d^{(1)}}^{(1)}(l^{(1)}),\nu_{d^{(1)}}^{(1)}} \\
& \cdots \\
& \sum_{\tilde{i}_1^{(L)}(1)=0}^{1}\tilde{p}_{\tilde{i}_1^{(L)}(1),\nu_1^{(L)}}\cdots\sum_{\tilde{i}_{d^{(L)}}^{(L)}(1)=0}^{1}\tilde{p}_{\tilde{i}_{d^{(L)}}^{(L)}(1),\nu_{d^{(L)}}^{(L)}} \\
& \cdots\sum_{\tilde{i}_1^{(L)}(l^{(L)})=0}^{1}\tilde{p}_{\tilde{i}_1^{(L)}(l^{(L)}),\nu_1^{(L)}}\cdots\sum_{\tilde{i}_{d^{(L)}}^{(L)}(l^{(L)})=0}^{1}\tilde{p}_{\tilde{i}_{d^{(L)}}^{(L)}(l^{(L)}),\nu_{d^{(L)}}^{(L)}}E\Big[\\
& -\log E^U\Big[\prod_{\nu=1}^{N}\Big\{\frac{\Gamma\big(\gamma_\nu U+\sum_{\lambda\in\Lambda(\nu)}\sum_{\tilde{\boldsymbol{i}}\in\tilde{I}_0^{(\lambda)}(\nu)}\sum_{j=1}^{l^{(\lambda)}}\tilde{\delta}^{(\lambda)}(\tilde{\boldsymbol{i}},(\tilde{i}_h^{(\lambda)}(j))_{h=1}^{d^{(\lambda)}})+r_\nu+a_{0,\nu}\big)}{\Gamma\big(\gamma_\nu U+\sum_{\lambda\in\Lambda(\nu)}l^{(\lambda)}+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu}\big)} \\
& \times\frac{\Gamma\big(\sum_{\lambda\in\Lambda(\nu)}l^{(\lambda)}+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu}\big)}{\Gamma\big(\sum_{\lambda\in\Lambda(\nu)}\sum_{\tilde{\boldsymbol{i}}\in\tilde{I}_0^{(\lambda)}(\nu)}\sum_{j=1}^{l^{(\lambda)}}\tilde{\delta}^{(\lambda)}(\tilde{\boldsymbol{i}},(\tilde{i}_h^{(\lambda)}(j))_{h=1}^{d^{(\lambda)}})+r_\nu+a_{0,\nu}\big)} \\
& \times\frac{\Gamma(\gamma_\nu U+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})\Gamma(r_\nu+a_{0,\nu})}{\Gamma(\gamma_\nu U+r_\nu+a_{0,\nu})\Gamma(r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})}\Big\}\Big]\Big],
\end{aligned}
$$

where $\tilde{\delta}^{(\lambda)}(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{i}}')=1$ if $\tilde{\boldsymbol{i}}=\tilde{\boldsymbol{i}}'$ and $=0$ if $\tilde{\boldsymbol{i}}\neq\tilde{\boldsymbol{i}}'$ for $\tilde{\boldsymbol{i}},\tilde{\boldsymbol{i}}'\in\{0,1\}^{d^{(\lambda)}}$ for $\lambda=1,\ldots,L$. Furthermore, since

$$
\sum_{\lambda\in\Lambda(\nu)}\sum_{\tilde{\boldsymbol{i}}\in\tilde{I}_0^{(\lambda)}(\nu)}\sum_{j=1}^{l^{(\lambda)}}\tilde{\delta}^{(\lambda)}(\tilde{\boldsymbol{i}},(\tilde{i}_h^{(\lambda)}(j))_{h=1}^{d^{(\lambda)}})=\sum_{\lambda\in\Lambda(\nu)}\sum_{j=1}^{l^{(\lambda)}}\{1-\tilde{i}^{(\lambda)}_{h_\nu^{(\lambda)}}(j)\}
$$

for all $(((\tilde{i}_h^{(\lambda)}(j))_{h=1}^{d^{(\lambda)}})_{j=1,\dots,l^{(\lambda)}})_{\lambda=1,\dots,L} \in (\{0,1\}^{d^{(1)}} \times \cdots \times \{0,1\}^{d^{(1)}}) \times \cdots \times (\{0,1\}^{d^{(L)}} \times \cdots \times \{0,1\}^{d^{(L)}})$ for all $\nu = 1, \dots, N$, we can rewrite the risk difference as

$$
\begin{aligned}
\Delta^{(\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a})} = & \sum_{\tilde{i}_1^{(1)}(1)=0}^{1} \tilde{p}_{\tilde{i}_1^{(1)}(1),\nu_1^{(1)}} \cdots \sum_{\tilde{i}_{d^{(1)}}^{(1)}(1)=0}^{1} \tilde{p}_{\tilde{i}_{d^{(1)}}^{(1)}(1),\nu_{d^{(1)}}^{(1)}} \\
& \cdots \sum_{\tilde{i}_1^{(1)}(l^{(1)})=0}^{1} \tilde{p}_{\tilde{i}_1^{(1)}(l^{(1)}),\nu_1^{(1)}} \cdots \sum_{\tilde{i}_{d^{(1)}}^{(1)}(l^{(1)})=0}^{1} \tilde{p}_{\tilde{i}_{d^{(1)}}^{(1)}(l^{(1)}),\nu_{d^{(1)}}^{(1)}} \\
& \cdots \\
& \sum_{\tilde{i}_1^{(L)}(1)=0}^{1} \tilde{p}_{\tilde{i}_1^{(L)}(1),\nu_1^{(L)}} \cdots \sum_{\tilde{i}_{d^{(L)}}^{(L)}(1)=0}^{1} \tilde{p}_{\tilde{i}_{d^{(L)}}^{(L)}(1),\nu_{d^{(L)}}^{(L)}} \\
& \cdots \sum_{\tilde{i}_1^{(L)}(l^{(L)})=0}^{1} \tilde{p}_{\tilde{i}_1^{(L)}(l^{(L)}),\nu_1^{(L)}} \cdots \sum_{\tilde{i}_{d^{(L)}}^{(L)}(l^{(L)})=0}^{1} \tilde{p}_{\tilde{i}_{d^{(L)}}^{(L)}(l^{(L)}),\nu_{d^{(L)}}^{(L)}} E\Big[\\
& - \log E^U \Big[F\Big(U, (((\tilde{i}_h^{(\lambda)}(j))_{h=1}^{d^{(\lambda)}})_{j=1,\dots,l^{(\lambda)}})_{\lambda=1,\dots,L}, \Big(\sum_{\lambda\in\Lambda(\nu)} l^{(\lambda)} \Big)_{\nu=1}^{N} \Big) \Big] \Big], \qquad (4.6.7)
\end{aligned}
$$

where

$$
\begin{aligned}
F(u, \tilde{\boldsymbol{i}}, \boldsymbol{k}) = \prod_{\nu=1}^{N} \Big[& \frac{\Gamma\big(\gamma_\nu u + \sum_{\lambda\in\Lambda(\nu)} \sum_{j=1}^{l^{(\lambda)}} \{1 - \tilde{i}_{h_\nu^{(\lambda)}}^{(\lambda)}(j)\} + r_\nu + a_{0,\nu}\big)}{\Gamma(\gamma_\nu u + k_\nu + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu})} \\
& \times \frac{\Gamma(k_\nu + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu})}{\Gamma\big(\sum_{\lambda\in\Lambda(\nu)} \sum_{j=1}^{l^{(\lambda)}} \{1 - \tilde{i}_{h_\nu^{(\lambda)}}^{(\lambda)}(j)\} + r_\nu + a_{0,\nu}\big)} \\
& \times \frac{\Gamma(\gamma_\nu u + r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu})\Gamma(r_\nu + a_{0,\nu})}{\Gamma(\gamma_\nu u + r_\nu + a_{0,\nu})\Gamma(r_\nu + a_{0,\nu} + X_{\cdot,\nu} + a_{\cdot,\nu})} \Big]
\end{aligned}
$$

for $u \in (0,\infty)$, $\tilde{\boldsymbol{i}} = (((\tilde{i}_h^{(\lambda)}(j))_{h=1}^{d^{(\lambda)}})_{j=1,\dots,l^{(\lambda)}})_{\lambda=1,\dots,L} \in (\{0,1\}^{d^{(1)}} \times \cdots \times \{0,1\}^{d^{(1)}}) \times \cdots \times (\{0,1\}^{d^{(L)}} \times \cdots \times \{0,1\}^{d^{(L)}})$, and $\boldsymbol{k} = (k_\nu)_{\nu=1}^{N} \in {\mathbb{N}_0}^N$.

Now fix $\lambda^* = 1, \dots, L$, $h^* = 1, \dots, d^{(\lambda^*)}$, and $j^* = 1, \dots, l^{(\lambda^*)}$. For each $(j,h,\lambda) \in \mathbb{N} \times \mathbb{N} \times \{1,\dots,L\}$ satisfying $j \le l^{(\lambda)}$, $h \le d^{(\lambda)}$, and $(j,h,\lambda) \neq (j^*,h^*,\lambda^*)$, fix $\tilde{i}_h^{(\lambda)}(j) \in \{0,1\}$. Let $\nu^* = \nu_{h^*}^{(\lambda^*)}$. For $u \in (0,\infty)$, $\tilde{i} \in \{0,1\}$, and $\boldsymbol{k} \in {\mathbb{N}_0}^N$, let $F^*(u,\tilde{i},\boldsymbol{k})$ denote $F(u, ((\tilde{i}_h^{(\lambda)}(j))_{h=1}^{d^{(\lambda)}})_{j=1,\dots,l^{(\lambda)}})_{\lambda=1,\dots,L}, \boldsymbol{k})$ with $\tilde{i}_{h^*}^{(\lambda^*)}(j^*) = \tilde{i}$. For each $\nu = 1, \dots, N$, let $\tilde{s}_\nu^*(\tilde{i})$ denote $\sum_{\lambda\in\Lambda(\nu)} \sum_{j=1}^{l^{(\lambda)}} \{1 - \tilde{i}_{h_\nu^{(\lambda)}}^{(\lambda)}(j)\}$ with $\tilde{i}_{h^*}^{(\lambda^*)}(j^*) = \tilde{i}$ for $\tilde{i} \in \{0,1\}$. Finally, fix $\boldsymbol{k} = (k_\nu)_{\nu=1}^{N} \in {\mathbb{N}_0}^N$ such that $\tilde{s}_\nu^*(\tilde{i}) \le k_\nu \le \sum_{\lambda\in\Lambda(\nu)} l^{(\lambda)}$ for all $\nu = 1, \dots, N$ for any $\tilde{i} \in \{0,1\}$. Then, by Lemma

4.6.1,

$$
\begin{aligned}
&\sum_{\tilde{i}=0}^{1} \tilde{p}_{\tilde{i},\nu^*} E[-\log E^U[F^*(U,\tilde{i},\boldsymbol{k})]] \\
&= E[-\log E^U[F^*(U,0,\boldsymbol{k})]] + \tilde{p}_{1,\nu^*} E\Big[\log \frac{E^U[F^*(U,0,\boldsymbol{k})]}{E^U[F^*(U,1,\boldsymbol{k})]}\Big] \\
&= E\Big[-\log \frac{\int_0^\infty F^*(u,0,\boldsymbol{k})d\mu(u)}{\int_0^\infty d\mu(u)}\Big] + \tilde{p}_{1,\nu^*} E\Big[\log \frac{\int_0^\infty F^*(u,0,\boldsymbol{k})d\mu(u)}{\int_0^\infty F^*(u,1,\boldsymbol{k})d\mu(u)}\Big] \\
&= E\Big[-\log \frac{\int_0^\infty F^*(u,0,\boldsymbol{k})d\mu(u)}{\int_0^\infty d\mu(u)}\Big] + E\Big[\frac{X_{\cdot,\nu^*}}{r_{\nu^*}+X_{\cdot,\nu^*}-1} \\
&\quad \times \log\Big\{\int_0^\infty F^*(u,0,\boldsymbol{k})\frac{\gamma_{\nu^*}u+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}d\mu(u) \\
&\quad / \int_0^\infty F^*(u,1,\boldsymbol{k})\frac{\gamma_{\nu^*}u+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}d\mu(u)\Big\}\Big].
\end{aligned}
$$

In the following, if φ is a continuous function from $(0,\infty)$ to $[0,\infty)$, we write

$$
\begin{aligned}
&\int_0^\infty d\tilde{\mu}(u) = \int_0^\infty F^*(u,1,\boldsymbol{k})\frac{\gamma_{\nu^*}u+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}d\mu(u), \\
&\int_0^\infty \varphi(u)d\tilde{\mu}(u) = \int_0^\infty \varphi(u)F^*(u,1,\boldsymbol{k})\frac{\gamma_{\nu^*}u+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}d\mu(u), \quad \text{and} \\
&\widetilde{E}^U[\varphi(U)] = \int_0^\infty \varphi(u)d\tilde{\mu}(u)/\int_0^\infty d\tilde{\mu}(u).
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\sum_{\tilde{i}=0}^{1} \tilde{p}_{\tilde{i},\nu^*} E[-\log E^U[F^*(U,\tilde{i},\boldsymbol{k})]] \\
&= E\Big[-\log \frac{\int_0^\infty d\tilde{\mu}(u)}{\int_0^\infty d\mu(u)} - \log \frac{\int_0^\infty F^*(u,0,\boldsymbol{k})d\mu(u)}{\int_0^\infty d\tilde{\mu}(u)} + \frac{X_{\cdot,\nu^*}}{r_{\nu^*}+X_{\cdot,\nu^*}-1}\log \widetilde{E}^U\Big[\frac{F^*(U,0,\boldsymbol{k})}{F^*(U,1,\boldsymbol{k})}\Big]\Big] \\
&= E\Big[-\log \frac{\int_0^\infty d\tilde{\mu}(u)}{\int_0^\infty d\mu(u)} - \log \widetilde{E}^U\Big[\frac{F^*(U,0,\boldsymbol{k})}{F^*(U,1,\boldsymbol{k})}\frac{k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{\gamma_{\nu^*}U+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big] \\
&\quad + \frac{X_{\cdot,\nu^*}}{r_{\nu^*}+X_{\cdot,\nu^*}-1}\log \widetilde{E}^U\Big[\frac{F^*(U,0,\boldsymbol{k})}{F^*(U,1,\boldsymbol{k})}\Big]\Big].
\end{aligned} \tag{4.6.8}
$$

Notice that for all $u \in (0,\infty)$,

$$
\begin{aligned}
\frac{F^*(u,0,\boldsymbol{k})}{F^*(u,1,\boldsymbol{k})} &= \prod_{\nu=1}^{N} \frac{\Gamma(\gamma_\nu u+\tilde{s}^*_\nu(0)+r_\nu+a_{0,\nu})\Gamma(\tilde{s}^*_\nu(1)+r_\nu+a_{0,\nu})}{\Gamma(\gamma_\nu u+\tilde{s}^*_\nu(1)+r_\nu+a_{0,\nu})\Gamma(\tilde{s}^*_\nu(0)+r_\nu+a_{0,\nu})} \\
&= \frac{\Gamma(\gamma_{\nu^*}u+\tilde{s}^*_{\nu^*}(0)+r_{\nu^*}+a_{0,\nu^*})\Gamma(\tilde{s}^*_{\nu^*}(1)+r_{\nu^*}+a_{0,\nu^*})}{\Gamma(\gamma_{\nu^*}u+\tilde{s}^*_{\nu^*}(1)+r_{\nu^*}+a_{0,\nu^*})\Gamma(\tilde{s}^*_{\nu^*}(0)+r_{\nu^*}+a_{0,\nu^*})} \\
&= \frac{\gamma_{\nu^*}u+\tilde{s}^*_{\nu^*}(1)+r_{\nu^*}+a_{0,\nu^*}}{\tilde{s}^*_{\nu^*}(1)+r_{\nu^*}+a_{0,\nu^*}}
\end{aligned}
$$

since $\tilde{s}^*_{\nu^*}(0) = \tilde{s}^*_{\nu^*}(1) + 1$. It follows that

$$\begin{aligned}
&\log \widetilde{E}^U\Big[\frac{F^*(U,0,\boldsymbol{k})}{F^*(U,1,\boldsymbol{k})}\frac{k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{\gamma_{\nu^*}U+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big]\\
&=\log \widetilde{E}^U\Big[\frac{k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{\tilde{s}^*_{\nu^*}(1)+r_{\nu^*}+a_{0,\nu^*}}\frac{\gamma_{\nu^*}U+\tilde{s}^*_{\nu^*}(1)+r_{\nu^*}+a_{0,\nu^*}}{\gamma_{\nu^*}U+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big]\\
&=\log \widetilde{E}^U\Big[\Big\{1+\frac{k_{\nu^*}-\tilde{s}^*_{\nu^*}(1)+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{\tilde{s}^*_{\nu^*}(1)+r_{\nu^*}+a_{0,\nu^*}}\Big\}\Big\{1-\frac{k_{\nu^*}-\tilde{s}^*_{\nu^*}(1)+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{\gamma_{\nu^*}U+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big\}\Big]\\
&=\log \widetilde{E}^U\Big[1+\frac{\{k_{\nu^*}-\tilde{s}^*_{\nu^*}(1)+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1\}\gamma_{\nu^*}U}{\{\tilde{s}^*_{\nu^*}(1)+r_{\nu^*}+a_{0,\nu^*}\}(\gamma_{\nu^*}U+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1)}\Big] \qquad (4.6.9)
\end{aligned}$$

and that

$$\begin{aligned}
&\frac{X_{\cdot,\nu^*}}{r_{\nu^*}+X_{\cdot,\nu^*}-1}\log \widetilde{E}^U\Big[\frac{F^*(U,0,\boldsymbol{k})}{F^*(U,1,\boldsymbol{k})}\Big]\\
&=\frac{X_{\cdot,\nu^*}}{r_{\nu^*}+X_{\cdot,\nu^*}-1}\log \widetilde{E}^U\Big[1+\frac{\gamma_{\nu^*}U}{\tilde{s}^*_{\nu^*}(1)+r_{\nu^*}+a_{0,\nu^*}}\Big]\\
&\le \log \widetilde{E}^U\Big[1+\frac{X_{\cdot,\nu^*}}{r_{\nu^*}+X_{\cdot,\nu^*}-1}\frac{\gamma_{\nu^*}U}{\tilde{s}^*_{\nu^*}(1)+r_{\nu^*}+a_{0,\nu^*}}\Big], \qquad (4.6.10)
\end{aligned}$$

where the inequality follows since $0 \le X_{\cdot,\nu^*}/(r_{\nu^*}+X_{\cdot,\nu^*}-1) \le 1$ by assumption. By integration by parts,

$$\begin{aligned}
&(\alpha+1)\int_0^\infty u d\tilde{\mu}(u) = \int_0^\infty \Big[(\alpha+1)u^\alpha e^{-\beta u}\Big\{\prod_{\nu=1}^N \frac{\Gamma(\gamma_\nu u+r_\nu+a_{0,\nu})\Gamma(r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})}{\Gamma(\gamma_\nu u+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})\Gamma(r_\nu+a_{0,\nu})}\Big\}\\
&\quad\times F^*(u,1,\boldsymbol{k})\frac{\gamma_{\nu^*}u+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big]du\\
&=\int_0^\infty \Big(u^{\alpha+1}e^{-\beta u}\Big\{\prod_{\nu=1}^N \frac{\Gamma(\gamma_\nu u+r_\nu+a_{0,\nu})\Gamma(r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})}{\Gamma(\gamma_\nu u+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})\Gamma(r_\nu+a_{0,\nu})}\Big\}\\
&\quad\times F^*(u,1,\boldsymbol{k})\frac{\gamma_{\nu^*}u+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\\
&\quad\times\Big[\beta+\sum_{\nu=1}^N \gamma_\nu\{\psi(\gamma_\nu u+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})-\psi(\gamma_\nu u+r_\nu+a_{0,\nu})\}\\
&\quad-\sum_{\nu=1}^N \gamma_\nu\{\psi(\gamma_\nu u+\tilde{s}^*_\nu(1)+r_\nu+a_{0,\nu})-\psi(\gamma_\nu u+k_\nu+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})\\
&\quad+\psi(\gamma_\nu u+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})-\psi(\gamma_\nu u+r_\nu+a_{0,\nu})\}\\
&\quad-\frac{\gamma_{\nu^*}}{\gamma_{\nu^*}u+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big]\Big)du\\
&=\int_0^\infty \Big(u^2\Big[\beta+\sum_{\nu=1}^N \gamma_\nu\{\psi(\gamma_\nu u+k_\nu+r_\nu+a_{0,\nu}+X_{\cdot,\nu}+a_{\cdot,\nu})-\psi(\gamma_\nu u+\tilde{s}^*_\nu(1)+r_\nu+a_{0,\nu})\}\\
&\quad-\frac{\gamma_{\nu^*}}{\gamma_{\nu^*}u+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big]\Big)d\tilde{\mu}(u).
\end{aligned}$$

Therefore, by Lemma 7 of Hamura and Kubokawa (2020b),

$$
\begin{aligned}
(\alpha+1)\int_0^\infty u d\tilde{\mu}(u) &\geq \int_0^\infty u^2[\beta+\gamma_{\nu^*}\{\psi(\gamma_{\nu^*}u+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1) \\
&\quad -\psi(\gamma_{\nu^*}u+\tilde{s}^*_{\nu^*}(1)+r_{\nu^*}+a_{0,\nu^*})\}]d\tilde{\mu}(u) \\
&\geq \int_0^\infty u^2\Big\{\beta+\gamma_{\nu^*}\frac{k_{\nu^*}-\tilde{s}^*_{\nu^*}(1)+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{\gamma_{\nu^*}u+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big\}d\tilde{\mu}(u) \\
&\geq (\beta+\gamma_{\nu^*})\int_0^\infty u^2\frac{k_{\nu^*}-\tilde{s}^*_{\nu^*}(1)+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{\gamma_{\nu^*}u+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}d\tilde{\mu}(u),
\end{aligned}
$$

where the third inequality follows since $k_{\nu^*}\geq \tilde{s}^*_{\nu^*}(0)=\tilde{s}^*_{\nu^*}(1)+1$, and this implies that

$$
\widetilde{E}^U\Big[\frac{U^2}{\gamma_{\nu^*}U+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big]\leq \frac{(\alpha+1)/(\beta+\gamma_{\nu^*})}{k_{\nu^*}-\tilde{s}^*_{\nu^*}(1)+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\widetilde{E}^U[U]. \tag{4.6.11}
$$

When $X_{\nu^*}\geq 1$, we have, by (4.3.4),

$$
\Big\{\frac{(\alpha+1)\gamma_{\nu^*}}{\beta+\gamma_{\nu^*}}-a_{\cdot,\nu^*}\Big\}(r_{\nu^*}-1)\leq X_{\nu^*}\Big\{-\frac{(\alpha+1)\gamma_{\nu^*}}{\beta+\gamma_{\nu^*}}-k_{\nu^*}-a_{0,\nu^*}\Big\},
$$

which implies that

$$
\gamma_{\nu^*}\frac{(\alpha+1)/(\beta+\gamma_{\nu^*})}{k_{\nu^*}-\tilde{s}^*_{\nu^*}(1)+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\leq 1-\frac{X_{\cdot,\nu^*}}{r_{\nu^*}+X_{\cdot,\nu^*}-1}\frac{k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{k_{\nu^*}-\tilde{s}^*_{\nu^*}(1)+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1} \tag{4.6.12}
$$

since $k_{\nu^*}\geq \tilde{s}^*_{\nu^*}(1)+1$. From (4.6.11) and (4.6.12), it follows that when $X_{\nu^*}\geq 1$,

$$
\begin{aligned}
&\widetilde{E}^U\Big[\frac{\gamma_{\nu^*}U^2}{\gamma_{\nu^*}U+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big] \\
&\leq \gamma_{\nu^*}\frac{(\alpha+1)/(\beta+\gamma_{\nu^*})}{k_{\nu^*}-\tilde{s}^*_{\nu^*}(1)+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\widetilde{E}^U[U] \\
&\leq \Big\{1-\frac{X_{\cdot,\nu^*}}{r_{\nu^*}+X_{\cdot,\nu^*}-1}\frac{k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}{k_{\nu^*}-\tilde{s}^*_{\nu^*}(1)+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big\}\widetilde{E}^U[U],
\end{aligned}
$$

which can be rewritten as

$$
\begin{aligned}
&\frac{X_{\cdot,\nu^*}}{r_{\nu^*}+X_{\cdot,\nu^*}-1}\widetilde{E}^U\Big[\frac{U}{k_{\nu^*}-\tilde{s}^*_{\nu^*}(1)+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big] \\
&\leq \widetilde{E}^U\Big[\frac{U}{k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big(1-\frac{\gamma_{\nu^*}U}{\gamma_{\nu^*}U+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1}\Big)\Big]
\end{aligned}
$$

or

$$
\begin{aligned}
&\widetilde{E}^U\Big[\frac{X_{\cdot,\nu^*}}{r_{\nu^*}+X_{\cdot,\nu^*}-1}\frac{\gamma_{\nu^*}U}{\tilde{s}^*_{\nu^*}(1)+r_{\nu^*}+a_{0,\nu^*}}\Big] \\
&\leq \widetilde{E}^U\Big[\frac{\{k_{\nu^*}-\tilde{s}^*_{\nu^*}(1)+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1\}\gamma_{\nu^*}U}{\{\tilde{s}^*_{\nu^*}(1)+r_{\nu^*}+a_{0,\nu^*}\}(\gamma_{\nu^*}U+k_{\nu^*}+r_{\nu^*}+a_{0,\nu^*}+X_{\cdot,\nu^*}+a_{\cdot,\nu^*}-1)}\Big].
\end{aligned} \tag{4.6.13}
$$

Thus, by (4.6.8), (4.6.9), (4.6.10), and (4.6.13),

$$\sum_{\tilde{i}=0}^{1} \tilde{p}_{\tilde{i},\nu^*} E[-\log E^U[F^*(U,\tilde{i},\boldsymbol{k})]] < E\Big[-\log \frac{\int_0^\infty d\tilde{\mu}(u)}{\int_0^\infty d\mu(u)}\Big]$$
$$= E[-\log E^U[F^*(U,1,\boldsymbol{k}-\boldsymbol{e}_{\nu^*}^{(N)})]]. \tag{4.6.14}$$

Finally, applying (4.6.14) to (4.6.7) sequentially, we obtain

$$\Delta^{(\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{a}_0,\boldsymbol{a})} < \cdots < 0.$$

This completes the proof. □

Proof of Theorem 4.5.1. By (4.5.1), (4.5.2), and (4.5.3),

$$R(\boldsymbol{p},\hat{g}^{(\pi)}) = E\Big[\log\Big\{\prod_{\nu=1}^{n}\Big({p_{0,\nu}}^{s_\nu}\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{Y_{i,\nu}}\Big)\Big\}\Big]$$
$$+ E\Big[-\log\frac{\int_D \pi(\boldsymbol{p})\big\{\prod_{\nu=1}^{N}\big({p_{0,\nu}}^{s_\nu+r_\nu}\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{Y_{i,\nu}+X_{i,\nu}}\big)\big\}d\boldsymbol{p}}{\int_D \pi(\boldsymbol{p})\big\{\prod_{\nu=1}^{N}\big({p_{0,\nu}}^{r_\nu}\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{X_{i,\nu}}\big)\big\}d\boldsymbol{p}}\Big], \tag{4.6.15}$$

where $Y_{1,\nu} = \cdots = Y_{m_\nu,\nu} = 0$ if $\nu \in \{1,\dots,N\} \cap [n+1,\infty)$. The first term on the right of (4.6.15) is

$$E\Big[\log\Big\{\prod_{\nu=1}^{n}\Big({p_{0,\nu}}^{s_\nu}\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{Y_{i,\nu}}\Big)\Big\}\Big]$$
$$= \sum_{\nu=1}^{n}\Big(s_\nu\log p_{0,\nu} + \sum_{i=1}^{m_\nu} s_\nu\frac{p_{i,\nu}}{p_{0,\nu}}\log p_{i,\nu}\Big)$$
$$= \sum_{\nu=1}^{n} s_\nu\sum_{k=1}^{\infty}\frac{1}{k}\Big(-{p_{\cdot,\nu}}^{k} + {p_{\cdot,\nu}}^{k}\sum_{i=1}^{m_\nu} k\frac{p_{i,\nu}}{p_{\cdot,\nu}}\log p_{i,\nu}\Big)$$
$$= \sum_{\nu=1}^{n} s_\nu\sum_{k=1}^{\infty}\frac{1}{k}\sum_{(w_i)_{i=1}^{m_\nu}\in\mathcal{W}_{\nu,k}}\frac{k!}{\prod_{i=1}^{m_\nu}w_i!}\Big\{-\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{w_i} + \Big(\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{w_i}\Big)\sum_{i=1}^{m_\nu} w_i\log p_{i,\nu}\Big\}. \tag{4.6.16}$$

On the other hand, since t_ν is a constant if $\nu \in \{1,\dots,N\} \cap [n+1,\infty)$,

$$E\Big[-\log\frac{\int_D \pi(\boldsymbol{p})\big\{\prod_{\nu=1}^{N}\big({p_{0,\nu}}^{s_\nu+r_\nu}\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{Y_{i,\nu}+X_{i,\nu}}\big)\big\}d\boldsymbol{p}}{\int_D \pi(\boldsymbol{p})\big\{\prod_{\nu=1}^{N}\big({p_{0,\nu}}^{r_\nu}\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{X_{i,\nu}}\big)\big\}d\boldsymbol{p}}\Big] = \int_0^1\Big\{\frac{\partial}{\partial\tau}E[-\log G(\tau,\boldsymbol{Z}(\tau))]\Big\}d\tau$$
$$= \int_0^1 E\Big[\sum_{\nu=1}^{n} t_\nu'(\tau)\Big\{\sum_{k=1}^{Z_{\cdot,\nu}(\tau)}\frac{1}{t_\nu(\tau)+k-1} + \log p_{0,\nu}\Big\}\{-\log G(\tau,\boldsymbol{Z}(\tau))\} - \frac{\frac{\partial G}{\partial\tau}(\tau,\boldsymbol{Z}(\tau))}{G(\tau,\boldsymbol{Z}(\tau))}\Big]d\tau, \tag{4.6.17}$$

where

$$G(\tau,((z_{i,\nu})_{i=1}^{m_\nu})_{\nu=1,\dots,N}) = \int_D \pi(\boldsymbol{p})\Big[\prod_{\nu=1}^{N}\Big\{{p_{0,\nu}}^{t_\nu(\tau)}\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{z_{i,\nu}}\Big\}\Big]d\boldsymbol{p}$$

for $((z_{i,\nu})_{i=1}^{m_\nu})_{\nu=1,\dots,N} \in {\mathbb{N}_0}^{m_1} \times \cdots \times {\mathbb{N}_0}^{m_N}$ and where $Z_{\cdot,\nu}(\tau) = \sum_{i=1}^{m_\nu} Z_{i,\nu}(\tau)$ for $\nu = 1, \dots, N$ for each $\tau \in [0,1]$.

Fix $\tau \in [0,1]$. Then

$$
\begin{aligned}
&E\Big[\Big\{\frac{\partial G}{\partial \tau}(\tau, \boldsymbol{Z}(\tau))\Big\}/G(\tau, \boldsymbol{Z}(\tau))\Big] \\
&= E\Big[\int_D \pi(\boldsymbol{p})\Big[\Big\{\sum_{\nu=1}^{N} {t_\nu}'(\tau)\log p_{0,\nu}\Big\}\prod_{\nu=1}^{N}\Big\{{p_{0,\nu}}^{t_\nu(\tau)}\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{Z_{i,\nu}(\tau)}\Big\}\Big]d\boldsymbol{p}/G(\tau, \boldsymbol{Z}(\tau))\Big] \\
&= -\sum_{\nu=1}^{n}{t_\nu}'(\tau)\sum_{k=1}^{\infty}\frac{1}{k}E\Big[\int_D \pi(\boldsymbol{p})\Big[{p_{\cdot,\nu}}^k\prod_{\nu'=1}^{N}\Big\{{p_{0,\nu'}}^{t_{\nu'}(\tau)}\prod_{i=1}^{m_{\nu'}}{p_{i,\nu'}}^{Z_{i,\nu'}(\tau)}\Big\}\Big]d\boldsymbol{p}/G(\tau, \boldsymbol{Z}(\tau))\Big] \\
&= -\sum_{\nu=1}^{n}{t_\nu}'(\tau)\sum_{k=1}^{\infty}\frac{1}{k}\sum_{(w_i)_{i=1}^{m_\nu}\in\mathcal{W}_{\nu,k}}\frac{k!}{\prod_{i=1}^{m_\nu}w_i!}E\Big[\int_D \pi(\boldsymbol{p})\Big[\Big(\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{w_i}\Big)\prod_{\nu'=1}^{N}\Big\{{p_{0,\nu'}}^{t_{\nu'}(\tau)}\prod_{i=1}^{m_{\nu'}}{p_{i,\nu'}}^{Z_{i,\nu'}(\tau)}\Big\}\Big]d\boldsymbol{p} \\
&\quad /G(\tau, \boldsymbol{Z}(\tau))\Big]. \qquad (4.6.18)
\end{aligned}
$$

On the other hand, by Lemmas 2.1 and 2.2 of Hamura and Kubokawa (2019a), we have for any $\nu = 1, \dots, n$,

$$
\begin{aligned}
&E\Big[\Big\{\sum_{k=1}^{Z_{\cdot,\nu}(\tau)}\frac{1}{t_\nu(\tau)+k-1}+\log p_{0,\nu}\Big\}\{-\log G(\tau, \boldsymbol{Z}(\tau))\}\Big] \\
&= E\Big[\Big\{\sum_{k=1}^{Z_{\cdot,\nu}(\tau)}\frac{1}{k}\frac{Z_{\cdot,\nu}(\tau)\cdots\{Z_{\cdot,\nu}(\tau)-k+1\}}{\{t_\nu(\tau)+Z_{\cdot,\nu}(\tau)-1\}\cdots\{t_\nu(\tau)+Z_{\cdot,\nu}(\tau)-k\}}+\log p_{0,\nu}\Big\}\{-\log G(\tau, \boldsymbol{Z}(\tau))\}\Big] \\
&= \sum_{k=1}^{\infty}\frac{1}{k}{p_{\cdot,\nu}}^k E[E[-\log G(\tau, \boldsymbol{Z}(\tau))|\boldsymbol{Z}_\cdot(\tau)+k\boldsymbol{e}_\nu^{(N)}]-\{-\log G(\tau, \boldsymbol{Z}(\tau))\}],
\end{aligned}
$$

where $\boldsymbol{Z}_\cdot(\tau) = (Z_{\cdot,\nu}(\tau))_{\nu=1}^N$. Now, fix $k \in \mathbb{N}$. Let $\boldsymbol{W}_\nu$, $\nu = 1, \dots, N$, be mutually independent multinomial variables such that for each $\nu = 1, \dots, N$, the probability mass function of $\boldsymbol{W}_\nu | Z_{\cdot,\nu}(\tau)$ is given by

$$
\frac{Z_{\cdot,\nu}(\tau)!}{\prod_{i=1}^{m_\nu}w_{i,\nu}!}\prod_{i=1}^{m_\nu}\Big(\frac{p_{i,\nu}}{p_{\cdot,\nu}}\Big)^{w_{i,\nu}}
$$

for $(w_{i,\nu})_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,Z_{\cdot,\nu}(\tau)}$. Let $\boldsymbol{W}_\nu^*$, $\nu = 1, \dots, N$, be independent multinomial variable with mass functions

$$
\frac{k!}{\prod_{i=1}^{m_\nu}w_{i,\nu}^*!}\prod_{i=1}^{m_\nu}\Big(\frac{p_{i,\nu}}{p_{\cdot,\nu}}\Big)^{w_{i,\nu}^*},
$$

$(w^*_{i,\nu})_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,k}$, $\nu = 1, \ldots, N$, respectively. Then, for any $\nu = 1, \ldots, N$,

$$
\begin{aligned}
&E[-\log G(\tau, \boldsymbol{Z}(\tau))|\boldsymbol{Z}_\cdot(\tau) + k\boldsymbol{e}_\nu^{(N)}] \\
&= E[-\log G(\tau, (\boldsymbol{W}_{\nu'} + \delta^{(N)}_{\nu,\nu'}\boldsymbol{W}^*_{\nu'})_{\nu'=1,\ldots,N})|\boldsymbol{Z}_\cdot(\tau)] \\
&= \sum_{(w^*_{i,\nu})_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,k}} \frac{k!}{\prod_{i=1}^{m_\nu} w^*_{i,\nu}!} \Big\{ \prod_{i=1}^{m_\nu} \Big(\frac{p_{i,\nu}}{p_{\cdot,\nu}}\Big)^{w^*_{i,\nu}} \Big\} E[-\log G(\tau, (\boldsymbol{W}_{\nu'} + \delta^{(N)}_{\nu,\nu'}(w^*_{i,\nu})_{i=1}^{m_\nu})_{\nu'=1,\ldots,N})|\boldsymbol{Z}_\cdot(\tau)] \\
&= \sum_{(w^*_{i,\nu})_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,k}} \frac{k!}{\prod_{i=1}^{m_\nu} w^*_{i,\nu}!} \Big\{ \prod_{i=1}^{m_\nu} \Big(\frac{p_{i,\nu}}{p_{\cdot,\nu}}\Big)^{w^*_{i,\nu}} \Big\} E[-\log G(\tau, (\boldsymbol{Z}_{\nu'}(\tau) + \delta^{(N)}_{\nu,\nu'}(w^*_{i,\nu})_{i=1}^{m_\nu})_{\nu'=1,\ldots,N})|\boldsymbol{Z}_\cdot(\tau)]
\end{aligned}
$$

and therefore

$$
\begin{aligned}
&E[E[-\log G(\tau, \boldsymbol{Z}(\tau))|\boldsymbol{Z}_\cdot(\tau) + k\boldsymbol{e}_\nu^{(N)}]] \\
&= \frac{1}{{p_{\cdot,\nu}}^k} \sum_{(w_i)_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,k}} \frac{k!}{\prod_{i=1}^{m_\nu} w_i!} \Big(\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big) E[-\log G(\tau, (\boldsymbol{Z}_{\nu'}(\tau) + \delta^{(N)}_{\nu,\nu'}(w_i)_{i=1}^{m_\nu})_{\nu'=1,\ldots,N})].
\end{aligned}
$$

Since k is arbitrarily chosen, it follows that

$$
\begin{aligned}
&E\Big[\Big\{ \sum_{k=1}^{Z_{\cdot,\nu}(\tau)} \frac{1}{t_\nu(\tau) + k - 1} + \log p_{0,\nu} \Big\} \{-\log G(\tau, \boldsymbol{Z}(\tau))\}\Big] \\
&= \sum_{k=1}^{\infty} \frac{1}{k} \Big\{ \sum_{(w_i)_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,k}} \frac{k!}{\prod_{i=1}^{m_\nu} w_i!} \Big(\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big) E[-\log G(\tau, (\boldsymbol{Z}_{\nu'}(\tau) + \delta^{(N)}_{\nu,\nu'}(w_i)_{i=1}^{m_\nu})_{\nu'=1,\ldots,N})] \\
&\quad - \sum_{(w_i)_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,k}} \frac{k!}{\prod_{i=1}^{m_\nu} w_i!} \Big(\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big) E[-\log G(\tau, \boldsymbol{Z}(\tau))] \Big\} \\
&= \sum_{k=1}^{\infty} \frac{1}{k} \sum_{(w_i)_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,k}} \frac{k!}{\prod_{i=1}^{m_\nu} w_i!} \Big(\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big) E\Big[-\log \frac{G(\tau, (\boldsymbol{Z}_{\nu'}(\tau) + \delta^{(N)}_{\nu,\nu'}(w_i)_{i=1}^{m_\nu})_{\nu'=1,\ldots,N})}{G(\tau, \boldsymbol{Z}(\tau))} \Big].
\end{aligned}
\tag{4.6.19}
$$

Finally, combining (4.6.15), (4.6.16), (4.6.17), (4.6.18), and (4.6.19), we obtain

$$
\begin{aligned}
R(\boldsymbol{p},\hat{g}^{(\pi)}) &= \int_0^1 \Big[\sum_{\nu=1}^{n} t_\nu{}'(\tau)\sum_{k=1}^{\infty}\frac{1}{k}\sum_{(w_i)_{i=1}^{m_\nu}\in\mathcal{W}_{\nu,k}}\Big\{\frac{k!}{\prod_{i=1}^{m_\nu} w_i!}\\
&\quad\times\Big(-\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i}+\Big(\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i}\Big)\sum_{i=1}^{m_\nu} w_i\log p_{i,\nu}\\
&\quad+\Big(\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i}\Big)E\Big[-\log\frac{G(\tau,(\boldsymbol{Z}_{\nu'}(\tau)+\delta^{(N)}_{\nu,\nu'}(w_i)_{i=1}^{m_\nu})_{\nu'=1,\dots,N})}{G(\tau,\boldsymbol{Z}(\tau))}\Big]\\
&\quad+E\Big[\int_D \pi(\boldsymbol{p})\Big[\Big(\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i}\Big)\prod_{\nu'=1}^{N}\Big\{{p_{0,\nu'}}^{t_{\nu'}(\tau)}\prod_{i=1}^{m_{\nu'}}{p_{i,\nu'}}^{Z_{i,\nu'}(\tau)}\Big\}\Big]d\boldsymbol{p}/G(\tau,\boldsymbol{Z}(\tau))\Big]\Big)\Big\}\Big]d\tau\\
&= \int_0^1 \Big[\sum_{\nu=1}^{n} t_\nu{}'(\tau)\sum_{k=1}^{\infty}\frac{1}{k}\sum_{(w_i)_{i=1}^{m_\nu}\in\mathcal{W}_{\nu,k}}\Big\{\frac{k!}{\prod_{i=1}^{m_\nu} w_i!}\\
&\quad\times E\Big[L^{\mathrm{KL}}\Big(\frac{G(\tau,(\boldsymbol{Z}_{\nu'}(\tau)+\delta^{(N)}_{\nu,\nu'}(w_i)_{i=1}^{m_\nu})_{\nu'=1,\dots,N})}{G(\tau,\boldsymbol{Z}(\tau))},\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{w_i}\Big)\Big]\Big\}\Big]d\tau.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&R(\boldsymbol{p},\hat{g}^{(\pi)})\\
&=\int_0^1\Big\{\sum_{\nu=1}^{n} t_\nu{}'(\tau)\sum_{k=1}^{\infty}\frac{1}{k}\sum_{(w_i)_{i=1}^{m_\nu}\in\mathcal{W}_{\nu,k}}\frac{k!}{\prod_{i=1}^{m_\nu} w_i!}E\Big[L^{\mathrm{KL}}\Big(E_\pi\Big[\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{w_i}\Big|\boldsymbol{Z}(\tau)\Big],\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{w_i}\Big)\Big]\Big\}d\tau,
\end{aligned}
$$

which is the desired result. □

Proof of Corollary 4.5.1. By Theorem 4.5.1, we have

$$
\begin{aligned}
&R(\boldsymbol{p},\hat{g}^{(\pi_{M,\tilde{\boldsymbol{\gamma}},\boldsymbol{a}_0,\boldsymbol{a}})})-R(\boldsymbol{p},\hat{g}^{(\pi_{\boldsymbol{a}_0,\boldsymbol{a}})})\\
&=\int_0^1\Big\{\sum_{\nu=1}^{n} t_\nu{}'(\tau)\sum_{k=1}^{\infty}\frac{1}{k}\sum_{(w_i)_{i=1}^{m_\nu}\in\mathcal{W}_{\nu,k}}\frac{k!}{\prod_{i=1}^{m_\nu} w_i!}E\Big[L^{\mathrm{KL}}\Big(E_{\pi_{M,\tilde{\boldsymbol{\gamma}},\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{w_i}\Big|\boldsymbol{Z}(\tau)\Big],\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{w_i}\Big)\\
&\quad-L^{\mathrm{KL}}\Big(E_{\pi_{\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{w_i}\Big|\boldsymbol{Z}(\tau)\Big],\prod_{i=1}^{m_\nu}{p_{i,\nu}}^{w_i}\Big)\Big]\Big\}d\tau.
\end{aligned}
$$

Fix $\tau \in [0,1]$, $\nu = 1, \dots, n$, and $k \in \mathbb{N}$. Then

$$\begin{aligned}
&\sum_{(w_i)_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,k}} \frac{k!}{\prod_{i=1}^{m_\nu} w_i!} E\Big[L^{\mathrm{KL}}\Big(E_{\pi_{M,\tilde{\boldsymbol{\gamma}},\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big| \boldsymbol{Z}(\tau) \Big], \prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big) \\
&\quad - L^{\mathrm{KL}}\Big(E_{\pi_{\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big| \boldsymbol{Z}(\tau) \Big], \prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big) \Big] \\
&= \sum_{(w_i)_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,k}} \frac{k!}{\prod_{i=1}^{m_\nu} w_i!} E\Big[E_{\pi_{M,\tilde{\boldsymbol{\gamma}},\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big| \boldsymbol{Z}(\tau) \Big] - E_{\pi_{\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big| \boldsymbol{Z}(\tau) \Big] \\
&\quad - \Big(\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big) \log \Big\{ E_{\pi_{M,\tilde{\boldsymbol{\gamma}},\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big| \boldsymbol{Z}(\tau) \Big] / E_{\pi_{\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big| \boldsymbol{Z}(\tau) \Big] \Big\} \Big].
\end{aligned}$$

Note that

$$\begin{aligned}
&\sum_{(w_i)_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,k}} \frac{k!}{\prod_{i=1}^{m_\nu} w_i!} E\Big[E_{\pi_{M,\tilde{\boldsymbol{\gamma}},\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big| \boldsymbol{Z}(\tau) \Big] - E_{\pi_{\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big| \boldsymbol{Z}(\tau) \Big] \Big] \\
&= E[E_{\pi_{M,\tilde{\boldsymbol{\gamma}},\boldsymbol{a}_0,\boldsymbol{a}}}[{p_{\cdot,\nu}}^k | \boldsymbol{Z}(\tau)] - E_{\pi_{\boldsymbol{a}_0,\boldsymbol{a}}}[{p_{\cdot,\nu}}^k | \boldsymbol{Z}(\tau)]]
\end{aligned}$$

and that for all $(w_i)_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,k}$,

$$\begin{aligned}
&E_{\pi_{M,\tilde{\boldsymbol{\gamma}},\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big| \boldsymbol{Z}(\tau) \Big] / E_{\pi_{\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big| \boldsymbol{Z}(\tau) \Big] \\
&= \frac{\displaystyle\int_0^\infty \Big\{ \prod_{\nu'=1}^{N} \frac{\Gamma(\tilde{\gamma}_{\nu'}(u) + t_{\nu'}(\tau) + a_{0,\nu'})}{\Gamma(\tilde{\gamma}_{\nu'}(u) + t_{\nu'}(\tau) + a_{0,\nu'} + Z_{\cdot,\nu'}(\tau) + a_{\cdot,\nu'} + \delta_{\nu,\nu'}^{(N)} k)} \Big\} dM(u)}{\displaystyle\int_0^\infty \Big\{ \prod_{\nu'=1}^{N} \frac{\Gamma(\tilde{\gamma}_{\nu'}(u) + t_{\nu'}(\tau) + a_{0,\nu'})}{\Gamma(\tilde{\gamma}_{\nu'}(u) + t_{\nu'}(\tau) + a_{0,\nu'} + Z_{\cdot,\nu'}(\tau) + a_{\cdot,\nu'})} \Big\} dM(u)} \\
&\quad \Big/ \frac{\displaystyle\prod_{\nu'=1}^{N} \frac{\Gamma(t_{\nu'}(\tau) + a_{0,\nu'})}{\Gamma(t_{\nu'}(\tau) + a_{0,\nu'} + Z_{\cdot,\nu'}(\tau) + a_{\cdot,\nu'} + \delta_{\nu,\nu'}^{(N)} k)}}{\displaystyle\prod_{\nu'=1}^{N} \frac{\Gamma(t_{\nu'}(\tau) + a_{0,\nu'})}{\Gamma(t_{\nu'}(\tau) + a_{0,\nu'} + Z_{\cdot,\nu'}(\tau) + a_{\cdot,\nu'})}} \\
&= E_{\pi_{M,\tilde{\boldsymbol{\gamma}},\boldsymbol{a}_0,\boldsymbol{a}}}[{p_{\cdot,\nu}}^k | \boldsymbol{Z}(\tau)] / E_{\pi_{\boldsymbol{a}_0,\boldsymbol{a}}}[{p_{\cdot,\nu}}^k | \boldsymbol{Z}(\tau)],
\end{aligned}$$

where $Z_{\cdot,\nu'}(\tau) = \sum_{i=1}^{m_{\nu'}} Z_{i,\nu'}(\tau)$ for $\nu' = 1, \dots, N$. It follow that

$$\begin{aligned}
&\sum_{(w_i)_{i=1}^{m_\nu} \in \mathcal{W}_{\nu,k}} \frac{k!}{\prod_{i=1}^{m_\nu} w_i!} E\Big[L^{\mathrm{KL}}\Big(E_{\pi_{M,\tilde{\boldsymbol{\gamma}},\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big| \boldsymbol{Z}(\tau) \Big], \prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big) \\
&\quad - L^{\mathrm{KL}}\Big(E_{\pi_{\boldsymbol{a}_0,\boldsymbol{a}}}\Big[\prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big| \boldsymbol{Z}(\tau) \Big], \prod_{i=1}^{m_\nu} {p_{i,\nu}}^{w_i} \Big) \Big] \\
&= E[L^{\mathrm{KL}}(E_{\pi_{M,\tilde{\boldsymbol{\gamma}},\boldsymbol{a}_0,\boldsymbol{a}}}[{p_{\cdot,\nu}}^k | \boldsymbol{Z}(\tau)], {p_{\cdot,\nu}}^k) - L^{\mathrm{KL}}(E_{\pi_{\boldsymbol{a}_0,\boldsymbol{a}}}[{p_{\cdot,\nu}}^k | \boldsymbol{Z}(\tau)], {p_{\cdot,\nu}}^k)].
\end{aligned}$$

This completes the proof. □

Chapter 5

Bayesian Predictive Density Estimation for a Chi-Squared Model Using Information from a Normal Observation with Unknown Mean and Variance

5.1 Introduction

Suppose that $\boldsymbol{X}$ and V are independently distributed according to the normal and Chi-squared distributions $\mathrm{N}_p(\boldsymbol{\mu}, (r_0/\eta)I_p)$ and $(r_0'/\eta)\chi^2(n_1)$ with densities

$$p(\boldsymbol{x}|\boldsymbol{\mu},\eta) = \frac{(\eta/r_0)^{p/2}}{(2\pi)^{p/2}} \exp\Big(-\frac{\eta/r_0}{2}||\boldsymbol{x}-\boldsymbol{\mu}||^2\Big), \quad \boldsymbol{x}\in\mathbb{R}^p, \quad \text{and}$$

$$p_1(v|\eta) = \frac{(1/2)^{n_1/2}}{\Gamma(n_1/2)} v^{n_1/2-1}(\eta/r_0')^{n_1/2} \exp\Big(-\frac{\eta/r_0'}{2}v\Big), \quad v\in(0,\infty),$$

respectively, for known $p\in\mathbb{N}=\{1,2,\dots\}$ and $r_0, r_0', n_1 > 0$ and unknown $\boldsymbol{\mu}\in\mathbb{R}^p$ and $\eta\in(0,\infty)$. Suppose that for known $s_0', n_2 > 0$, W is an unobservable Chi-squared variable with distribution $(s_0'/\eta)\chi^2(n_2)$ which is independent of $(\boldsymbol{X}, V)$. We consider the problem of estimating the density of W, namely

$$p_2(w|\eta) = \frac{(1/2)^{n_2/2}}{\Gamma(n_2/2)} w^{n_2/2-1}(\eta/s_0')^{n_2/2} \exp\Big(-\frac{\eta/s_0'}{2}w\Big), \quad w\in(0,\infty),$$

on the basis of the observation of $(\boldsymbol{X}, V)$ under the Kullback-Leibler loss. The risk function of a predictive density $\hat{p}_2(\cdot;\boldsymbol{X},V)$ is

$$R((\boldsymbol{\mu},\eta),\hat{p}_2) = E^{(\boldsymbol{X},V,W)}_{(\boldsymbol{\mu},\eta)}\Big[\log\frac{p_2(W|\eta)}{\hat{p}_2(W;\boldsymbol{X},V)}\Big].$$

Such a situation arises, for example, if $\boldsymbol{X}_1,\dots,\boldsymbol{X}_{N_1}$ and $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_{N_2}$ are independently distributed as $\mathrm{N}_p(\boldsymbol{\mu},(1/\eta)I_p)$ and if we want to estimate the predictive density of $\sum_{i=1}^{N_2}||\boldsymbol{Y}_i -$

$\overline{\boldsymbol{Y}}||^2$, where $\overline{\boldsymbol{Y}} = (1/N_2)\sum_{i=1}^{N_2} \boldsymbol{Y}_i$, based on the sufficient statistics $\overline{\boldsymbol{X}} = (1/N_1)\sum_{i=1}^{N_1} \boldsymbol{X}_i$ and $\sum_{i=1}^{N_1} ||\boldsymbol{X}_i - \overline{\boldsymbol{X}}||^2$. On the other hand, since r_0'/η and n_1 can be any positive real numbers, V may be viewed as a gamma variable. Throughout this chapter, however, we assume that $r_0 = r_0' = s_0' = 1$ for simplicity.

For a prior $\pi(\boldsymbol{\mu}, \eta)$ for the unknown parameters $(\boldsymbol{\mu}, \eta)$, the associated Bayesian predictive density $\hat{p}_2^{(\pi)}(\cdot; \boldsymbol{X}, V)$ is given by

$$\begin{aligned}\hat{p}_2^{(\pi)}(w; \boldsymbol{x}, v) &= E_\pi^{(\boldsymbol{\mu},\eta)|(\boldsymbol{X},V)}[p_2(w|\eta)|(\boldsymbol{X}, V) = (\boldsymbol{x}, v)] \\ &= \frac{(1/2)^{n_2/2}}{\Gamma(n_2/2)} w^{n_2/2-1} E_\pi^{\eta|(\boldsymbol{X},V)}\Big[\eta^{n_2/2} \exp\Big(-\frac{\eta}{2}v\Big)\Big|(\boldsymbol{X}, V) = (\boldsymbol{x}, v)\Big].\end{aligned}$$

The Jeffreys prior for the model where only V is observed is $\pi_0(\boldsymbol{\mu}, \eta) = \eta^{-1}$, which corresponds to the unbiased estimator V/n_1 of the variance $1/\eta$ in the sense that $1/E_{\pi_0}[\eta|\boldsymbol{X}, V] = V/n_1$. As in Liang and Barron (2004), it can be shown that $\hat{p}_2^{(\pi_0)}(\cdot; \boldsymbol{X}, V)$ is uniformly optimal among the predictive densities which are equivariant with respect to the transformations of Section 2 of Stein (1964). In particular, for any $a_0 < n_1/2$, it improves upon $\hat{p}_2^{(\pi_{a_0})}(\cdot; \boldsymbol{X}, V)$ for $\pi_{a_0}(\boldsymbol{\mu}, \eta) = \eta^{-a_0-1}$, which, when $a_0 = -p/2$, coincides with the Jeffreys prior for the present model where both $\boldsymbol{X}$ and V are observed. In this chapter, as in Maruyama and Strawderman (2012), we consider the hierarchical shrinkage prior

$$\pi_{b,a}(\boldsymbol{\mu}, \eta) = \int_0^1 \pi_{b,a}(\boldsymbol{\mu}, \gamma, \eta) d\gamma, \tag{5.1.1}$$

where

$$\begin{aligned}\pi_{b,a}(\boldsymbol{\mu}, \gamma, \eta) &= \mathrm{N}_p(\boldsymbol{\mu}|\boldsymbol{0}_p, [\{(1-\gamma)/\gamma\}/\eta] I_p)(1-\gamma)^{b-1}\gamma^{-a-1}\eta^{-a-1} \\ &= \frac{(1-\gamma)^{b-p/2-1}\gamma^{p/2-a-1}\eta^{p/2-a-1}}{(2\pi)^{p/2}} \exp\Big(-\frac{\eta}{2}\frac{\gamma}{1-\gamma}||\boldsymbol{\mu}||^2\Big)\end{aligned}$$

for $b > 0$ and $a < p/2$. We compare the two predictive densities $\hat{p}_2^{(\pi_0)}(\cdot; \boldsymbol{X}, V)$ and $\hat{p}_2^{(\pi_{b,a})}(\cdot; \boldsymbol{X}, V)$. In particular, in Section 5.3, we obtain conditions under which $\hat{p}_2^{(\pi_{b,a})}(\cdot; \boldsymbol{X}, V)$ dominates $\hat{p}_2^{(\pi_0)}(\cdot; \boldsymbol{X}, V)$.

An important feature of the problem is that the distribution of $\boldsymbol{X}$ depends on the unknown location parameter $\boldsymbol{\mu}$ while the distribution of W does not depend on $\boldsymbol{\mu}$. As will be shown later, $\hat{p}_2^{(\pi_0)}(\cdot; \boldsymbol{X}, V)$ is a function only of V but $\hat{p}_2^{(\pi_{b,a})}(\cdot; \boldsymbol{X}, V)$ does depend on $\boldsymbol{X}$. Thus, dominance of $\hat{p}_2^{(\pi_{b,a})}(\cdot; \boldsymbol{X}, V)$ over $\hat{p}_2^{(\pi_0)}(\cdot; \boldsymbol{X}, V)$ is analogous to the result of Stein (1964) that when estimating the variance $1/\eta$ under the standardized squared error loss, the unbiased estimator V/n_1 can be improved upon by using additional information from $\boldsymbol{X}$.

Although Stein (1964) considered a truncated estimator, it was shown by Brewster and Zidek (1974) that the unbiased estimator is dominated by a smooth generalized Bayes estimator also. Kubokawa (1994) showed that these improved estimators can be derived through the unified method of Integral Expression of Risk Difference (IERD). Maruyama (1998) gave a class of priors including that of Brewster and Zidek (1974) to improve on the unbiased estimator when the mean of the normal distribution is equal to zero. Related hierarchical priors have been shown to be useful in estimating location parameters in the presence of an unknown scale parameter (Maruyama and Strawderman (2005, 2020a, 2020b)).

Bayesian predictive densities have been widely studied in the literature since Aitchison (1975) showed their superiority to plug-in predictive densities. Komaki (2001) proved for a normal model with unknown mean that the Bayesian predictive density against the uniform prior is dominated by that against a shrinkage prior as in estimation problems. Parallels between estimation and prediction were investigated by George, Liang and Xu (2006, 2012) and Brown, George and Xu (2008) in terms of minimaxity and admissibility. Kato (2009) and Boisbunon and Maruyama (2014) considered the case of unknown mean and variance. Prediction for a 2×2 Wishart model was considered by Komaki (2009). Prediction for a gamma model when the scale parameter is restricted to an interval was considered by L'Moudden, Marchand, Kortbi and Strawderman (2017).

5.2 Bayesian Predictive Densities

In this section, the Bayesian predictive densities with respect to the priors π_0 and $\pi_{b,a}$ given in Section 5.1 are derived. The choice of the hyperparameter b in $\pi_{b,a}$ is discussed.

We first consider $\hat{p}_2^{(\pi_0)}(\cdot; \boldsymbol{X}, V)$ for the noninformative prior $\pi_0(\boldsymbol{\mu}, \eta) = \eta^{-1}$.

Proposition 5.2.1 *The Bayesian predictive density $\hat{p}_2^{(\pi_0)}(\cdot; \boldsymbol{X}, V)$ is given by*

$$\hat{p}_2^{(\pi_0)}(w; \boldsymbol{X}, V) = \frac{1}{B(n_1/2, n_2/2)} \frac{V^{n_1/2} w^{n_2/2-1}}{(V+w)^{(n_1+n_2)/2}}.$$

We note that this predictive density does not depend on $\boldsymbol{X}$. Moreover, it is identical to the predictive density with respect to the observation $V \sim (1/\eta)\chi^2(n_1)$ and the prior $\eta \sim \eta^{-1}$. Its superiority to the corresponding plug-in predictive density is discussed in Aitchison (1975).

On the other hand, $\hat{p}_2^{(\pi_{b,a})}(\cdot; \boldsymbol{X}, V)$ actually depends on the normal variable $\boldsymbol{X}$.

Proposition 5.2.2 *The Bayesian predictive density $\hat{p}_2^{(\pi_{b,a})}(\cdot; \boldsymbol{X}, V)$ for the hierarchical prior $\pi_{b,a}$ in (5.1.1) is given by*

$$\hat{p}_2^{(\pi_{b,a})}(w|\boldsymbol{X}, V) = \frac{w^{n_2/2-1}}{B(n_1/2+p/2-a, n_2/2)} \frac{\displaystyle\int_0^1 \frac{(1-\gamma)^{b-1}\gamma^{p/2-a-1}}{(V+w+\gamma||\boldsymbol{X}||^2)^{(n_1+n_2)/2+p/2-a}} d\gamma}{\displaystyle\int_0^1 \frac{(1-\gamma)^{b-1}\gamma^{p/2-a-1}}{(V+\gamma||\boldsymbol{X}||^2)^{n_1/2+p/2-a}} d\gamma}.$$

Because of the integrals in the above expression, the risk function of $\hat{p}_2^{(\pi_{b,a})}(\cdot; \boldsymbol{X}, V)$ is hard to evaluate in general.

If we choose $b = n_1/2$, then the integral in the denominator can be simplified to

$$\frac{B(n_1/2, p/2-a)}{V^{n_1/2}(V+||\boldsymbol{X}||^2)^{p/2-a}} \tag{5.2.1}$$

by Lemma 2 of Boisbunon and Maruyama (2014). This choice corresponds to that in Section 2.1 of Maruyama and Strawderman (2005). On the other hand, in this case, the integral in the numerator becomes, by Lemma 2 of Boisbunon and Maruyama (2014),

$$\frac{1}{(V+w)^{(n_1+n_2)/2}(V+w+||\boldsymbol{X}||^2)^{p/2-a}} \int_0^1 (1-\gamma)^{n_1/2-1}\gamma^{p/2-a-1}\Big(1 - \frac{||\boldsymbol{X}||^2}{V+w+||\boldsymbol{X}||^2}\gamma\Big)^{n_2/2} d\gamma \tag{5.2.2}$$

and involves the hypergeometric function, which shows the greater complexity of the prediction problem. However, the above integral can be evaluated as in the proof of Lemma A2 of Boisbunon and Maruyama (2014), which is crucial for our proof of Theorem 5.3.1 for general n_2.

There is another case where we can analytically examine the risk function of $\hat{p}_2^{(\pi_{b,a})}(\cdot; \boldsymbol{X}, V)$. Suppose that $b = 1$. Then $\hat{p}_2^{(\pi_{b,a})}(\cdot; \boldsymbol{X}, V)$ becomes, by Lemma 5.5.1 in the Appendix,

$$\begin{aligned}\hat{p}_2^{(\pi_{1,a})}(w; \boldsymbol{X}, V) &= \frac{w^{n_2/2-1}}{B(n_1/2+p/2-a, n_2/2)} \frac{\displaystyle\int_0^1 \frac{\gamma^{p/2-a-1}}{(V+w+\gamma||\boldsymbol{X}||^2)^{(n_1+n_2)/2+p/2-a}} d\gamma}{\displaystyle\int_0^1 \frac{\gamma^{p/2-a-1}}{(V+\gamma||\boldsymbol{X}||^2)^{n_1/2+p/2-a}} d\gamma} \\ &= \hat{p}_2^{(\pi_0)}(w; \boldsymbol{X}, V) \frac{\displaystyle\int_0^{||\boldsymbol{X}||^2/(V+w+||\boldsymbol{X}||^2)} \frac{\gamma^{p/2-a-1}(1-\gamma)^{(n_1+n_2)/2-1}}{B(p/2-a,(n_1+n_2)/2)} d\gamma}{\displaystyle\int_0^{||\boldsymbol{X}||^2/(V+||\boldsymbol{X}||^2)} \frac{\gamma^{p/2-a-1}(1-\gamma)^{n_1/2-1}}{B(p/2-a,n_1/2)} d\gamma}. \end{aligned} \tag{5.2.3}$$

Therefore,

$$\lim_{||\boldsymbol{x}||^2 \to \infty} \hat{p}_2^{(\pi_{1,a})}(w; \boldsymbol{x}, V) = \hat{p}_2^{(\pi_0)}(w; \boldsymbol{X}, V),$$

which shows that we can apply the method of IERD of Kubokawa (1994). In order to prove Theorem 5.3.2 given later, we use the expression (5.2.3) and apply the argument of Kato (2009). Finally, it is interesting to note that for $a = p/2-1$, the Bayesian predictive density $\hat{p}_2^{(\pi_{1,a})}(\cdot; \boldsymbol{X}, V)$ can be expressed in closed form as

$$\hat{p}_2^{(\pi_{1,p/2-1})}(w; \boldsymbol{X}, V) = \frac{n_1+n_2}{n_1\Gamma(n_2/2)} \frac{V^{n_1/2} w^{n_2/2-1}}{(V+w)^{(n_1+n_2)/2}} \frac{1-\Big(\dfrac{V+w}{V+w+||\boldsymbol{X}||^2}\Big)^{(n_1+n_2)/2}}{1-\Big(\dfrac{V}{V+||\boldsymbol{X}||^2}\Big)^{n_1/2}}. \tag{5.2.4}$$

That we can obtain this simple predictive density is one of the important features of our prediction problem.

5.3 Dominance Conditions

In this section, we provide sufficient conditions for $\hat{p}_2^{(\pi_{b,a})}(\cdot; \boldsymbol{X}, V)$ to dominate $\hat{p}_2^{(\pi_0)}(\cdot; \boldsymbol{X}, V)$ in the two cases $b = n_1/2$ and $b = 1$. In particular, conditions on the other hyperparameter a are obtained.

We first consider the case $b = n_1/2$. Let

$$(c_1, c_2) = \begin{cases} \Big(\dfrac{\Gamma(n_1/2)\Gamma((n_1+n_2)/2+p/2-a)}{\Gamma((n_1+n_2)/2)\Gamma(n_1/2+p/2-a)} - 1, 1\Big), & \text{if } n_2 \le 2, \\ \Big(\dfrac{p/2-a}{(n_1+n_2)/2-1}, \dfrac{n_2}{2}\Big), & \text{if } n_2 > 2. \end{cases}$$

Theorem 5.3.1 *Suppose that $b = n_1/2$ and $a < p/2$. If the inequality*

$$\frac{p/2-a}{c_2}\Big\{\psi\Big(\frac{n_1+n_2}{2}+\frac{p}{2}\Big)-\psi\Big(\frac{n_1}{2}+\frac{p}{2}\Big)\Big\} \\ \le \int_0^1 (1-\rho)^{(n_1+n_2)/2+p/2-1}\frac{1}{\rho}\Big\{1-\frac{1}{(1+c_1\rho)^{(n_1+n_2)/2}}\Big\}d\rho \tag{5.3.1}$$

is satisfied, then $R((\boldsymbol{\mu},\eta),\hat{p}_2^{(\pi_{b,a})}) \le R((\boldsymbol{\mu},\eta),\hat{p}_2^{(\pi_0)})$ for all $\boldsymbol{\mu}\in\mathbb{R}^p$ and $\eta\in(0,\infty)$. Equality can hold only if $\boldsymbol{\mu}=\mathbf{0}_p$.

The integral appearing in the right-hand side of (5.3.1) is not a big problem. First, we can numerically calculate the integral since it does not involve the unknown parameters. Second, the integral can actually be evaluated analytically to obtain simpler sufficient conditions.

Corollary 5.3.1 *Assume that $b = n_1/2$ and $a < p/2$.*

(i) *If*

$$\psi\Big(\frac{n_1+n_2}{2}+\frac{p}{2}\Big)-\psi\Big(\frac{n_1}{2}+\frac{p}{2}\Big) \\ \le \frac{c_2}{p/2-a}\frac{n_1+n_2+p+2}{n_1+n_2+p}\Big[1-\frac{1}{\{1+2c_1/(n_1+n_2+p+2)\}^{(n_1+n_2)/2}}\Big],$$

then $\hat{p}_2^{(\pi_{b,a})}(\cdot;\boldsymbol{X},V)$ dominates $\hat{p}_2^{(\pi_0)}(\cdot;\boldsymbol{X},V)$.

(ii) *Suppose that either $n_2 \le 2$ and*

$$\psi\Big(\frac{n_1+n_2}{2}+\frac{p}{2}\Big)-\psi\Big(\frac{n_1}{2}+\frac{p}{2}\Big) < \frac{(n_1+n_2)c_2}{n_1+n_2+p}\psi\Big(\frac{n_1+n_2}{2}\Big)-\psi\Big(\frac{n_1}{2}\Big)$$

or $n_2 > 2$ and

$$\psi\Big(\frac{n_1+n_2}{2}+\frac{p}{2}\Big)-\psi\Big(\frac{n_1}{2}+\frac{p}{2}\Big) < \frac{(n_1+n_2)c_2}{n_1+n_2+p}\frac{2}{n_1+n_2-2}.$$

Then $\hat{p}_2^{(\pi_{b,a})}(\cdot;\boldsymbol{X},V)$ dominates $\hat{p}_2^{(\pi_0)}(\cdot;\boldsymbol{X},V)$ for any $0 \le a < p/2$ sufficiently close to $p/2$.

When $n_2 = 2$, condition (5.3.1) is actually necessary and sufficient for $\hat{p}_2^{(\pi_{n_1/2,a})}(\cdot;\boldsymbol{X},V)$ to dominate $\hat{p}_2^{(\pi_0)}(\cdot;\boldsymbol{X},V)$.

Corollary 5.3.2 *Assume that $b = n_1/2$, $a < p/2$, and $n_2 = 2$.*

(i) *$\hat{p}_2^{(\pi_{b,a})}(\cdot;\boldsymbol{X},V)$ dominates $\hat{p}_2^{(\pi_0)}(\cdot;\boldsymbol{X},V)$ if and only if*

$$\frac{p/2-a}{n_1/2+p/2} \le \int_0^1 (1-\rho)^{n_1/2+p/2}\frac{1}{\rho}\Big(1-\frac{1}{[1+\{(p/2-a)/(n_1/2)\}\rho]^{n_1/2+1}}\Big)d\rho. \tag{5.3.2}$$

(ii) *When $n_1 = 2$, $\hat{p}_2^{(\pi_{b,a})}(\cdot;\boldsymbol{X},V)$ dominates $\hat{p}_2^{(\pi_0)}(\cdot;\boldsymbol{X},V)$ if and only if $0 \le a < p/2$.*

Next we consider the case of $b = 1$.

Theorem 5.3.2 *Assume that $b = 1$, $0 \leq a < p/2$, and $n_1 > 2$. Then $R((\boldsymbol{\mu}, \eta), \hat{p}_2^{(\pi_{b,a})}) \leq R((\boldsymbol{\mu}, \eta), \hat{p}_2^{(\pi_0)})$ for all $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\eta \in (0, \infty)$. Equality holds if and only if $\boldsymbol{\mu} = \mathbf{0}_p$ and $a = 0$.*

For the special case of (5.2.4), we can obtain another sufficient condition.

Theorem 5.3.3 *Suppose that $b = 1$ and $a = p/2 - 1$ for $p \geq 2$. Then $R((\boldsymbol{\mu}, \eta), \hat{p}_2^{(\pi_{b,a})}) \leq R((\boldsymbol{\mu}, \eta), \hat{p}_2^{(\pi_0)})$ for all $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\eta \in (0, \infty)$. Equality holds if and only if $p = 2$ and $\boldsymbol{\mu} = \mathbf{0}_p$.*

5.4 Simulation Study

In this section, we investigate through simulation the numerical performance of the risk functions of the Bayesian predictive densities $\hat{p}_2^{\mathrm{O}}(\cdot; \boldsymbol{X}, V) = \hat{p}_2^{(\pi_0)}(\cdot; \boldsymbol{X}, V)$ and $\hat{p}_2^{(b,a)}(\cdot; \boldsymbol{X}, V) = \hat{p}_2^{(\pi_{b,a})}(\cdot; \boldsymbol{X}, V)$ for $b \in \{n_1/2, 1\}$ and $a \in \{0, p/2 - 1\}$. We consider the following cases: (i) $(n_1, n_2) = (3, 3)$; (ii) $(n_1, n_2) = (3, 5)$; (iii) $(n_1, n_2) = (5, 3)$; (iv) $(n_1, n_2) = (5, 5)$. We set $p = 14$. When $b = 1$, the conditions of Theorem 5.3.2 are satisfied for both $a = 0$ and $a = p/2 - 1$. On the other hand, when $b = n_1/2$, the condition of part (i) of Corollary 5.3.1 is satisfied if $a = p/2 - 1$ but not if $a = 0$, which can be verified numerically.

The risk function of $\hat{p}_2^{\mathrm{O}}(\cdot; \boldsymbol{X}, V)$ is a constant independent of the unknown parameters $(\boldsymbol{\mu}, \eta)$ while that of $\hat{p}_2^{(b,a)}(\cdot; \boldsymbol{X}, V)$ depends on $(\boldsymbol{\mu}, \eta)$ only through $\theta = \eta ||\boldsymbol{\mu}||^2$. For $\theta \in \{0, 20, 40, 60\}$, we obtain approximated values of the risk function of $\hat{p}_2^{(b,a)}(\cdot; \boldsymbol{X}, V)$ by the Monte Carlo simulation with $100,000$ replications. The integrals are calculated via the Monte Carlo simulation with $10,000$ replications.

The results are illustrated in Figure 5.1. The constant risk of $\hat{p}_2^{\mathrm{O}}(\cdot; \boldsymbol{X}, V)$ is not the same for each case. For each $b \in \{n_1/2, 1\}$, the risk values of $\hat{p}_2^{(b,p/2-1)}(\cdot; \boldsymbol{X}, V)$ are smaller than those of $\hat{p}_2^{(b,0)}(\cdot; \boldsymbol{X}, V)$ when $\theta = 0$ but larger when $\theta = 60$. The risk values of $\hat{p}_2^{(n_1/2,0)}(\cdot; \boldsymbol{X}, V)$ are larger than those of $\hat{p}_2^{(1,0)}(\cdot; \boldsymbol{X}, V)$ when $\theta = 0$ but smaller when $\theta = 60$; on the other hand, the risk values of $\hat{p}_2^{(n_1/2,p/2-1)}(\cdot; \boldsymbol{X}, V)$ are close to those of $\hat{p}_2^{(1,p/2-1)}(\cdot; \boldsymbol{X}, V)$ for all $\theta \in \{0, 20, 40, 60\}$. Since by Theorem 5.3.2 the values of the risk functions of $\hat{p}_2^{\mathrm{O}}(\cdot; \boldsymbol{X}, V)$ and $\hat{p}_2^{(1,0)}(\cdot; \boldsymbol{X}, V)$ at $\theta = 0$ coincide, that the blue triangles are not on the horizontal lines when $\theta = 0$ will be due to Monte Carlo error. Finally, $\hat{p}_2^{(n_1/2,0)}(\cdot; \boldsymbol{X}, V)$ does not seem to dominate $\hat{p}_2^{\mathrm{O}}(\cdot; \boldsymbol{X}, V)$ with the value of a too small, for the black squares lie far above the horizontal lines when $\theta = 0$.

5.5 Appendix

Useful lemmas are given in Section 5.5.1. Propositions 5.2.1 and 5.2.2, Theorems 5.3.1, 5.3.2, and 5.3.3, and Corollaries 5.3.1 and 5.3.2 are proved in Section 5.5.2. Let $\mathbb{N}_0 = \{0, 1, 2, \dots\}$.

5.5.1 Lemmas

Lemma 5.5.1 *For any $\xi_1, \xi_2, c > 0$, it holds that*

$$\int_0^1 \frac{\gamma^{\xi_1 - 1}}{(1 + c\gamma)^{\xi_1 + \xi_2}} d\gamma = \frac{1}{c^{\xi_1}} \int_0^{c/(1+c)} \gamma^{\xi_1 - 1} (1 - \gamma)^{\xi_2 - 1} d\gamma.$$

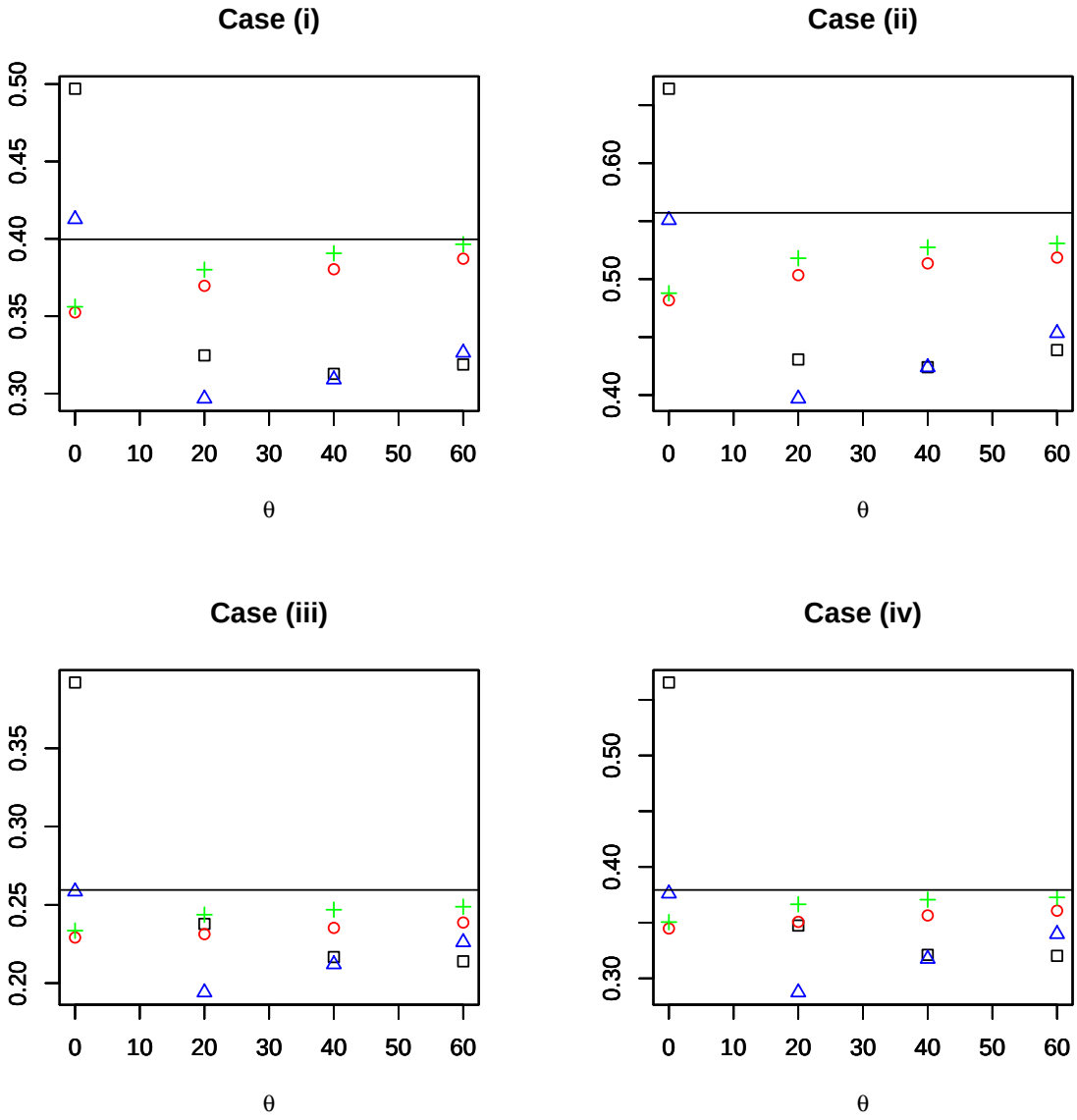

Figure 5.1: Risks of the predictive densities $\hat{p}_2^{\mathrm{O}}(\cdot; \boldsymbol{X}, V)$ and $\hat{p}_2^{(b,a)}(\cdot; \boldsymbol{X}, V)$ in the following cases: (i) $(n_1, n_2) = (3, 3)$; (ii) $(n_1, n_2) = (3, 5)$; (iii) $(n_1, n_2) = (5, 3)$; (iv) $(n_1, n_2) = (5, 5)$. We set $p = 14$. The horizontal lines show the constant risk of $\hat{p}_2^{\mathrm{O}}(\cdot; \boldsymbol{X}, V)$. The black squares, red circles, blue triangles, and green pluses correspond to $(b, a) = (n_1/2, 0), (n_1/2, p/2-1), (1, 0), (1, p/2-1)$, respectively.

Proof. We have

$$\int_0^1 \frac{\gamma^{\xi_1-1}}{(1+c\gamma)^{\xi_1+\xi_2}}d\gamma = \int_0^c \frac{\lambda^{\xi_1-1}/c^{\xi_1}}{(1+\lambda)^{\xi_1+\xi_2}}d\lambda = \frac{1}{c^{\xi_1}}\int_0^{c/(1+c)} \gamma^{\xi_1-1}(1-\gamma)^{\xi_2-1}d\gamma,$$

which is the desired result. □

Lemma 5.5.2 *For any $\xi_1, \xi_{2,1}, \xi_{2,2}, c > 0$, we have*

$$\begin{aligned}
&\int_0^1 (1-\gamma)^{\xi_{2,1}-1}\gamma^{\xi_1-1}\Big(1-\frac{c}{1+c}\gamma\Big)^{\xi_{2,2}}d\gamma \\
&\geq B(\xi_{2,1}+\xi_{2,2},\xi_1) \times \begin{cases} 1+\Big\{\dfrac{\Gamma(\xi_{2,1})\Gamma(\xi_{2,1}+\xi_{2,2}+\xi_1)}{\Gamma(\xi_{2,1}+\xi_{2,2})\Gamma(\xi_{2,1}+\xi_1)}-1\Big\}\dfrac{1}{1+c}, & \text{if } \xi_{2,2}\leq 1, \\ \Big(1+\dfrac{\xi_1}{\xi_{2,1}+\xi_{2,2}-1}\dfrac{1}{1+c}\Big)^{\xi_{2,2}}, & \text{if } \xi_{2,2}>1. \end{cases}
\end{aligned}$$

Proof. Suppose first that $\xi_{2,2} \leq 1$. Then, by Lemma 3 of Boisbunon and Maruyama (2014), we have for all $\gamma \in (0,1)$

$$\Big(1-\frac{c}{1+c}\gamma\Big)^{\xi_{2,2}} \geq (1-\gamma)^{\xi_{2,2}} + \Big(1-\frac{c}{1+c}\Big)\{1-(1-\gamma)^{\xi_{2,2}}\}.$$

Therefore,

$$\begin{aligned}
&\int_0^1 (1-\gamma)^{\xi_{2,1}-1}\gamma^{\xi_1-1}\Big(1-\frac{c}{1+c}\gamma\Big)^{\xi_{2,2}}d\gamma \\
&\geq B(\xi_{2,1}+\xi_{2,2},\xi_1) + \frac{1}{1+c}\{B(\xi_{2,1},\xi_1) - B(\xi_{2,1}+\xi_{2,2},\xi_1)\} \\
&= B(\xi_{2,1}+\xi_{2,2},\xi_1)\Big[1+\Big\{\frac{\Gamma(\xi_{2,1})\Gamma(\xi_{2,1}+\xi_{2,2}+\xi_1)}{\Gamma(\xi_{2,1}+\xi_{2,2})\Gamma(\xi_{2,1}+\xi_1)}-1\Big\}\frac{1}{1+c}\Big].
\end{aligned}$$

Next suppose that $\xi_{2,2} > 1$. Then, by Jensen's inequality, it follows that

$$\begin{aligned}
&\int_0^1 (1-\gamma)^{\xi_{2,1}-1}\gamma^{\xi_1-1}\Big(1-\frac{c}{1+c}\gamma\Big)^{\xi_{2,2}}d\gamma \\
&= B(\xi_{2,1}+\xi_{2,2},\xi_1)\int_0^1 \frac{(1-\gamma)^{\xi_{2,1}-1}\gamma^{\xi_1-1}}{B(\xi_{2,1}+\xi_{2,2},\xi_1)}\Big(1-\gamma+\gamma-\frac{c}{1+c}\gamma\Big)^{\xi_{2,2}}d\gamma \\
&= B(\xi_{2,1}+\xi_{2,2},\xi_1)\int_0^1 \frac{(1-\gamma)^{\xi_{2,1}+\xi_{2,2}-1}\gamma^{\xi_1-1}}{B(\xi_{2,1}+\xi_{2,2},\xi_1)}\Big(1+\frac{1}{1+c}\frac{\gamma}{1-\gamma}\Big)^{\xi_{2,2}}d\gamma \\
&\geq B(\xi_{2,1}+\xi_{2,2},\xi_1)\Big(1+\frac{1}{1+c}\frac{\xi_1}{\xi_{2,1}+\xi_{2,2}-1}\Big)^{\xi_{2,2}}
\end{aligned}$$

This completes the proof. □

Lemma 5.5.3 *For any $\xi_1, \xi_2, c > 0$, we have*

$$\int_0^1 \{\log(1+c\rho)\}\frac{\rho^{\xi_1-1}(1-\rho)^{\xi_2-1}}{B(\xi_1,\xi_2)}d\rho = \int_0^1 \frac{(1-\rho)^{\xi_1+\xi_2-1}}{\rho}\Big\{1-\frac{1}{(1+c\rho)^{\xi_1}}\Big\}d\rho.$$

Proof. The hypergeometric function F satisfies

$$F(a', b'; c'; z') = \sum_{s=0}^{\infty} \frac{\Gamma(a'+s)\Gamma(b'+s)\Gamma(c')}{\Gamma(a')\Gamma(b')\Gamma(c'+s)s!}(z')^s$$
$$= \frac{\Gamma(c')}{\Gamma(b')\Gamma(c'-b')} \int_0^1 \frac{t^{b'-1}(1-t)^{c'-b'-1}}{(1-z't)^{a'}} dt$$

for $a' > 0$, $c' > b' > 0$, and $z' < 0$. Therefore,

$$\int_0^1 \{\log(1+c\rho)\} \frac{\rho^{\xi_1-1}(1-\rho)^{\xi_2-1}}{B(\xi_1,\xi_2)} d\rho - \int_0^1 \frac{(1-\rho)^{\xi_1+\xi_2-1}}{\rho} \Big\{1 - \frac{1}{(1+c\rho)^{\xi_1}}\Big\} d\rho$$
$$= \int_0^1 \Big(\int_0^1 \frac{c\rho}{1+c\rho t} dt \Big) \frac{\rho^{\xi_1-1}(1-\rho)^{\xi_2-1}}{B(\xi_1,\xi_2)} d\rho - \int_0^1 \frac{(1-\rho)^{\xi_1+\xi_2-1}}{\rho} \Big\{ \int_0^1 \frac{\xi_1 c\rho}{(1+c\rho t)^{\xi_1+1}} dt \Big\} d\rho$$
$$= \int_0^1 \Big\{ \frac{1}{B(\xi_1,\xi_2)} \int_0^1 \frac{c\rho^{\xi_1}(1-\rho)^{\xi_2-1}}{1+ct\rho} d\rho - \xi_1 \int_0^1 \frac{c(1-\rho)^{\xi_1+\xi_2-1}}{(1+ct\rho)^{\xi_1+1}} d\rho \Big\} dt$$
$$= \frac{\xi_1 c}{\xi_1+\xi_2} \int_0^1 \{F(1, \xi_1+1; \xi_1+\xi_2+1; -ct) - F(\xi_1+1, 1; \xi_1+\xi_2+1; -ct)\} dt = 0,$$

which proves Lemma 5.5.3. □

Lemma 5.5.4 *For any $\xi_1, \xi_2 > 0$, we have*

$$\psi(\xi_1) - \psi(\xi_2) = \sum_{i=0}^{\infty} \frac{\xi_1 - \xi_2}{(i+\xi_1)(i+\xi_2)}.$$

Proof. Let $C = \lim_{i\to\infty} \big(\sum_{j=1}^{i} 1/j - \log i \big)$. Then

$$\psi(\xi_1) - \psi(\xi_2) = \Big\{ -\frac{1}{\xi_1} - C + \sum_{i=1}^{\infty} \Big(\frac{1}{i} - \frac{1}{i+\xi_1} \Big) \Big\} - \Big\{ -\frac{1}{\xi_2} - C + \sum_{i=1}^{\infty} \Big(\frac{1}{i} - \frac{1}{i+\xi_2} \Big) \Big\}$$
$$= \sum_{i=0}^{\infty} \Big(\frac{1}{i+\xi_2} - \frac{1}{i+\xi_1} \Big) = \sum_{i=0}^{\infty} \frac{\xi_1 - \xi_2}{(i+\xi_1)(i+\xi_2)},$$

which shows Lemma 5.5.4. □

Lemma 5.5.5 *Let $\xi_1 > 0$ and $1 < \xi_{2,1} < \xi_{2,2}$. Let, for $i \in \{1, 2\}$,*

$$F_i(q) = \int_0^q \frac{\gamma^{\xi_1-1}(1-\gamma)^{\xi_{2,i}-1}}{B(\xi_1, \xi_{2,i})} d\gamma, \quad q \in (0,1).$$

(i) *${F_2}^{-1}(\omega)/{F_1}^{-1}(\omega)$ is nondecreasing in $\omega \in (0,1)$.*

(ii) *There exist $0 < \underline{\omega} < \overline{\omega} < 1$ such that ${F_2}^{-1}(\omega)/{F_1}^{-1}(\omega)$ is strictly increasing in $\omega \in (\underline{\omega}, \overline{\omega})$.*

Proof. Part (i) follows from Lemma 2 of Kato (2009). For part (ii), we need only show that ${F_2}^{-1}(\omega)/{F_1}^{-1}(\omega)$ is not constant in $\omega \in (0,1)$. Suppose that there exists $C_0 \in \mathbb{R}$ such that ${F_2}^{-1}(\omega)/{F_1}^{-1}(\omega) = C_0$ for all $\omega \in (0,1)$. Then $C_0 = \lim_{\omega\to 1}\{{F_2}^{-1}(\omega)/{F_1}^{-1}(\omega)\} = 1$. Therefore, we have that ${F_2}^{-1} = {F_1}^{-1}$ and hence that $F_2 = F_1$. This is a contradiction. □

Lemma 5.5.6 *Let $h \in \mathbb{N}$ and $\xi \geq 1$. Then for all $\tau > 0$,*

$$\frac{\partial}{\partial \tau} \frac{\Gamma((h+1)\tau)\Gamma(\tau+\xi)}{\Gamma(\tau)\Gamma((h+1)\tau+\xi)} \begin{cases} = 0, & \text{if } \xi = 1, \\ < 0, & \text{if } \xi > 1. \end{cases}$$

Proof. By Gauss's multiplication formula, we have

$$\begin{aligned}
&\frac{\Gamma((h+1)\tau)\Gamma(\tau+\xi)}{\Gamma(\tau)\Gamma((h+1)\tau+\xi)} \\
&= \frac{\Gamma(\tau+\xi)}{\Gamma(\tau)} \frac{(2\pi)^{\{1-(h+1)\}/2}(h+1)^{(h+1)\tau-1/2}}{(2\pi)^{\{1-(h+1)\}/2}(h+1)^{(h+1)\tau+\xi-1/2}} \frac{\prod_{i=0}^{h}\Gamma(\tau+i/(h+1))}{\prod_{i=0}^{h}\Gamma(\tau+\xi/(h+1)+i/(h+1))} \\
&= \frac{1}{(h+1)^{\xi}} \frac{\Gamma(\tau+\xi)}{\Gamma(\tau+(\xi+h)/(h+1))} \prod_{i=1}^{h} \frac{\Gamma(\tau+i/(h+1))}{\Gamma(\tau+(\xi+i-1)/(h+1))}
\end{aligned}$$

for all $\tau > 0$. Therefore, by Lemma 5.5.4,

$$\begin{aligned}
&\frac{\partial}{\partial \tau} \log \frac{\Gamma((h+1)\tau)\Gamma(\tau+\xi)}{\Gamma(\tau)\Gamma((h+1)\tau+\xi)} \\
&= \psi(\tau+\xi) - \psi\Big(\tau + \frac{\xi+h}{h+1}\Big) + \sum_{i=1}^{h} \Big\{ \psi\Big(\tau + \frac{i}{h+1}\Big) - \psi\Big(\tau + \frac{\xi+i-1}{h+1}\Big) \Big\} \\
&= \frac{(\xi-1)h}{h+1} \sum_{j=0}^{\infty} \Big\{ \frac{1}{(j+\tau+\xi)\big(j+\tau+\frac{\xi+h}{h+1}\big)} - \frac{1}{h} \sum_{i=1}^{h} \frac{1}{\big(j+\tau+\frac{i}{h+1}\big)\big(j+\tau+\frac{\xi+i-1}{h+1}\big)} \Big\}
\end{aligned}$$

for all $\tau > 0$. Fix $j \in \mathbb{N}_0$ and $\tau > 0$. Then, by Jensen's inequality,

$$\begin{aligned}
&\frac{1}{(j+\tau+\xi)\big(j+\tau+\frac{\xi+h}{h+1}\big)} - \frac{1}{h} \sum_{i=1}^{h} \frac{1}{\big(j+\tau+\frac{i}{h+1}\big)\big(j+\tau+\frac{\xi+i-1}{h+1}\big)} \\
&\leq \frac{1}{(j+\tau+\xi)\{j+\tau+(\xi+h)/(h+1)\}} - \frac{1}{(j+\tau+1/2)\{j+\tau+(\xi-1)/(h+1)+1/2\}} \\
&= \frac{(-\xi)(j+\tau)+(1/2)\{(\xi-1)/(h+1)+1/2\}-\xi(\xi+h)/(h+1)}{(j+\tau+\xi)\{j+\tau+(\xi+h)/(h+1)\}(j+\tau+1/2)\{j+\tau+(\xi-1)/(h+1)+1/2\}} < 0.
\end{aligned}$$

This completes the proof. □

5.5.2 Proofs

Proof of Proposition 5.2.1. Since the joint posterior density of $(\boldsymbol{\mu}, \eta)$ is proportional to

$$\eta^{n_1/2+p/2-1} \exp\Big(-\frac{\eta}{2}V\Big) \exp\Big(-\frac{\eta}{2}||\boldsymbol{X}-\boldsymbol{\mu}||^2\Big),$$

the marginal posterior of η is proportional to

$$\eta^{n_1/2+p/2-1} \exp\Big(-\frac{\eta}{2}V\Big) \int_{\mathbb{R}^p} \exp\Big(-\frac{\eta}{2}||\boldsymbol{X}-\boldsymbol{\mu}||^2\Big) d\boldsymbol{\mu} = (2\pi)^{p/2} \eta^{n_1/2-1} \exp\Big(-\frac{\eta}{2}V\Big).$$

Therefore, the posterior mean of $p_2(w|\eta)$ is

$$\begin{aligned}\hat{p}_2^{(\pi_0)}(w|\boldsymbol{X},V) &= \frac{(1/2)^{n_2/2}}{\Gamma(n_2/2)}w^{n_2/2-1}\frac{\int_0^\infty \eta^{(n_1+n_2)/2-1}e^{-\eta(V+w)/2}d\eta}{\int_0^\infty \eta^{n_1/2-1}e^{-\eta V/2}d\eta}\\ &= \frac{(1/2)^{n_2/2}}{\Gamma(n_2/2)}w^{n_2/2-1}\frac{\Gamma((n_1+n_2)/2)/\{(V+w)/2\}^{(n_1+n_2)/2}}{\Gamma(n_1/2)/(V/2)^{n_1/2}},\end{aligned}$$

which is the desired result. □

Proof of Proposition 5.2.2. Let $\pi_{b,a}(\gamma) = (1-\gamma)^{b-1}\gamma^{-a-1}$ for $\gamma \in (0,1)$. Then the joint posterior density of $(\boldsymbol{\mu},\eta)$ is proportional to

$$\int_0^1 \pi_{b,a}(\gamma)\Big(\frac{\gamma}{1-\gamma}\Big)^{p/2}\eta^{n_1/2+p-a-1}\exp\Big(-\frac{\eta}{2}V\Big)\exp\Big\{-\frac{\eta}{2}\Big(\frac{\gamma}{1-\gamma}||\boldsymbol{\mu}||^2+||\boldsymbol{X}-\boldsymbol{\mu}||^2\Big)\Big\}d\gamma.$$

Note that

$$\frac{\gamma}{1-\gamma}||\boldsymbol{\mu}||^2+||\boldsymbol{X}-\boldsymbol{\mu}||^2 = \frac{||\boldsymbol{\mu}-(1-\gamma)\boldsymbol{X}||^2}{1-\gamma}+\gamma||\boldsymbol{X}||^2.$$

Then the marginal posterior of η is proportional to

$$\begin{aligned}&\int_0^1 \pi_{b,a}(\gamma)\Big(\frac{\gamma}{1-\gamma}\Big)^{p/2}\eta^{n_1/2+p-a-1}\exp\Big(-\frac{\eta}{2}V\Big)\Big(\int_{\mathbb{R}^p}\exp\Big[-\frac{\eta}{2}\Big\{\frac{||\boldsymbol{\mu}-(1-\gamma)\boldsymbol{X}||^2}{1-\gamma}+\gamma||\boldsymbol{X}||^2\Big\}\Big]d\boldsymbol{\mu}\Big)d\gamma\\ &= (2\pi)^{p/2}\int_0^1 \pi_{b,a}(\gamma)\gamma^{p/2}\eta^{n_1/2+p/2-a-1}\exp\Big\{-\frac{\eta}{2}(V+\gamma||\boldsymbol{X}||^2)\Big\}d\gamma.\end{aligned}$$

Therefore, the Bayesian predictive density $\hat{p}_2^{(\pi_{b,a})}(\cdot|\boldsymbol{X},V)$ is given by

$$\begin{aligned}\frac{\hat{p}_2^{(\pi_{b,a})}(w|\boldsymbol{X},V)}{\frac{(1/2)^{n_2/2}}{\Gamma(n_2/2)}w^{n_2/2-1}} &= \frac{\int_0^1 \pi_{b,a}(\gamma)\gamma^{p/2}\Big[\int_0^\infty \eta^{(n_1+n_2)/2+p/2-a-1}\exp\Big\{-\frac{\eta}{2}(V+w+\gamma||\boldsymbol{X}||^2)\Big\}d\eta\Big]d\gamma}{\int_0^1 \pi_{b,a}(\gamma)\gamma^{p/2}\Big[\int_0^\infty \eta^{n_1/2+p/2-a-1}\exp\Big\{-\frac{\eta}{2}(V+\gamma||\boldsymbol{X}||^2)\Big\}d\eta\Big]d\gamma}\\ &= \frac{\int_0^1 \pi_{b,a}(\gamma)\gamma^{p/2}\frac{\Gamma((n_1+n_2)/2+p/2-a)}{\{(1/2)(V+w+\gamma||\boldsymbol{X}||^2)\}^{(n_1+n_2)/2+p/2-a}}d\gamma}{\int_0^1 \pi_{b,a}(\gamma)\gamma^{p/2}\frac{\Gamma(n_1/2+p/2-a)}{\{(1/2)(V+\gamma||\boldsymbol{X}||^2)\}^{n_1/2+p/2-a}}d\gamma},\end{aligned}$$

from which the desired result follows. □

Proof of Theorem 5.3.1. Let $\Delta = R((\boldsymbol{\mu},\eta),\hat{p}_2^{(\pi_{n_1/2,a})}) - R((\boldsymbol{\mu},\eta),\hat{p}_2^{(\pi_0)})$. By Propositions 5.2.1 and 5.2.2 and by (5.2.1) and (5.2.2), we have

$$\begin{aligned}\Delta &= E_{(\boldsymbol{\mu},\eta)}^{(\boldsymbol{X},V,W)}\Big[\log\frac{\hat{p}_2^{(\pi_0)}(W;\boldsymbol{X},V)}{\hat{p}_2^{(\pi_{n_1/2,a})}(W;\boldsymbol{X},V)}\Big]\\ &= E_{(\boldsymbol{\mu},\eta)}^{(\boldsymbol{X},V,W)}\Big[\log B\Big(\frac{n_1+n_2}{2},\frac{p}{2}-a\Big)+\Big(\frac{p}{2}-a\Big)\log\frac{V+W+||\boldsymbol{X}||^2}{V+||\boldsymbol{X}||^2}\\ &\quad -\log\int_0^1(1-\gamma)^{n_1/2-1}\gamma^{p/2-a-1}\Big(1-\frac{||\boldsymbol{X}||^2}{V+W+||\boldsymbol{X}||^2}\gamma\Big)^{n_2/2}d\gamma\Big].\end{aligned}$$

It follows from Lemma 5.5.2 that for all $\boldsymbol{x} \in \mathbb{R}^p$, $v \in (0,\infty)$, and $w \in (0,\infty)$,

$$\begin{aligned}&\int_0^1 (1-\gamma)^{n_1/2-1}\gamma^{p/2-a-1}\Big(1-\frac{||\boldsymbol{x}||^2}{v+w+||\boldsymbol{x}||^2}\gamma\Big)^{n_2/2}d\gamma\\&\geq B\Big(\frac{n_1+n_2}{2},\frac{p}{2}-a\Big)\Big(1+c_1\frac{v+w}{v+w+||\boldsymbol{x}||^2}\Big)^{c_2}.\end{aligned}$$

Therefore,

$$\begin{aligned}\Delta &\leq E^{(\boldsymbol{X},V,W)}_{(\boldsymbol{\mu},\eta)}\Big[\Big(\frac{p}{2}-a\Big)\log\frac{V+W+||\boldsymbol{X}||^2}{V+||\boldsymbol{X}||^2}-c_2\log\Big(1+c_1\frac{V+W}{V+W+||\boldsymbol{X}||^2}\Big)\Big]\\&= E^{(\boldsymbol{X},V,W)}_{(\boldsymbol{\mu},\eta)}\Big[\Big(\frac{p}{2}-a\Big)\log\frac{\eta V+\eta W+||\sqrt{\eta}\boldsymbol{X}||^2}{\eta V+||\sqrt{\eta}\boldsymbol{X}||^2}-c_2\log\Big(1+c_1\frac{\eta V+\eta W}{\eta V+\eta W+||\sqrt{\eta}\boldsymbol{X}||^2}\Big)\Big].\end{aligned}$$

Let $k = n_1/2$, $l = n_2/2$, $m = p/2$, and $m' = m-a = p/2-a$. Let $Z \sim \mathrm{Po}(\theta/2)$ for $\theta = \eta||\boldsymbol{\mu}||^2$ and let $\tilde{V}$, $\tilde{W}$, and $\tilde{T}$ be independently distributed as $\chi^2(n_1)$, $\chi^2(n_2)$, and $\chi^2(p+2Z)$, respectively. Then since $(\eta V, \eta W, ||\sqrt{\eta}\boldsymbol{X}||^2) \overset{\mathrm{d}}{=} (\tilde{V},\tilde{W},\tilde{T})$ and since the expectation of the logarithm of a Chi-squared variable with $\nu > 0$ degrees of freedom is $\log 2 + \psi(\nu/2)$, it follows that

$$\begin{aligned}\Delta &\leq E^Z_\theta\Big[E^{(\tilde{T},\tilde{V},\tilde{W})|Z}_\theta\Big[m'\log\frac{\tilde{V}+\tilde{W}+\tilde{T}}{\tilde{V}+\tilde{T}}-c_2\log\Big(1+c_1\frac{\tilde{V}+\tilde{W}}{\tilde{V}+\tilde{W}+\tilde{T}}\Big)\Big|Z\Big]\Big]\\&= E^Z_\theta[D_1(Z)+D_2(Z)],\end{aligned}\tag{5.5.1}$$

where

$$D_1(z) = m'\{\psi(k+l+m+z)-\psi(k+m+z)\},\quad z\in\mathbb{N}_0,$$

and

$$D_2(z) = E^{\rho_Z|Z}_\theta[-c_2\log(1+c_1\rho_Z)|Z=z],\quad z\in\mathbb{N}_0,$$

for a random variable ρ_Z such that $\rho_Z|Z \sim \mathrm{Beta}(k+l, m+Z)$. By Lemma 5.5.3,

$$\begin{aligned}D_2(z) &= -c_2\int_0^1\{\log(1+c_1\rho)\}\frac{\rho^{k+l-1}(1-\rho)^{m+z-1}}{B(k+l,m+z)}d\rho\\&= -c_2\int_0^1\frac{(1-\rho)^{k+l+m+z-1}}{\rho}\Big\{1-\frac{1}{(1+c_1\rho)^{k+l}}\Big\}d\rho\end{aligned}$$

for all $z \in \mathbb{N}_0$. Therefore, by Lemma 5.5.4, $\lim_{z\to\infty}\{D_1(z)+D_2(z)\} = 0$. Fix $z \in \mathbb{N}_0$. Then

$$\begin{aligned}&\{D_1(z+1)+D_2(z+1)\}-\{D_1(z)+D_2(z)\}\\&= m'\Big(\frac{1}{k+l+m+z}-\frac{1}{k+m+z}\Big)-c_2\int_0^1\frac{(1-\rho)^{k+l+m+z-1}(-\rho)}{\rho}\Big\{1-\frac{1}{(1+c_1\rho)^{k+l}}\Big\}d\rho\\&= -\frac{lm'}{(k+m+z)(k+l+m+z)}+c_2\int_0^1(1-\rho)^{k+l+m+z-1}\Big\{1-\frac{1}{(1+c_1\rho)^{k+l}}\Big\}d\rho.\end{aligned}$$

Therefore,

$$D_1(z+1)+D_2(z+1) \gtreqless D_1(z)+D_2(z)\quad\text{if and only if}\quad f(k+m+z)\gtreqless lm'/c_2$$

for the function f defined by

$$f(\zeta) = \zeta(\zeta + l) \int_0^1 (1-\rho)^{\zeta+l-1} \Big\{ 1 - \frac{1}{(1+c_1\rho)^{k+l}} \Big\} d\rho, \quad \zeta \in (0, \infty).$$

Furthermore, by integration by parts

$$\begin{aligned} f(\zeta) &= \zeta \int_0^1 (1-\rho)^{\zeta+l} \frac{\partial}{\partial \rho} \Big\{ 1 - \frac{1}{(1+c_1\rho)^{k+l}} \Big\} d\rho = \int_0^1 \zeta(1-\rho)^{\zeta-1}(1-\rho)^{l+1} \frac{(k+l)c_1}{(1+c_1\rho)^{k+l+1}} d\rho \\ &= \Big[-(1-\rho)^{\zeta}(1-\rho)^{l+1} \frac{(k+l)c_1}{(1+c_1\rho)^{k+l+1}} \Big]_0^1 + \int_0^1 (1-\rho)^{\zeta} \frac{\partial}{\partial \rho} \Big\{ (1-\rho)^{l+1} \frac{(k+l)c_1}{(1+c_1\rho)^{k+l+1}} \Big\} d\rho \end{aligned}$$

for all $\zeta \in (0, \infty)$ and thus f is an increasing function. Finally, $D_1(0) + D_2(0) \leq 0$ by assumption. Hence, we conclude that $D_1(z) + D_2(z) \leq 0$ for all $z \in \mathbb{N}_0$ with strict inequality for some $z \in \mathbb{N}_0$. This completes the proof. □

Proof of Corollary 5.3.1. Let $k = n_1/2$, $l = n_2/2$, $m = p/2$, $m' = p/2 - a$. We show that

$$\frac{m'}{c_2} \{ \psi(k+l+m) - \psi(k+m) \} \leq \int_0^1 (1-\rho)^{k+l+m-1} g(\rho) d\rho, \tag{5.5.2}$$

where

$$(c_1, c_2) = \begin{cases} \Big(\dfrac{\Gamma(k)\Gamma(k+l+m')}{\Gamma(k+l)\Gamma(k+m')} - 1, 1 \Big), & \text{if } l \leq 1, \\ \Big(\dfrac{m'}{k+l-1}, l \Big), & \text{if } l > 1, \end{cases}$$

and where $g \colon (0,1) \to [0, \infty)$ is the function defined by

$$g(\rho) = \frac{1}{\rho} \Big\{ 1 - \frac{1}{(1+c_1\rho)^{k+l}} \Big\} = \frac{1}{\rho} \int_0^1 \Big[\frac{\partial}{\partial t} \Big\{ \frac{-1}{(1+c_1\rho t)^{k+l}} \Big\} \Big] dt = \int_0^1 \frac{(k+l)c_1}{(1+c_1\rho t)^{k+l+1}} dt, \quad \rho \in (0,1). \tag{5.5.3}$$

For part (i), since for all $\rho \in (0,1)$

$$g'(\rho) = \int_0^1 \frac{-(k+l+1)(k+l){c_1}^2 t}{(1+c_1\rho t)^{k+l+2}} dt,$$

g is a convex function. Therefore, by Jensen's inequality,

$$\begin{aligned} \int_0^1 (1-\rho)^{k+l+m-1} g(\rho) d\rho &= B(1, k+l+m) \int_0^1 \frac{\rho^{1-1}(1-\rho)^{k+l+m-1}}{B(1, k+l+m)} g(\rho) d\rho \\ &\geq B(1, k+l+m) g\Big(\frac{1}{k+l+m+1} \Big) \\ &= \frac{1}{k+l+m} \int_0^1 \frac{(k+l)c_1}{[1 + \{c_1/(k+l+m+1)\}t]^{k+l+1}} dt \\ &= \frac{k+l+m+1}{k+l+m} \Big[1 - \frac{1}{\{1 + c_1/(k+l+m+1)\}^{k+l}} \Big], \end{aligned}$$

the right-hand side of which is greater than or equal to the left-hand side of (5.5.2) by assumption.

To prove part (ii), note that

$$\begin{aligned}\lim_{m'\to 0}\frac{g(\rho)}{m'} &= \lim_{m'\to 0}\frac{c_1}{m'}\int_0^1 \frac{k+l}{(1+c_1\rho t)^{k+l+1}}dt = (k+l)\lim_{m'\to 0}\frac{c_1}{m'}\\ &= (k+l)\begin{cases}\dfrac{\partial}{\partial m'}\Big|_{m'=0}\dfrac{\Gamma(k)\Gamma(k+l+m')}{\Gamma(k+l)\Gamma(k+m')}, & \text{if } l\le 1,\\ \dfrac{1}{k+l-1}, & \text{if } l>1,\end{cases}\\ &= (k+l)\begin{cases}\psi(k+l)-\psi(k), & \text{if } l\le 1,\\ \dfrac{1}{k+l-1}, & \text{if } l>1,\end{cases}\end{aligned}$$

Then

$$\begin{aligned}&\lim_{m'\to 0}\frac{c_2}{m'}\int_0^1 (1-\rho)^{k+l+m-1}g(\rho)d\rho\\ &= c_2\Big\{\int_0^1 (1-\rho)^{k+l+m-1}d\rho\Big\}(k+l)\begin{cases}\psi(k+l)-\psi(k), & \text{if } l\le 1,\\ \dfrac{1}{k+l-1}, & \text{if } l>1,\end{cases}\\ &= \frac{(k+l)c_2}{k+l+m}\begin{cases}\psi(k+l)-\psi(k), & \text{if } l\le 1,\\ \dfrac{1}{k+l-1}, & \text{if } l>1,\end{cases}\end{aligned}$$

from which the desired result follows. □

Proof of Corollary 5.3.2. Let Δ and $D_1(z), D_2(z)$, $z\in\mathbb{N}_0$, be defined as in the proof of Theorem 5.3.1. For part (i), note that equality holds in (5.5.1) when $n_2=2$. Then if $\hat{p}_2^{(\pi_{n_1/2,a})}$ dominates $\hat{p}_2^{(\pi_0)}$, we have $\Delta|_{\boldsymbol{\mu}=\mathbf{0}_p}\le 0$, which implies $D_1(0)+D_2(0)\le 0$. This proves the "only if" part. The "if" part follows from Theorem 5.3.1. For part (ii), note that by (5.5.3), the right-hand side of (5.3.2) divided by $p/2-a$ is

$$\int_0^1 (1-\rho)^{n_1/2+p/2}\Big(\int_0^1 \frac{(n_1+2)/n_1}{[1+\{(p-2a)/n_1\}\rho t]^{n_1/2+2}}dt\Big)d\rho.$$

Since the above integral is increasing in a, we need only show that equality holds in (5.3.2) when $a=0$. Suppose that $n_1=2$ and that $a=0$. Let $m=p/2$. Then, by integration by parts,

$$\begin{aligned}&\frac{1}{p/2-a}\int_0^1 (1-\rho)^{n_1/2+p/2}\frac{1}{\rho}\Big(1-\frac{1}{[1+\{(p/2-a)/(n_1/2)\}\rho]^{n_1/2+1}}\Big)d\rho\Big\}\\ &= \int_0^1 \frac{(1-\rho)^{m+1}}{m\rho}\Big\{1-\frac{1}{(1+m\rho)^2}\Big\}d\rho = \int_0^1 (1-\rho)^{m+1}\Big\{\frac{1}{1+m\rho}+\frac{1}{(1+m\rho)^2}\Big\}d\rho\\ &= \int_0^1 \frac{(1-\rho)^{m+1}}{1+m\rho}d\rho+\frac{1}{m}-\frac{m+1}{m}\int_0^1\frac{(1-\rho)^m}{1+m\rho}d\rho = \frac{1}{m}-\frac{1}{m}\int_0^1(1-\rho)^m d\rho = \frac{1}{m+1},\end{aligned}$$

which equals $1/(n_1/2+p/2)$. Thus, we have proved the desired result. □

Proof of Theorem 5.3.2. By (5.2.3), we have

$$R((\boldsymbol{\mu},\eta),\hat{p}_2^{(\pi_{1,a})}) - R((\boldsymbol{\mu},\eta),\hat{p}_2^{(\pi_0)}) = E_{(\boldsymbol{\mu},\eta)}^{(\boldsymbol{X},V,W)}\Big[\log\frac{\hat{p}_2^{(\pi_0)}(W;\boldsymbol{X},V)}{\hat{p}_2^{(\pi_{1,a})}(W;\boldsymbol{X},V)}\Big] = \Delta(n_1+n_2) - \Delta(n_1),$$

where, for each $n \in \{n_1, n_1+n_2\}$,

$$\Delta(n) = E_{(\boldsymbol{\mu},\eta)}^{(\boldsymbol{X},U_n)}\Big[-\log\int_0^{||\boldsymbol{X}||^2/(U_n+||\boldsymbol{X}||^2)}\frac{\gamma^{p/2-a-1}(1-\gamma)^{n/2-1}}{B(p/2-a,n/2)}d\gamma\Big]$$

for the random variable U_n which is V if $n = n_1$ and $V+W$ if $n = n_1+n_2$. Let θ, Z, and $\tilde{T}$ be defined as in the proof of Theorem 5.3.1. Then for $n \in \{n_1, n_1+n_2\}$, $\Delta(n)$ can be written as

$$\Delta(n) = E_\theta^Z[D(n;Z)],$$

where

$$D(n;z) = E_\theta^{(\tilde{T},\tilde{U}_n)|Z}\Big[-\log\int_0^{\tilde{T}/(\tilde{U}_n+\tilde{T})}\frac{\gamma^{p/2-a-1}(1-\gamma)^{n/2-1}}{B(p/2-a,n/2)}d\gamma\Big|Z=z\Big], \quad z \in \mathbb{N}_0,$$

for an independent variable $\tilde{U}_n \sim \chi^2(n)$.

Fix $z \in \mathbb{N}_0$. Then for each $n \in \{n_1, n_1+n_2\}$, since $\{\tilde{T}/(\tilde{U}_n+\tilde{T})\}|(Z=z) \sim \mathrm{Beta}(p/2+z, n/2)$, it follows that

$$\begin{aligned} D(n;z) &= -\int_0^1\Big\{\log\int_0^q\frac{\gamma^{p/2-a-1}(1-\gamma)^{n/2-1}}{B(p/2-a,n/2)}d\gamma\Big\}\frac{q^{p/2+z-1}(1-q)^{n/2-1}}{B(p/2+z,n/2)}dq \\ &= -\int_0^1(\log\omega)\frac{B(p/2-a,n/2)}{B(p/2+z,n/2)}\{{F_n}^{-1}(\omega)\}^{z+a}d\omega, \end{aligned}$$

where

$$F_n(q) = \int_0^q\frac{\gamma^{p/2-a-1}(1-\gamma)^{n/2-1}}{B(p/2-a,n/2)}d\gamma$$

for $q \in (0,1)$. Therefore, $D(n_1+n_2;z) \lesseqgtr D(n_1;z)$ if and only if

$$\int_0^1(\log\omega)\Big[1-C(z)\Big\{\frac{{F_{n_1}}^{-1}(\omega)}{{F_{n_1+n_2}}^{-1}(\omega)}\Big\}^{z+a}\Big]dP_z(\omega) \gtreqless 0, \tag{5.5.4}$$

where

$$C(z) = \frac{B(p/2-a,n_1/2)}{B(p/2+z,n_1/2)}\Big/\frac{B(p/2-a,(n_1+n_2)/2)}{B(p/2+z,(n_1+n_2)/2)}$$

and where P_z is the probability measure with density

$$\frac{B(p/2-a,(n_1+n_2)/2)}{B(p/2+z,(n_1+n_2)/2)}\{{F_{n_1+n_2}}^{-1}(\omega)\}^{z+a}, \quad \omega \in (0,1).$$

Since $a < p/2$ and $n_1 > 2$ by assumption, it follows from Lemma 5.5.5 that ${F_{n_1+n_2}}^{-1}(\omega)/{F_{n_1}}^{-1}(\omega)$ is nondecreasing in $\omega \in (0,1)$ and strictly increasing in $\omega \in (\underline{\omega}, \overline{\omega})$ for some $0 < \underline{\omega} < \overline{\omega} < 1$. Thus, since

$$\int_0^1 \Big[1 - C(z)\Big\{\frac{{F_{n_1}}^{-1}(\omega)}{{F_{n_1+n_2}}^{-1}(\omega)}\Big\}^{z+a}\Big] dP(\omega) = 0,$$

the left-hand side of (5.5.4) is, by the covariance inequality, greater than zero if $z + a > 0$ and equal to zero if $z + a = 0$, from which the desired result follows. □

Proof of Theorem 5.3.3. Let θ and Z be defined as in the proof of Theorem 5.3.1. Then, by the proof of Theorem 5.3.2,

$$\begin{aligned}
&R((\boldsymbol{\mu}, \eta), \hat{p}_2^{(\pi_{1,p/2-1})}) - R((\boldsymbol{\mu}, \eta), \hat{p}_2^{(\pi_0)}) \\
&= E_\theta^Z\Big[\Big[-\int_0^1 \Big\{\log \int_0^q \frac{(1-\gamma)^{n/2-1}}{B(1, n/2)} d\gamma\Big\} \frac{q^{p/2+Z-1}(1-q)^{n/2-1}}{B(p/2+Z, n/2)} dq\Big]_{n=n_1}^{n=n_1+n_2}\Big] \\
&= E_\theta^Z\Big[\Big[-\int_0^1 [\log\{1-(1-q)^{n/2}\}] \frac{q^{p/2+Z-1}(1-q)^{n/2-1}}{B(p/2+Z, n/2)} dq\Big]_{n=n_1}^{n=n_1+n_2}\Big] \\
&= \sum_{h=1}^{\infty} \frac{1}{h} E_\theta^Z\Big[\Big[\int_0^1 \frac{q^{p/2+Z-1}(1-q)^{(h+1)(n/2)-1}}{B(p/2+Z, n/2)} dq\Big]_{n=n_1}^{n=n_1+n_2}\Big].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&R((\boldsymbol{\mu}, \eta), \hat{p}_2^{(\pi_{1,p/2-1})}) - R((\boldsymbol{\mu}, \eta), \hat{p}_2^{(\pi_0)}) \\
&= \sum_{h=1}^{\infty} \frac{1}{h} E_\theta^Z\Big[\Big[\frac{B(p/2+Z, (h+1)(n/2))}{B(p/2+Z, n/2)}\Big]_{n=n_1}^{n=n_1+n_2}\Big] \\
&= \sum_{h=1}^{\infty} \frac{1}{h} E_\theta^Z\Big[\Big[\frac{\Gamma(p/2+Z+n/2)\Gamma((h+1)(n/2))}{\Gamma(p/2+Z+(h+1)(n/2))\Gamma(n/2)}\Big]_{n=n_1}^{n=n_1+n_2}\Big] \\
&= \sum_{h=1}^{\infty} \frac{1}{h} E_\theta^Z\Big[\int_{n_1/2}^{(n_1+n_2)/2} \Big\{\frac{\partial}{\partial\tau} \frac{\Gamma(p/2+Z+\tau)\Gamma((h+1)\tau)}{\Gamma(p/2+Z+(h+1)\tau)\Gamma(\tau)}\Big\} d\tau\Big].
\end{aligned}$$

Thus, by Lemma 5.5.6, we have $R((\boldsymbol{\mu}, \eta), \hat{p}_2^{(\pi_{1,p/2-1})}) \le R((\boldsymbol{\mu}, \eta), \hat{p}_2^{(\pi_0)})$. Equality holds if and only if $p = 2$ and $\boldsymbol{\mu} = \mathbf{0}_p$. This completes the proof. □

Part III

Fully Bayesian Posterior Inference

Chapter 6

On Global-Local Shrinkage Priors for Count Data

6.1 Introduction

High-dimensional count data appears in a variety of scientific fields including genetics, epidemiology and social science. It is frequently observed in such data that many of those counts are very small and nearly zero except for some outliers. For example, in crime statistics where we divide the area of interest into small sub-regions, the number of occurrences of specific crime is likely to be small or zero in many sub-regions, while it is still important to detect "hotspots", i.e., the regions of the unexplained high crime rate. In this context, the Poisson-gamma model is obviously inappropriate, for the gamma prior shrinks all the observations uniformly, including the large signals that should be kept unshrunk, which might result in overlooking such meaningful regions. The desirable prior should account for both small and large signals and realize the flexible shrinkage effects on Poisson rates.

The prior of this type has been studied as global-local shrinkage prior for the Gaussian observations. The sparse signals of high-dimensional continuous observations are detected by the horseshoe prior, which exhibits the aforementioned property of shrinkage, being comparable to the variable selection (Carvalho et al. (2010)). It is extended to the three-parameter beta distribution for more flexible modeling of sparseness (Armagan et al. (2011)). In hierarchical models, such priors have been adopted for random effect distributions in small area estimation (Tang et al. (2018)) or default Bayesian analysis (Bhadra et al. (2016)). For recent developments, see, for example, Bhadra et al. (2019) and the references therein.

While extensively studied for Gaussian data, the global-local shrinkage priors have not been fully developed for count data, although Poisson likelihood models with hierarchical structure are widely used in applications such as disease mapping (see, for example, Wakefield (2006) and Lawson (2013)). The theory related to the Poisson likelihoods has been well developed (e.g., Brown et al. (2013) and Yano et al. (2019)), but not necessarily from the viewpoint of global-local shrinkage. The standard Bayesian models for count data is of Poisson-gamma type; the gamma prior for the Poisson rate shows the similarity to the global-local shrinkage prior if one assumes further hierarchical prior on the gamma scale parameters. In this context, the use of heavy-tailed hierarchical priors has already been practiced (e.g., Zhu et al. (2019)), but the research on the general, statistical property of such priors has been limited. The theoretical

properties of the Bayes estimators of those models have been investigated partially by Datta and Dunson (2016) with the focus on (a generalized version of) the three-parameter beta prior for the analysis of zero-inflated count data. Our research is also concerned with the global-local shrinkage for count data, but especially from the rigorous viewpoint of heavy-tail property, which ensures the large signals are less or not at all affected by the shrinkage effect.

The objective of our research is to consider the effect of the hyperprior on the Bayes estimators (posterior means) of Poisson rates in terms of the robustness property. In doing so, we first define the concept of tail-robustness for the Bayes estimators mathematically. A robust Bayes estimator should keep large signals unshrunk, while retaining the strong shrink effect on small signals towards prior means, which is the *tail-robustness* we assess by our main theorem. In Section 6.2, Theorem 6.2.1 and Corollary 6.2.1 give sufficient conditions for the tail-robustness.

Requiring the tail-robustness for the Bayes estimators helps us restrict the class of priors we should use. The conditions in Theorem 6.2.1 reveal the importance of local shrinkage, or the individual scale parameter of gamma distribution customized for each Poisson rate, and support the use of two classes of hyperpriors proposed in Section 6.3: the inverse-gamma prior and the newly-introduced extremely heavy-tailed prior. The inverse-gamma prior is a well-known distribution and easy to be integrated into the model. The asymptotic bias for large signals is shown to be negligible, hence the inverse-gamma prior is "approximately" tail-robust. The extremely heavily-tailed prior is a new class of probability distributions, whose density function is derived so as to satisfy the conditions for tail-robustness. In contrast to the inverse-gamma prior, this prior is exactly tail-robust. Both priors are conditionally conjugate for most of parameters in the model, which allows the fast and efficient posterior analysis by Gibbs sampler.

In the numerical study, we observe the properties of tail-robustness theoretically guaranteed for those priors, while the standard Poisson-gamma model suffers from the overly-shrunk Bayes estimators for outliers. The difference of two proposed priors are empirically confirmed in this numerical study; the inverse-gamma prior is better in the point estimations for small signals, having more shrinkage effect toward prior mean, while the extremely heavy-tailed prior is successful in quantifying the uncertainty for large counts, as shown in the coverage rates of posterior credible intervals. Despite this difference, both priors perform almost equally in the analysis of the actual crime data in Japan by detecting the hotspots of crimes that are overlooked in the Poisson-gamma models.

The rest of this chapter is organized as follows. In Section 6.2, we consider theoretical argument regarding tail robustness and derive sufficient conditions for local priors to hold tail robustness. In Section 6.3, we propose two local priors and provide efficient posterior computation algorithms using Gibbs sampling. We also discuss some properties of the implied marginal priors and posteriors of Poisson rate. Section 6.4 is devoted to the numerical experiments for the extensive comparison of the proposed priors and other commonly-used priors/estimators under the various settings. The application to the real data of crimes in Tokyo metropolitan area, Japan, is discussed in Section 6.5. The step-by-step sampling algorithm, technical details regarding the proofs of Theorem 6.2.1 and Proposition 6.7.1, motivation of the EH prior, and other computational issues in the main text are given in the Appendix. Finally, R code implementing the proposed method is available at GitHub repository (`https://github.com/sshonosuke/GLSP-count`).

6.2 Tail-Robustness Under Count Response

6.2.1 Hierarchical models for count data

Our model has the following hierarchical representation; the m observations $y_1, \ldots, y_m$ are conditionally independent and modelled by, for $i = 1, \ldots, m$,

$$y_i|\lambda_i \sim \mathrm{Po}(\eta_i\lambda_i), \quad \lambda_i|u_i \sim \mathrm{Ga}(\alpha, \beta/u_i), \quad u_i \sim \pi(u_i), \tag{6.2.1}$$

where $\mathrm{Po}(\eta_i\lambda_i)$ is the Poisson distribution with rate $\eta_i\lambda_i$, and $\mathrm{Ga}(\alpha, \beta/u_i)$ the gamma distribution with shape α and rate β/u_i, whose (conditional) mean is $\alpha/(\beta/u_i)$. In addition, $\eta_i \in (0, \infty)$ is offset and known, $(\alpha, \beta) \in (0, \infty)^2$ are the hyperparameters, and $u_i \in (0, \infty)$ is a local scale parameter. The offset term, η_i, can be any known constant in general; in practice, it is flexibly modeled by regression with the log link function, as we examine in Section 6.5. In what follows, we assume $\eta_i = 1$ for simplicity. The two rate parameters of the gamma prior, β and u_i^{-1}, control the global and local shrinkage effects, respectively. Under this model, the Bayes estimator of Poisson rate λ_i we consider is the posterior mean

$$\begin{aligned} \tilde{\lambda}_i &= E\Big[\frac{u_i}{\beta + u_i}(\alpha + y_i)\Big|y_i\Big], \\ &= y_i - E\Big[\frac{\beta}{\beta + u_i}\Big(y_i - \frac{\alpha u_i}{\beta}\Big)\Big|y_i\Big] \end{aligned} \tag{6.2.2}$$

where the expectation is taken with respect to the marginal posterior of u_i, so that the conditional posterior mean of λ_i shrinks y_i toward the prior mean $\alpha u_i/\beta$. Throughout this chapter, we consider proper priors for u_i only. The use of improper priors for u_i results in the improper marginal of λ_i, and the posterior distribution of λ_i would not successfully reflect the prior information, failing to shrink the Bayes estimator satisfactory.

6.2.2 Tail-robustness of the posterior mean

The appropriate choice of prior $\pi(u_i)$ is discussed in terms of the shrinkage effect realized in the Bayes estimator $\tilde{\lambda}_i$. As stated in the introduction, the estimator should not be shrunk toward prior mean when the large signal is observed. This property is named as the tail-robustness (e.g. Carvalho et al. (2010)). The tail-robustness is mathematically defined as the property that

$$\lim_{y_i \to \infty} |\tilde{\lambda}_i - y_i| = 0. \tag{6.2.3}$$

This means that the (mean) absolute error loss tends to zero as $y_i \to \infty$. For fixed u_i, the Bayes estimator $(\alpha + y_i)/(1 + \beta/u_i)$ clearly loses the tail-robustness, which motivates the study of hierarchical prior for u_i. Throughout this chapter, our primal interest is in this property defined in (6.2.3), but we note that there have been other definitions of tail-robustness related to various loss functions. We discuss in details the difference of tail-robustness concepts in the Appendix.

To consider the tail-robustness, the next theorem is useful in evaluating the asymptotic bias $\lim_{y_i \to \infty}(\tilde{\lambda}_i - y_i)$ for a variety of priors.

Theorem 6.2.1 *Assume that $\pi(\cdot)$ is strictly positive and continuously differentiable. Suppose that $\pi(\cdot)$ satisfies the following two conditions:*

$$\int_0^1 |u\pi'(u)|du < \infty, \tag{A1}$$

$$\xi \equiv \lim_{u\to\infty} \frac{u\pi'(u)}{\pi(u)} \quad \textit{exists in } [-\infty, \infty]. \tag{A2}$$

Then the asymptotic bias of $\tilde{\lambda}_i$ is $1+\xi$, that is,

$$\lim_{y_i\to\infty} (\tilde{\lambda}_i - y_i) = 1 + \xi.$$

The asymptotic bias of $\tilde{\lambda}_i$ under $y_i \to \infty$ can be characterized by the tail behaviour of the mixing distribution $\pi(\cdot)$. This condition is similar to but significantly different from that of Gaussian response (e.g. Tang et al. (2018)). It is immediate from Theorem 6.2.1 that $\xi = -1$ is the sufficient condition for the estimator to be tail-robust, which is summarized in the following corollary.

Corollary 6.2.1 *Under the conditions (A1) and*

$$\lim_{u\to\infty} \frac{u\pi'(u)}{\pi(u)} = -1, \tag{A3}$$

the Bayes estimator $\tilde{\lambda}_i$ is tail-robust and satisfies $|\tilde{\lambda}_i - y_i| \to 0$ as $y_i \to \infty$.

The crucial assumption in the above corollary is (A3), which describes the desirable tail behavior of the marginal prior distribution of λ_i. In fact, (A3) is sufficient for $\psi(u) = u\pi(u)$ to be slowly varying as $u \to \infty$, i.e., $\lim_{u\to\infty} \psi(\kappa u)/\psi(u) = 1$ for all $\kappa > 0$ (e.g., see Seneta (1976), equation (1.11)). It further implies that, for the marginal prior $p(\lambda_i) = \int_0^\infty \mathrm{Ga}(\lambda_i|\alpha, \beta/u_i)\pi(u_i)du_i$, we have $\lambda_i p(\lambda_i) \sim \lambda_i \pi(\lambda_i)$ as $\lambda_i \to \infty$ under the regularity condition that justifies the interchange of the limit and integral. In other words, under this assumption, the marginal densities of λ_i and u_i are asymptotically equivalent in the tail as density functions.

An example of priors that satisfies assumption (A3) is $\pi(u) \propto 1/u$. In many cases, (A3) requires priors to be of this from; see Section 6.7.5. However, this prior is improper. In other words, $\pi(\cdot)$ have to be as heavy-tailed as improper priors for $\tilde{\lambda}_i$ to be tail-robust. On the other hand, (A1) is merely a technical requirement for the proof.

One notable feature of Corollary 6.2.1 is that the sufficient conditions for the tail-robustness, (A1) and (A3), are independent of the values of hyperparmeters α and β. This setting about hyperparameters is a great contrary to other approaches, e.g., Proposition 1 of Datta and Dunson (2016) where the tail-robustness is discussed for the limiting values of hyperparameters, i.e., $\beta \to \infty$ or $\beta \to 0$.

6.3 Global-Local Shrinkage Priors for Count Data

6.3.1 Proposed priors

Under the hierarchical model (6.2.1), we propose two families of priors for u_i. Each of them is indexed by a hyperparameter $\gamma \in (0, \infty)$, which will be estimated in practice.

The first prior is the inverse gamma (IG) prior given by

$$\pi_{\mathrm{IG}}(u_i;\gamma) = \frac{\gamma^\gamma}{\Gamma(\gamma)} \frac{1}{{u_i}^{1+\gamma}} e^{-\gamma/u_i}, \tag{6.3.1}$$

where $\gamma > 0$. This is the density of the $\mathrm{IG}(\gamma,\gamma)$ distribution. It is clearly proper and conditionally conjugate, which simplifies the posterior computation by Markov chain Monte Carlo methods. From Theorem 6.2.1, it holds that $\lim_{y_i\to\infty}(\tilde{\lambda}_i - y_i) = -\gamma$, indicating that the IG prior approximately satisfies the tail-robustness when γ is small. The shape and rate parameters are identical in (6.3.1), so that we have $E[1/u_i] = 1$, and the global parameter β can be interpreted as the marginal rate parameter of the gamma distribution for λ_i, i.e., the global shrinkage factor.

Next, we newly introduce a conjugate prior. The extremely heavy-tailed (EH) prior is defined by the density

$$\pi_{\mathrm{EH}}(u_i;\gamma) = \frac{\gamma}{1+u_i} \frac{1}{\{1+\log(1+u_i)\}^{1+\gamma}} \tag{6.3.2}$$

for $\gamma > 0$. The EH prior can be seen as a modification of the scaled-beta prior; the details on the connection to the EH prior is discussed in the Appendix. The additional logarithm function in (6.3.2) contributes to the integrability of the density function. The use of log-term is often seen in the literature of decision-theoretic statistical theory (for example, see Maruyama and Strawderman (2020a), Remark 4.1). This prior is proper because

$$\int_0^\infty \pi_{\mathrm{EH}}(u;\gamma)du = \Big[-\{1+\log(1+u)\}^{-\gamma} \Big]_0^\infty = 1.$$

The notable property of the EH prior is that it satisfies the condition of Corollary 6.2.1;

$$\frac{u\pi_{\mathrm{EH}}{}'(u;\gamma)}{\pi_{\mathrm{EH}}(u;\gamma)} = u\Big\{ -\frac{1}{1+u} - \frac{1+\gamma}{1+\log(1+u)} \frac{1}{1+u} \Big\} \to -1$$

as $u \to \infty$. Hence, the EH prior is exactly tail-robust.

The densities and tail-behaviors of the proposed priors are summarized in Table 6.1 together with those of the Gauss hypergeometric (GH) prior considered in Datta and Dunson (2016). The GH prior is dependent on the global rate parameter β, but its density tail (the asymptotic functional form of density as $u_i \to \infty$) is independent of β and identical to that of the half-Cauchy prior (Carvalho et al. (2010)). The density tail of the EH prior is heavier than those of the GH and IG priors regardless of γ. This difference originates from the log-term of the EH density and contributes to the exact tail-robustness of the EH prior.

	Density kernel of u_i	Density tail as $u_i \to \infty$
$\mathrm{GH}(1/2,1/2,\gamma,1/\beta)$	${u_i}^{-1/2}(1+u_i)^{-\gamma}(\beta+u_i)^{\gamma-1}$	${u_i}^{-3/2}$
$\mathrm{IG}(\gamma,\gamma)$	${u_i}^{-(\gamma+1)}e^{-\gamma/u_i}$	${u_i}^{-(\gamma+1)}$
$\mathrm{EH}(\gamma)$	$(1+u_i)^{-1}\{1+\log(1+u_i)\}^{-(1+\gamma)}$	${u_i}^{-1}(\log u_i)^{-(1+\gamma)}$

Table 6.1: Densities of GH, IG and EH priors

Finally, we note the parametrization by $\kappa = 1/(1+u) \in (0,1)$, which also clarifies the difference of the proposed priors from others. The implied density of the EH prior in the scale of

κ is $\pi_{\mathrm{EH}}(\kappa) = \gamma\kappa^{-1}/\{1+\log(1/\kappa)\}^{1+\gamma}$. This expression shows that the EH prior can be viewed as an extension of the improper beta prior, Beta$(0,1)$. The resulting EH prior is proper; the additional log-term in the density of the EH prior ensures the propriety. Other class of priors, including the half-Cauchy prior, remain in the class of beta distributions in κ-scale and do not involve the log-term in their densities.

6.3.2 Posterior computation

The computation of the Bayes estimator is based on the Markov chain Monte Carlo method. Because the proposed priors are mostly conditionally conjugate, sampling from most of the conditional posterior distributions is straightforward. In this subsection, we mention the essential strategies of the sampling methods. We provide the detailed step-by-step Gibbs sampling in the Appendix.

We first discuss the parameters $(\boldsymbol{\lambda}_{1:m}, \alpha, \beta)$, which are common to and always included in all the models regardless of the choice of prior for u_i. Note that we assign prior distributions for α and β in practice. In this research, we consider the gamma priors; $\alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha)$ and $\beta \sim \mathrm{Ga}(a_\beta, b_\beta)$. We adopt $a_\alpha = b_\alpha = a_\beta = b_\beta = 1$ as default choices, which will be used in the numerical studies in Sections 6.4 and 6.5. When the model is of Poisson-gamma type and the local parameters u_i are fixed, the posterior analysis can be done by sampling the above parameters. It is noted that the gamma prior for β is conditionally conjugate whereas the gamma prior for α is not. However, using the augmentation technique by Zhou and Carin (2013), we can derive an efficient Gibbs sampling method as provided in the Appendix.

For the model with the IG prior, the scale parameter u_i has a known conditional posterior, while the conditional posterior of the hyperparameter γ is difficult to directly sample from. Although several computationally-sophisticated options are available for the sampling of γ, we here simply use the random-walk Metropolis-Hastings method with uniform prior $\gamma \sim \mathrm{U}(\varepsilon_1, \varepsilon_2)$ for fixed small $\varepsilon_1 > 0$ and large $\varepsilon_2 > 0$. We set $\varepsilon_1 = 0.001$ and $\varepsilon_2 = 150$ as a default choice.

The new EH prior is not conditionally conjugate for u_i, despite its simple closed-form of the density function in (6.3.2). To develop an efficient sampling algorithm, we introduce a novel augmentation approach using two positive valued latent variables v_i and w_i, given by

$$\pi_{\mathrm{EH}}(u_i;\gamma) = \iint_{(0,\infty)^2} \pi_{\mathrm{EH}}(u_i, v_i, w_i;\gamma) dv_i dw_i,$$

where

$$\begin{aligned}\pi_{\mathrm{EH}}(u_i, v_i, w_i;\gamma) &= \mathrm{Ga}(u_i|1, v_i)\mathrm{Ga}(v_i|w_i, 1)\mathrm{Ga}(w_i|\gamma, 1)\\ &= \frac{w_i^{\gamma-1} v_i^{w_i}}{\Gamma(\gamma)\Gamma(w_i)} \exp\left\{-w_i - v_i(1+u_i)\right\}.\end{aligned}$$

Using the above expression, it is observed that the full conditional distribution of u_i is the generalized inverse Gaussian (GIG) distribution. We can also obtain familiar forms of the conditional posterior distributions of the other parameters, (v_i, w_i), where the details are given in the Appendix. For the shape parameter γ in the EH prior, we assign gamma prior $\gamma \sim \mathrm{Ga}(a_\gamma, b_\gamma)$ which is conditionally conjugate. We use $a_\gamma = b_\gamma = 1$ for simplicity as a default choice.

6.3.3 Marginal prior distributions for λ_i

In this section, the marginal density of λ_i is computed to consider its behavior in the limit of $\lambda_i \to \infty$ and $\lambda_i \to 0$. The tail property of the marginal density is the same as that of the prior of u_i. Information on the behavior of the marginal density of λ_i around zero is also important to understand the amount of shrinkage effect toward zero, which has not been discussed up to this point in this chapter. In general, the marginal prior distribution for λ_i is given by

$$\begin{aligned} p(\lambda_i;\alpha,\beta,\gamma) &= \int_0^\infty \frac{\beta^\alpha/{u_i}^\alpha}{\Gamma(\alpha)} {\lambda_i}^{\alpha-1} e^{-(\beta/u_i)\lambda_i} \pi(u_i;\gamma) du_i \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \frac{1}{x^\alpha} e^{-\beta/x} \pi(\lambda_i x;\gamma) dx. \end{aligned}$$

We continue the computation of this density for the two classes of priors: π_{IG} and π_{EH}.

For the IG prior $\pi(u_i;\gamma) = \pi_{\mathrm{IG}}(u_i;\gamma)$, we have

$$p(\lambda_i;\alpha,\beta,\gamma) = \frac{(\beta/\gamma)^\alpha}{B(\alpha,\gamma)} \frac{{\lambda_i}^{\alpha-1}}{\{1+(\beta/\gamma)\lambda_i\}^{\alpha+\gamma}}, \tag{6.3.3}$$

which implies the beta distribution, i.e.,

$$\frac{(\beta/\gamma)\lambda_i}{1+(\beta/\gamma)\lambda_i} \sim \mathrm{Beta}(\alpha,\gamma).$$

From (6.3.3), we have $p(\lambda_i;\alpha,\beta,\gamma) = O({\lambda_i}^{-1-\gamma})$ as $\lambda_i \to \infty$. For sufficiently small γ, the marginal prior of λ_i can be heavily-tailed, being almost equivalent to ${\lambda_i}^{-1}$ in the tail. This observation is coherent with the γ-dependent asymptotic bias of the Bayes estimator, $\lim_{y_i\to\infty}(\tilde{\lambda}_i - y_i) = -\gamma$. It should also be noted here that, due to the heavy tail of this density, the prior mean does not exist if $\gamma \le 1$. This can be confirmed by the direct computation or the fact that the prior mean of u_i is not finite under this condition. In this situation, it is difficult to interpret the prior from the viewpoint that the estimator is shrunk toward the prior mean. For those who prefer the prior with finite mean, we recommend the modification of the IG prior to IG$(\gamma+1,\gamma)$, $\gamma > 0$, which instead increases the asymptotic bias slightly to $-\gamma - 1$.

In contrast, the density at the origin depends on the value of α. In particular, $\lim_{\lambda_i\to 0} p(\lambda_i;\alpha,\beta,\gamma) = \infty$ for $\alpha < 1$, while the limit becomes a positive constant for $\alpha = 1$ and zero for $\alpha > 1$. This fact gives a clue to the interpretation of the choice of, or the posterior inference for, hyperparameter α.

For the EH prior, the marginal density is evaluated around zero as follows: For $\pi(u_i;\gamma) = \pi_{\mathrm{EH}}(u_i;\gamma)$,

$$\begin{aligned} p(\lambda_i;\alpha,\beta,\gamma) &= \frac{\beta^\alpha\gamma}{\Gamma(\alpha)} \int_0^\infty \frac{e^{-\beta/x}}{x^\alpha} \frac{1}{1+\lambda_i x} \frac{1}{\{1+\log(1+\lambda_i x)\}^{1+\gamma}} dx \\ &\to \frac{\beta^\alpha\gamma}{\Gamma(\alpha)} \int_0^\infty \frac{e^{-\beta/x}}{x^\alpha} dx \\ &= \begin{cases} (\alpha-1)^{-1}\beta\gamma & \text{if } \alpha > 1 \\ \infty & \text{if } \alpha \le 1 \end{cases} \end{aligned}$$

as $\lambda_i \to 0$ by the monotone convergence theorem. Thus, $\lim_{\lambda_i \to 0} p(\lambda_i; \alpha, \beta, \gamma) > 0$ and increasing as $\alpha \to 0$, implying the shrinkage of small signals toward the global prior mean. For the tail property, we have

$$\begin{aligned}\lim_{\lambda_i \to \infty} \frac{p(\lambda_i; \alpha, \beta, \gamma)}{\pi_{\text{EH}}(\lambda_i; \gamma)} &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \frac{e^{-\beta/x}}{x^\alpha} \Big[\lim_{\lambda_i \to \infty} \frac{1+\lambda_i}{1+\lambda_i x} \Big\{ \frac{1+\log(1+\lambda_i)}{1+\log(1+\lambda_i x)} \Big\}^{1+\gamma} \Big] dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \frac{e^{-\beta/x}}{x^{\alpha+1}} dx = 1.\end{aligned}$$

Therefore, $p(\lambda_i; \alpha, \beta, \gamma) \sim \pi_{\text{EH}}(\lambda_i; \gamma) \sim \gamma {\lambda_i}^{-1} (\log \lambda_i)^{-1-\gamma}$ as $\lambda_i \to \infty$, which means that the marginal prior $p(\lambda_i; \alpha, \beta, \gamma)$ is proper but has a sufficiently heavy tail so that the model can accommodate large signals. For the computation to verify the result above, see the Appendix.

The marginal distributions of λ_i with $\alpha = \beta = 2$ under the proposed IG and EH priors with $\gamma = 1$ and $\gamma = 0.5$ as well as the GH prior with $\gamma = 1$ are visually illustrated in Figure 6.1. It shows that the IG prior with $\gamma = 0.5$ has almost the same tail-behavior as the GH prior since the tail-behavior of the density of u_i under the IG prior with $\gamma = 1$ is equivalent that of GH as confirmed in Table 6.1. Moreover, the figure reveals that the density tail under the EH prior is heavier than those under the IG and GH priors, which is consistent with Table 6.1.

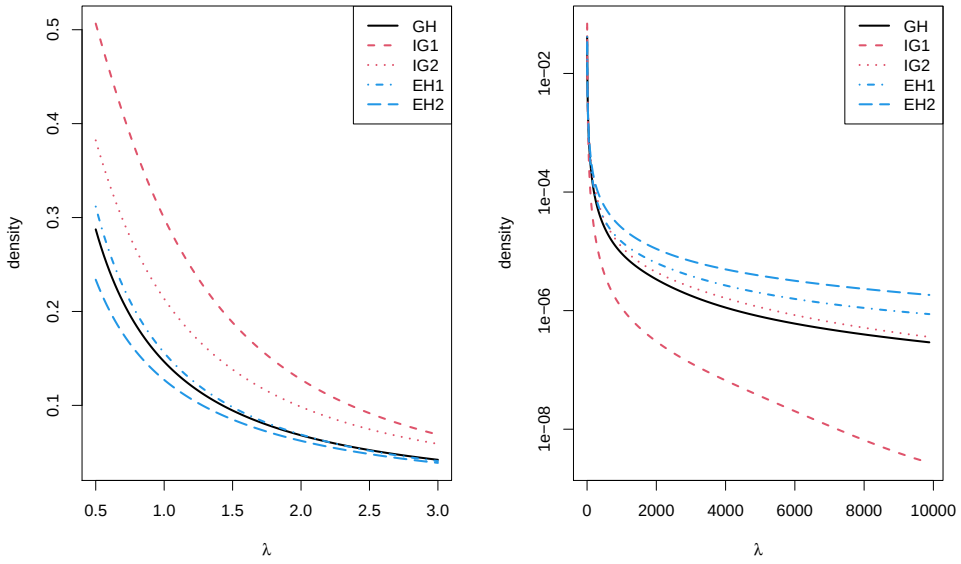

Figure 6.1: Left: Marginal densities of λ_i with $\alpha = \beta = 2$ under the Gauss hypergeometric prior (GH) with $\gamma = 1$, inverse-gamma priors with $\gamma = 1$ (IG1) and $\gamma = 0.5$ (IG2), and extremely heavily-tailed priors with $\gamma = 1$ (EH1) and $\gamma = 0.5$ (EH2). The GH and EH densities are evaluated by the Monte Carlo integration. Right: The marginal densities of the five prior distributions in the tail. The vertical axis is logarithmic.

6.3.4 Marginal posterior distributions for λ_i

We briefly describe the flexibility of the proposed prior distributions compared with the common gamma prior for λ_i. Since the conditional posterior distribution of λ_i given u_i is $\mathrm{Ga}(y_i+\alpha, 1+\beta/u_i)$ under the model (6.2.1), the marginal posterior distribution of λ_i is obtained as the mixture of the gamma distribution with respect to the marginal posterior distribution of u_i. Note that the use of the gamma prior distribution for λ_i leads to the posterior distribution $\mathrm{Ga}(y_i+\alpha, 1+\beta)$. We set $\alpha=\beta=2$ and show the marginal posterior density of λ_i with several values of y_i in Figure 6.2. It is observed that under the moderate signal such as $y_i=1$, the posterior distributions of λ_i are almost the same among the conventional gamma prior and the proposed global-local shrinkage priors. On the other hand, under large values of y_i, the posterior densities of the proposed methods are significantly different from one based on the gamma prior, which shows the flexibility of the proposed priors against large signals and is consistent with tail-robustness property given in Theorem 6.2.1. However, the posterior density with the conventional gamma prior is not sensitive to large signals, which leads to the over-shrinkage of estimators. As noted in the previous section, the hyperparameter γ in the inverse gamma (IG) distribution is directly related to the asymptotic bias, and Figure 6.2 shows that the IG prior with the smaller γ produces heavier-tailed posterior density functions than that with the larger γ.

6.4 Simulation Study

We here investigate the finite sample performance of the proposed method together with some existing methods. We generated the independent sequence of counts from $y_i \sim \mathrm{Po}(\lambda_i\eta_i)$ for $i=1,\ldots,m$ with $m=200$. The adjustment term η_i was generated from $\mathrm{U}(1,5)$, and assumed to be known. For the generating process for λ_i, we considered the mixture: $\lambda_i \sim (1-\omega)f_0+\omega f_1$, where f_0 and f_1 denote distributions of moderate and large signals, respectively. Note that ω denotes the proportion of large signals (outliers). For the settings of f_0 and f_1, we adopted the following four scenarios:

$$
\begin{aligned}
&\text{(I)} && f_0=\mathrm{Ga}(2,2), \quad f_1=\mathrm{Ga}(10,2)\\
&\text{(II)} && f_0=0.75\mathrm{Ga}(2,2)+0.25\delta(1), \quad f_1=\mathrm{Ga}(10,2)\\
&\text{(III)} && f_0=0.5\mathrm{Ga}(2,2)+0.5\delta(1), \quad f_1=\mathrm{Ga}(10,2)\\
&\text{(IV)} && f_0=\mathrm{U}(0,2), \quad f_1=4+|t_3|,
\end{aligned}
$$

where $\mathrm{U}(0,2)$ is the uniform distribution on $[0,2]$ and t_3 is the t-distribution with 3 degrees of freedom. In scenarios (II) and (III), the moderate signals are more concentrated around 1 and have less variation, in comparison to the continuous prior $\mathrm{Ga}(2,2)$ in scenario (I). We define the outlying and non-outlying values of λ_i's as those generated from f_1 and f_0, respectively. In each scenario, we considered two scenarios of ω, namely, $\omega=0.05$ and 0.1.

We considered the estimation of λ_i using the following six priors/methods:

- IG: The proposed method with inverse gamma prior for u_i.
- EH: The proposed method with extremely heavy tailed prior for u_i.
- GH: Gauss hyper-geometric prior proposed by Datta and Dunson (2016)

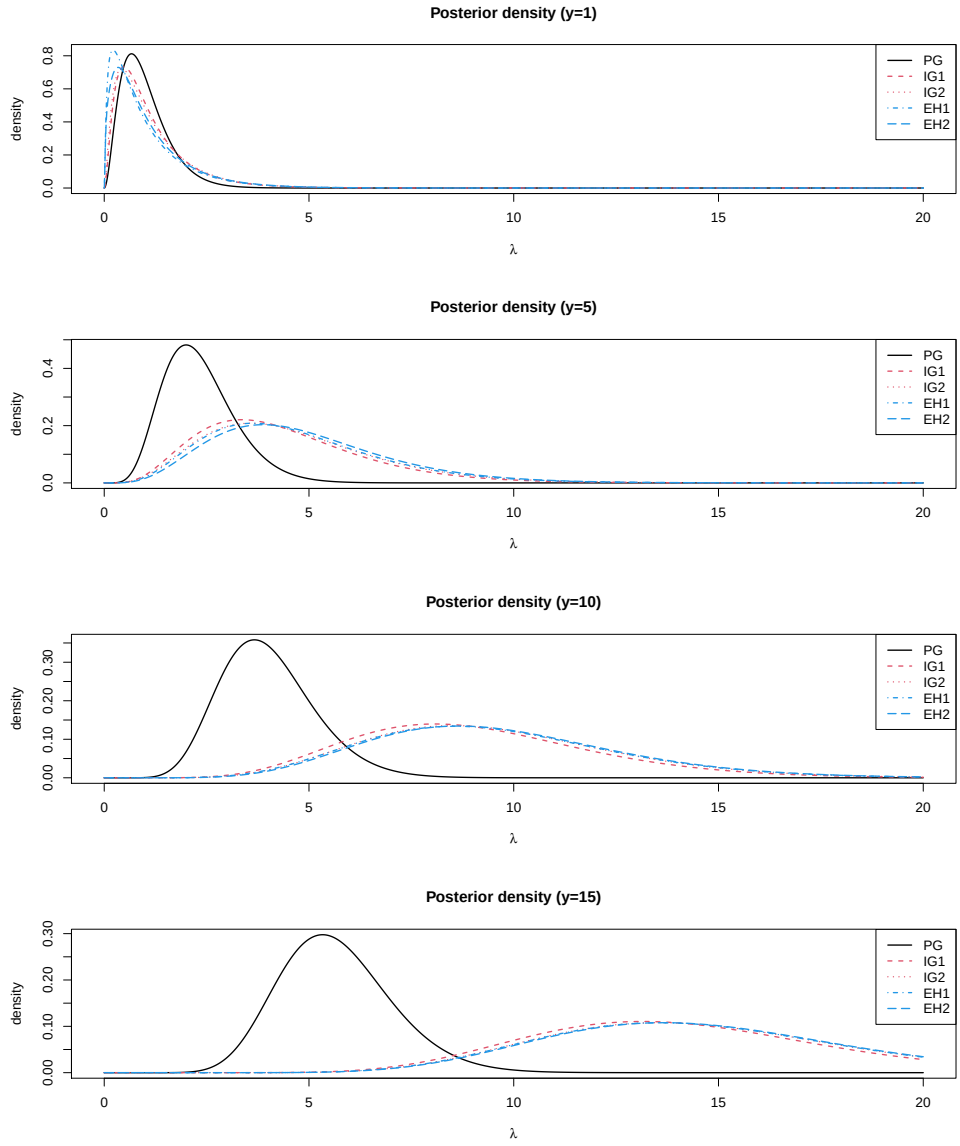

Figure 6.2: Marginal posterior distributions for λ_i with $\alpha = \beta = 2$ based on the conventional gamma prior (PG), the proposed inverse gamma prior with $\gamma = 1$ (IG1) and $\gamma = 0.5$ (IG2), and the proposed extremely heavy-tailed prior with $\gamma = 1$ (EH1) and $\gamma = 0.5$ (EH2). Each row corresponds to a difference value of $y_i \in \{1, 5, 10, 15\}$.

- PG: Using gamma distribution for λ_i, known as Poisson-gamma model.
- KW: Nonparametric empirical Bayes method (Kiefer and Wolfowitz (1956); Koenker and Mizera (2014)).
- ML: Maximum likelihood (non-shrinkage) estimator, i.e., y_i.

We assigned prior distributions for the hyperparameters in the two proposed methods, as illustrated in Section 6.3.2. In the GH method, the hyperparameters were estimated by the empirical Bayes method recommended in Datta and Dunson (2016), and then 3,000 posterior samples were generated directly from the posterior distribution of λ_i with the estimated hyperparameters. We assigned gamma priors for the hyperparameters in the PG method, and used the prior distributions given in Section 6.3.2 for the hyperparameter in the IG and EH methods. The three methods require the computation by Markov chain Monte Carlo method; for each dataset, we generated 3,000 posterior samples after discarding 500 samples as a burn-in period. We computed point estimates of λ_i, where we used the posterior mean as point estimation in the first four methods. The performance of these point estimators are evaluated by the mean squared errors (MSE) and mean absolute percentage error (MAPE) defined as the averaged values of $(\hat{\lambda}_i - \lambda_i)^2$ and $|\hat{\lambda}_i - \lambda_i|/\lambda_i$, respectively. These measures were calculated separately for outlying and non-outlying values of the true λ_i's. We also computed 95% credible intervals of λ_i based on the first four Bayesian methods, and evaluate the performance using the coverage probability (CP) and average length (AL). We repeated the process for 1,000 times to report the averages of MSE, MAPE, CP and AL below.

In Table 6.2, we presented the averaged values of the MSEs and MAPEs in all the scenarios. For non-outlying values, we can see that the PG and KW methods perform quite well while the proposed IG method is quite comparable. For non-outlying values, the performance of the three methods, IG, KW and ML are quite comparable and better than the other methods in MSE. However, it should be noted that the EH performs best in MAPE. These results would show that the shrinkage effects of the proposed methods successfully realized for small signals. On the other hand, for outlying values, the point estimates of both PG and KW methods tend to be worse than ML as predicted theoretically; the PG and KW methods are not tail-robust in general and are expected to produce over-shrunk estimates. In contrast, the proposed IG and EH methods as well as the GH method provides better performance than the PG and KW methods for outliers, as designed. Among the three methods, the EH method provides the best performance in all experiments, which is consistent with the tail-robustness property of the EH method, as discussed in Section 6.3, noting that the IG and GH methods does not necessarily hold the property.

In Table 6.3, we reported averaged values of the CPs and ALs of 95% credible intervals of the four Bayesian methods. It is observed that all the method provides reasonable CP for non-outlying values whereas the CP of the PG method is seriously smaller than the nominal level for outlying values, which also shows the serious over-shrinkage property of the PG method. On the other hand, the proposed methods and the GH method show much higher CPs, while the CP of the EH method is much closer to the nominal level than that of the IG method. It is also observed that the performance of the EH and GH methods are quite comparable in both CP and AL.

We checked the performance of the Markov chain Monte Carlo sampling algorithm for the IG, EH and IG methods under scenario (I) with $\omega = 0.1$. The averaged values of the inefficiency

factors of $\lambda_1, \ldots, \lambda_m$ under the IG, EH and PG methods were 1.17, 4.39 and 1.01, respectively. It shows that the resulting inefficiency factors seems acceptable, but that of the EH method is slightly higher than those of the other methods possibly because the number of latent parameters used in the Gibbs sampling of the EH method is large compared with the other methods. In the Appendix, we report the additional simulation studies with large sample size, namely, $m = 400$, and computation time of the four Bayesian methods.

Table 6.2: Averaged values of mean squared errors (MSE) and mean absolute percentage error (MAPE) in non-outlying (-n) and outlying (-o) areas under four scenarios with $m = 200$ and $\omega \in \{0.05, 0.1\}$.

Scenario	ω		IG	EH	GH	PG	KW	ML
(I)	0.05	MSE-n	0.24	0.28	0.42	0.25	0.26	0.40
		MSE-o	3.30	2.86	2.80	3.86	3.08	2.84
		MAPE-n	0.64	0.57	0.65	0.63	0.67	0.62
		MAPE-o	0.21	0.19	0.19	0.23	0.21	0.19
(I)	0.1	MSE-n	0.26	0.29	0.42	0.28	0.28	0.40
		MSE-o	2.99	2.76	2.69	3.01	2.58	2.73
		MAPE-n	0.64	0.58	0.65	0.63	0.67	0.61
		MAPE-o	0.20	0.19	0.19	0.20	0.19	0.19
(II)	0.05	MSE-n	0.22	0.27	0.43	0.23	0.23	0.40
		MSE-o	3.46	2.90	2.80	4.31	3.06	2.84
		MAPE-n	0.58	0.52	0.61	0.57	0.60	0.58
		MAPE-o	0.22	0.20	0.19	0.24	0.21	0.19
(II)	0.1	MSE-n	0.24	0.28	0.43	0.27	0.24	0.40
		MSE-o	3.05	2.79	2.78	3.13	2.60	2.81
		MAPE-n	0.59	0.54	0.62	0.59	0.62	0.58
		MAPE-o	0.20	0.19	0.19	0.20	0.19	0.19
(III)	0.05	MSE-n	0.19	0.26	0.43	0.21	0.18	0.40
		MSE-o	3.79	3.03	2.90	5.02	3.17	2.94
		MAPE-n	0.50	0.47	0.57	0.50	0.48	0.55
		MAPE-o	0.23	0.20	0.19	0.26	0.21	0.20
(III)	0.1	MSE-n	0.22	0.28	0.44	0.26	0.20	0.41
		MSE-o	3.09	2.78	2.80	3.25	2.54	2.82
		MAPE-n	0.53	0.50	0.58	0.53	0.51	0.55
		MAPE-o	0.20	0.19	0.19	0.21	0.19	0.19
(IV)	0.05	MSE-n	0.21	0.27	0.40	0.21	0.20	0.40
		MSE-o	2.38	1.97	2.01	2.71	2.52	2.07
		MAPE-n	22.06	14.75	12.49	20.71	24.00	0.63
		MAPE-o	0.25	0.22	0.22	0.27	0.25	0.22
(IV)	0.1	MSE-n	0.23	0.28	0.42	0.24	0.23	0.40
		MSE-o	2.12	1.95	2.02	2.14	2.03	2.07
		MAPE-n	2.79	2.05	1.74	2.59	2.66	0.63
		MAPE-o	0.23	0.22	0.22	0.23	0.21	0.22

Table 6.3: Coverage probabilities (CP) and average lengths (AL) of 95% credible intervals in non-outlying (n) and outlying (o) areas under four scenarios with $m = 200$ and $\omega \in \{0.05, 0.1\}$.

Scenario	ω		IG	EH	GH	PG	IG	EH	GH	PG
(I)	0.05	n	96.0	96.2	95.6	96.6	1.93	2.01	2.32	1.99
		o	88.1	91.7	94.3	80.8	5.57	5.81	6.27	4.83
	0.1	n	96.3	96.4	95.7	96.6	2.01	2.05	2.33	2.10
		o	90.7	92.4	94.8	88.7	5.71	5.83	6.25	5.20
(II)	0.05	n	96.2	96.3	95.5	96.9	1.90	2.02	2.36	1.98
		o	87.0	91.7	94.6	77.0	5.49	5.75	6.23	4.65
	0.1	n	96.4	96.4	95.5	96.8	2.00	2.07	2.37	2.12
		o	90.2	92.3	94.8	87.3	5.71	5.83	6.28	5.12
(III)	0.05	n	96.7	96.4	95.4	97.3	1.88	2.04	2.40	1.97
		o	84.8	90.9	94.1	69.9	5.42	5.73	6.23	4.47
	0.1	n	96.9	96.5	95.3	97.1	1.98	2.09	2.40	2.12
		o	89.8	92.2	94.8	86.0	5.69	5.82	6.27	5.03
(IV)	0.05	n	93.9	95.6	95.4	95.2	1.89	2.01	2.29	1.91
		o	84.5	91.4	94.3	77.5	4.35	4.83	5.33	3.80
	0.1	n	94.7	95.8	95.5	95.7	1.99	2.05	2.33	2.04
		o	88.0	91.6	94.6	85.8	4.51	4.85	5.32	4.13

6.5 Data Analysis

We apply the proposed method to the analysis of crime data by the generalized linear model with Poisson likelihood and random effects. This model has been adopted for various datasets in applied statistics; examples include the modeling of areal count data in disease mapping (Lawson (2013)). In such application, Poisson rate λ_i (defined below) is not just an adjustment of areal effects but the parameter of interest as the intrinsic relative risk of region i (e.g. Li et al. (2010)). Here we incorporate such idea of covariate adjustment into crime risk modeling.

The dataset consists of the numbers of police-recorded crime in Tokyo metropolitan area, provided by University of Tsukuba and publicly available online ("GIS database of number of police-recorded crime at O-aza, chome in Tokyo, 2009-2017", available at `https://commons.sk.tsukuba.ac.jp/data`). In this study, we focus on the number of violent crimes in $m = 2855$ local towns in Tokyo metropolitan area in 2015. For auxiliary information about each town, we adopted area (km^2), population densities in noon and night, density of foreign people, percentage of single-person household and average duration of residence, which all help adjustment of the crime risk. Let y_i be the observed count of violent crimes, a_i be area and $\boldsymbol{x}_i$ be the vector of standardized auxiliary information in the i-th local town. We are interested in the crime risk adjusted by the auxiliary information and, to this end, we employ the following Poisson regression model:

$$y_i|\lambda_i \sim \mathrm{Po}(\lambda_i \eta_i), \qquad \eta_i = \exp(\log a_i + \boldsymbol{x}_i{}^\top \boldsymbol{\delta}), \tag{6.5.1}$$

independently for $i = 1, \ldots, m$, where $\boldsymbol{\delta}$ is a vector of unknown regression coefficients. Under the model (6.5.1), the random effect for local town i, λ_i, can be interpreted as adjustment risk

factor that is not be explained by the auxiliary information. In most local towns, the offset term explains the variations of crime rates, hence the adjustment risk factor is expected to be small. Yet, the adjustment risk might be extremely high in some local towns, and we want to detect such districts. This is precisely where the global-local shrinkage priors fit, for which we employed the proposed the IG and EH priors for λ_i. We adopted $\mathrm{N}(0, 100)$ as a prior distribution of each component of $\boldsymbol{\delta}$; we found the following result was robust to the choice of prior variance. For posterior inference, we simply use a Gibbs sampling in which the posterior samples of $\lambda_1, \ldots, \lambda_m$ and $\boldsymbol{\delta}$ are iteratively drawn from their full conditional distributions. Conditional on δ, we can still use the posterior computation algorithm for λ_i provided in Section 6.3.2. On the other hand, given λ_i's, the full conditional distribution of $\boldsymbol{\delta}$ is not a familiar form. The detailed algorithm customized for sampling of $\boldsymbol{\delta}$ is based on the independent Metropolis-Hasting method and given in the Appendix. For comparison, we also applied the common gamma distribution for λ_i as considered in Section 6.4, which is again denoted by PG in what follows. Regarding other methods used in Section 6.4, the GH prior cannot be directly applied in this case since the specification method for hyperparameters recommended in Datta and Dunson (2016) is reasonable only when there is no adjustment terms. Similarly, the KW method is not applicable in this situation. Therefore, we will focus on the comparison of the proposed priors with the standard Gamma prior. In each Gibbs sampler, we generated 20,000 posterior samples after discarding 3,000 posterior samples as burn-in.

We first computed posterior means of risk factor λ_i based on the three methods. The spatial pattern of the estimates is shown in Figure 6.3. It is observed that the proposed two priors, IG and EH, produce almost the same estimates. We can confirm that the proposed EH method provides similar estimates of λ_i in most areas and successfully detected several local towns whose risk factors are extremely high. In contrast, such extreme towns are less emphasized, or not detected at all, by the PG method because the PG method seriously underestimates the true risk factors. More direct comparisons of estimates based on the purposed methods and the PG method are presented in Figure 6.4, which indicates the underestimation property of the PG method more clearly.

We then detected ten local towns with the largest posterior means of λ_i. For these towns, we computed 95% credible intervals of λ_i based on the three methods, as shown in the left panel of Figure 6.5. This panel clearly shows the over-shrinkage problem of the PG method in both point estimation (posterior means) and interval estimation (posterior credible intervals); the posterior credible intervals computed by the PG method tends to be narrow and further emphasizes the underestimated results. We also randomly selected another ten local towns with moderate estimates of λ_i and gave 95% credible intervals in the right panel of the same figure. The difference of the three methods is almost negligible for these towns. These observations exemplify that the proposed methods can avoid the over-shrinkage problem for large signals while their performance in the other towns are almost the same as the standard PG method.

6.6 Discussion

It should be emphasized again that the global-local shrinkage priors for sequence of counts developed in this article are based on the new concept of tail-robustness, that is clearly different from other definitions and non-trivial for many prior densities. We provided sufficient conditions for this desirable tail-robustness property and, specifically, proposed two tractable global-local

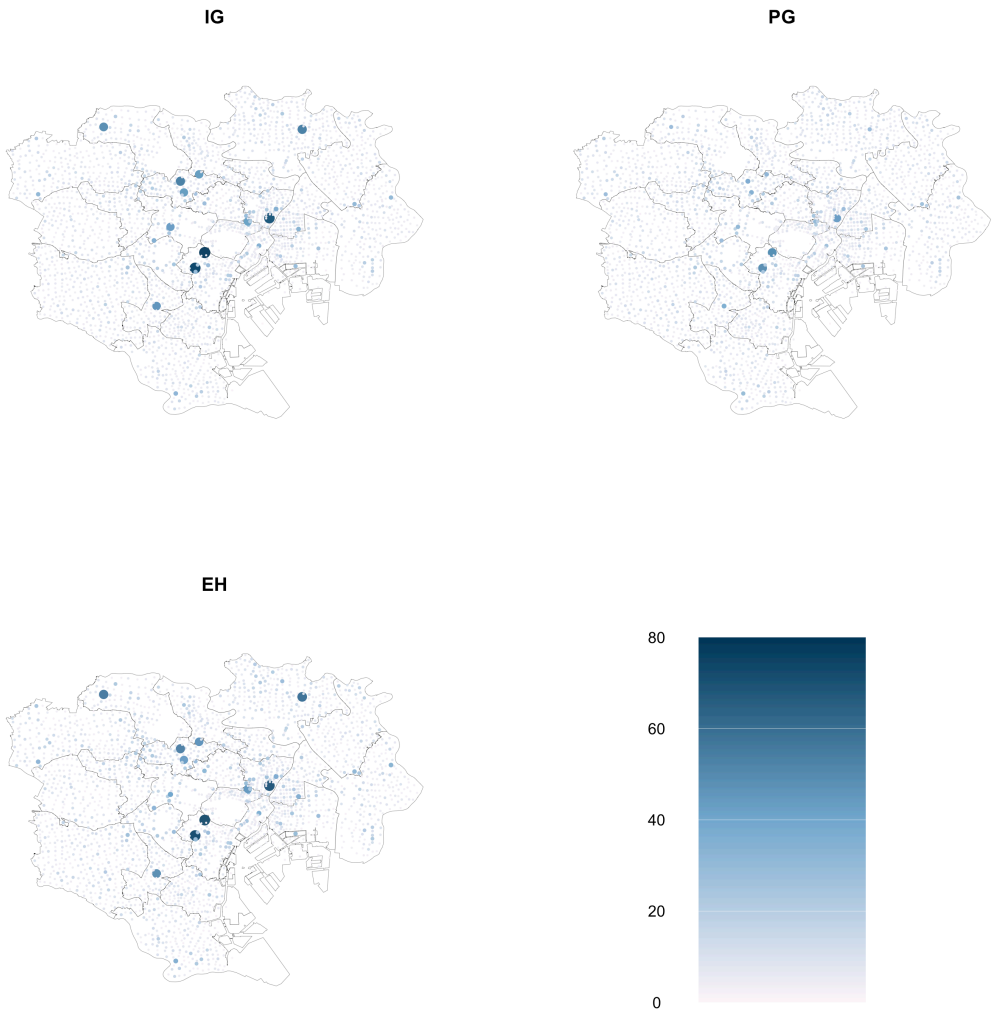

Figure 6.3: Posterior means of risk factors λ_i based on IG, EH and PG methods.

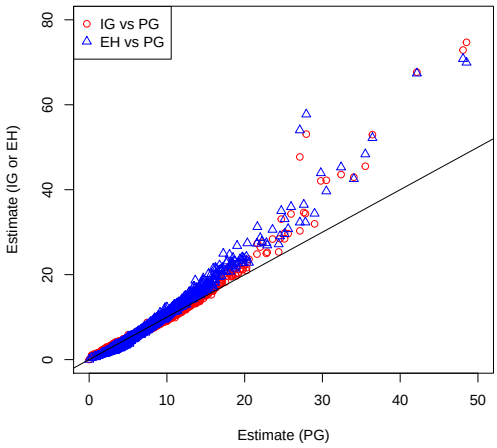

Figure 6.4: Scatter plot of posterior means of risk factors λ_i based on IG, EH and PG methods.

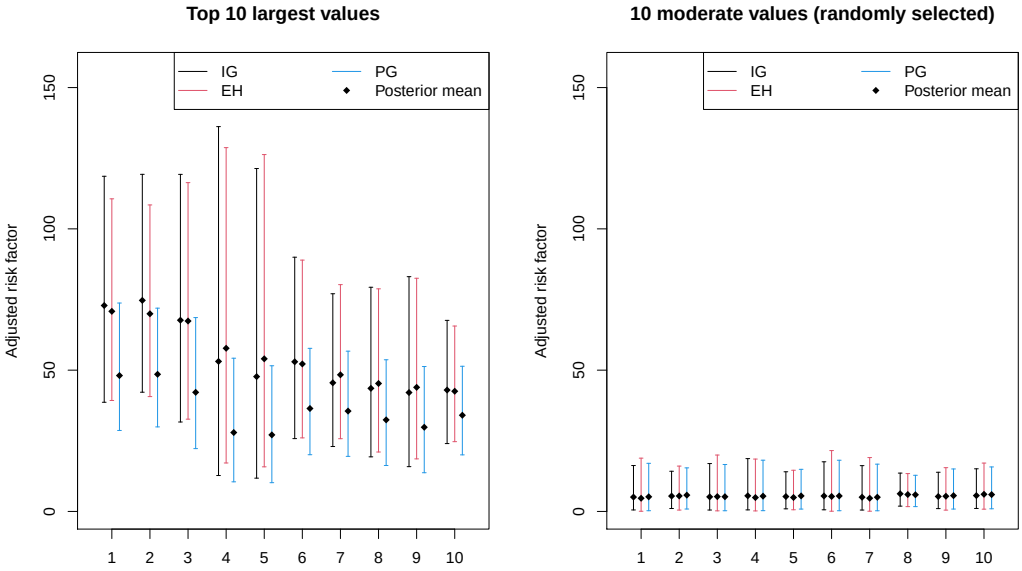

Figure 6.5: 95% credible intervals for areas with highest 10 posterior means (left) and for randomly selected 10 areas with moderate posterior means (right) of adjusted risk factors.

shrinkage priors. As illustrated by the simulated and real data examples, the models with these priors could actually show the tail-robustness as predicted by theory, and are expected to be applied in the studies of high-dimensional counts.

The settings of our study are critically dependent on the Poisson likelihoods, whose mean and variance are equal. Conditionally, both prior mean and variance, λ_i, are controlled by the common local parameter u_i under the gamma prior $\mathrm{Ga}(\alpha, \beta/u_i)$, affecting both the baseline of shrinkage and the amount of shrinkage. This property is not seen in the Gaussian case, where the local parameter appears in the prior variance only and controls the amount of local shrinkage, which makes the role of local parameters clear and interpretable. In this sense, the local parameter in (6.2.1) might be less interpretable, while it is also this setting that enables us to carry out posterior computation easily and has been studied intensively in the literature (e.g. Datta and Dunson (2016)). It would be an interesting future research to pursue an alternative setting for hierarchical modeling of sequence of counts under which the role of the local parameters is properly restricted and interpretable.

From the viewpoint of methodological research, this chapter is primarily focused on the point and interval estimation of the Poisson rate. The high-dimensional counts can be cast as other statistical problems such as multiple testing. The detailed investigation for such directions would extend the scope of this chapter, but we leave it to a valuable future study.

The newly-introduced EH prior is motivated as the probability distributions that satisfy the conditions given in Theorem 6.2.1, hence hold tail-robustness. However, the class of priors that meet those conditions is not limited to that of the EH priors. In theory, the priors with tail-robustness can be extended to

$$\pi(u_i) \propto \frac{{u_i}^{\gamma_1 - 1}}{(1+\gamma_2 u_i)^{\gamma_1}} \frac{1}{\{1+\gamma_3 \log(1+\gamma_4+u_i)\}^{1+\gamma_5}},$$

which is also proper and tail-robust. The hyperparameters $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)$ increases the flexibility of the model and could improve the EH prior equipped with a single parameter γ. However, the posterior inference under this prior is challenging due to the intractable normalizing constant that involves those hyperparameters. The full-Bayes inference for the hyperparameters is not as straightforward as that of the EH prior. The inference with fixed hyperparaemters is feasible by utilizing the same parameter augmentation in Section 6.3.2, but always raises the problem of hyperparameter tuning. We leave the development of this extension to the future work, which could be useful in more structured models for count data.

6.7 Appendix

6.7.1 Posterior computation algorithm

We here provide details of the posterior computation algorithm under the proposed two priors, IG and EH priors, under the hierarchical model (6.2.1). The contents consists of three parts, algorithms for sampling from the common parameters in (6.2.1), parameters related to the IG prior and parameters related to the EH prior.

Sampling of the common parameters $(\boldsymbol{\lambda}_{1:m}, \alpha, \beta)$.

- The full conditional of λ_i is $\mathrm{Ga}(y_i+\alpha, \eta_i+\beta/u_i)$ and $\lambda_1, \ldots, \lambda_m$ are mutually independent.

- The full conditional of β is $\mathrm{Ga}(m\alpha + a_\beta, \sum_{i=1}^m \lambda_i/u_i + b_\beta)$.
- The sampling of dispersion parameter α can be done in multiple ways. We take the strategy of Zhou and Carin (2013) by working on the conditional, negative binomial likelihood of α by marginalizing λ_i out. The conditional posterior density of α is proportional to

$$\psi_\alpha(\alpha)\prod_{i=1}^m \frac{\Gamma(y_i+\alpha)}{\Gamma(\alpha)}\left(\frac{\beta}{\beta+\eta_i u_i}\right)^\alpha = \psi_\alpha(\alpha)\prod_{i=1}^m \sum_{\nu_i=1}^{y_i} |s(y_i,\nu_i)|\alpha^{\nu_i}\left(1+\frac{\eta_i u_i}{\beta}\right)^{-\alpha},$$

  where $\psi_\alpha(\alpha)$ is the prior density of α and $s(y_i, \nu_i)$ is the Stirling's number of the second kind, and the summation collapses to one if $y_i = 0$. The integer-valued variable ν_i is considered a latent parameter that augments the model and allows Gibbs sampler. Thus, we need to sample from the full conditionals of α and $\boldsymbol{\nu}_{1:m}$.

  - The conditional of α is $\mathrm{Ga}(\sum_{i=1}^m \nu_i + a_\alpha, \sum_{i=1}^m \log(1+\eta_i u_i/\beta) + b_\alpha)$.
  - If $y_i = 0$, then $\nu_i = 0$ with probability one. Otherwise, the conditional posterior probability function of ν_i is proportional to $|s(y_i, \nu_i)|\alpha^{\nu_i}$, from which we can sample based on the distributional equation $\nu_i = \sum_{j=1}^{y_i} d_j$, where d_j $(j = 1, \ldots, y_i)$ are independent random variables distributed as $\mathrm{Ber}\,(\alpha/(j-1+\alpha))$.

Sampling of parameters related to IG prior

- The full conditional of u_i is $\mathrm{IG}(\gamma+\alpha, \gamma+\lambda_i\beta)$ and $u_1, \ldots, u_m$ are mutually independent.
- The full conditional of γ is proportional to

$$f_\gamma(\gamma) = \frac{\gamma^{m\gamma}}{\{\Gamma(\gamma)\}^m}\Big(\prod_{i=1}^m \frac{1}{u_i}\Big)^\gamma \exp\left(-\gamma\sum_{i=1}^m \frac{1}{u_i}\right) I(\varepsilon_1 \le \gamma \le \varepsilon_2).$$

  We sample the candidate of γ, denoted γ^*, from the distribution of $\min\{\varepsilon_2, \max\{\varepsilon_1, Z\}\}$, $Z \sim \mathrm{N}(\widetilde{\gamma}, \sigma^2)$, with tuning parameter $\sigma > 0$, where $\widetilde{\gamma}$ is the current value of γ, and accept it with probability $\min\{1, f_\gamma(\gamma^*)/f_\gamma(\tilde{\gamma})\}$, where we assume that the correction factor based on the asymmetric proposal density can be ignored.

Sampling parameters related to the EH prior

The latent parameters, (v_i, w_i), are marginalized out except for the sampling of u_i (Partially collapsed Gibbs sampler, van Dyk and Park (2019)).

- The full conditional of u_i is $\mathrm{GIG}(1-\alpha, 2v_i, 2\beta\lambda_i)$, where $\mathrm{GIG}(a, b, p)$ is the generalized inverse Gaussian distribution with density $\pi(x; a, b, p) \propto x^{p-1}\exp\{-(ax + b/x)/2\}$ for $x > 0$.
- The full conditional of $(v_i|w_i)$ is $\mathrm{Ga}(1+w_i, 1+u_i)$. The conditional posterior of w_i with v_i marginalized out is $\mathrm{Ga}(1+\gamma, 1+\log(1+u_i))$.
- Under gamma prior $\gamma \sim \mathrm{Ga}(a_\gamma, b_\gamma)$, the full conditional of γ (with (v_i, w_i) marginalized out) is $\mathrm{Ga}(a_\gamma + m, b_\gamma + \sum_{i=1}^m \log\{1+\log(1+u_i)\})$.

6.7.2 Lemmas

In this section, we provide two lemmas which will be used in the proof of Theorem 6.2.1 in the next section. The following lemma is useful for proving Proposition 6.7.1 as well.

Lemma 6.7.1 *Let $0 < M_1 < M_0 < \infty$. Let $h_0(\cdot)$ and $h_1(\cdot)$ be nonnegative integrable functions defined on (M_0, ∞) and $(0, M_1)$, respectively, and let $0 < \varphi(\cdot) < 1$ be a strictly increasing function defined on $(0, \infty)$. Suppose that $\int_{M_0}^{\infty} h_0(u)du > 0$. Then*

$$\lim_{y\to\infty} \int_0^{M_1} \{\varphi(u)\}^y h_1(u)du / \int_{M_0}^{\infty} \{\varphi(u)\}^y h_0(u)du = 0.$$

Proof. We have

$$\begin{aligned}
&\limsup_{y\to\infty} \int_0^{M_1} \{\varphi(u)\}^y h_1(u)du / \int_{M_0}^{\infty} \{\varphi(u)\}^y h_0(u)du \\
&\le \limsup_{y\to\infty} \Big\{\frac{\varphi(M_1)}{\varphi(M_0)}\Big\}^y \int_0^{M_1} h_1(u)du / \int_{M_0}^{\infty} h_0(u)du \\
&= 0
\end{aligned}$$

by assumption. □

Lemma 6.7.2 *The assumptions of Theorem 6.2.1 imply the following:*

$$\int_0^{\infty} \frac{|u\pi'(u)|}{(\beta+u)^{\alpha}}du < \infty, \tag{6.7.1}$$

$$\lim_{u\to\infty} \frac{u\pi(u)}{(\beta+u)^{\alpha}} = \lim_{u\to 0} \frac{u\pi(u)}{(\beta+u)^{\alpha}} = 0. \tag{6.7.2}$$

Proof. We first note that if $\pi(\cdot)$ is to be proper, we must have $\xi \in [-\infty, 0]$ since otherwise $\pi(\cdot)$ would be eventually increasing so that

$$\int_N^{\infty} \pi(u)du \ge \int_N^{\infty} \pi(N)du = \infty$$

for some $N > 0$. By (A1) of the main text, we have

$$\int_0^1 \frac{|u\pi'(u)|}{(\beta+u)^{\alpha}}du \le \int_0^1 \frac{|u\pi'(u)|}{\beta^{\alpha}}du < \infty.$$

If $\xi > -\infty$, then $|u\pi'(u)| = O(\pi(u))$ as $u \to \infty$ and hence

$$\int_1^{\infty} \frac{|u\pi'(u)|}{(\beta+u)^{\alpha}}du \le \int_1^{\infty} \frac{|u\pi'(u)|}{\beta^{\alpha}}du < \infty.$$

On the other hand, if $\xi = -\infty$, then there exists $N > 0$ such that $\pi'(u) < 0$ for all $u \ge N$ and therefore

$$\begin{aligned}
\infty &> \limsup_{M\to\infty} \int_N^M \pi(u)du = \limsup_{M\to\infty} \Big[M\pi(M) - N\pi(N) + \int_N^M \{-u\pi'(u)\}du\Big] \\
&\ge \limsup_{M\to\infty} \Big[- N\pi(N) + \int_N^M \{-u\pi'(u)\}du\Big] = -N\pi(N) + \int_N^{\infty} |u\pi'(u)|du
\end{aligned}$$

by integration by parts. Thus, (6.7.1) follows. To prove (6.7.2), note that for any $0 < \delta < 1 < M < \infty$, we have

$$\Big[u\pi(u)\Big]_{\delta}^{M} = \int_{\delta}^{M} \pi(u)du + \int_{\delta}^{M} u\pi'(u)du$$

by integration by parts. Then, since the right-hand side of the above equation converges as $\delta \to 0$ and as $M \to \infty$, there exist $c_0, \mathring{c} \in [0, \infty)$ such that $\lim_{u\to 0} u\pi(u) = c_0$ and $\lim_{u\to\infty} u\pi(u) = \mathring{c}$. If $c_0 > 0$ or $\mathring{c} > 0$, then $u^{-1} = O(\pi(u))$ as $u \to 0$ or as $u \to \infty$ in contradiction to the assumption that $\pi(\cdot)$ is proper. Thus, $c_0 = \mathring{c} = 0$ and (6.7.2) follows. □

6.7.3 Proof of Theorem 6.2.1

In this section, we prove Theorem 6.2.1.

Proof of Theorem 6.2.1. We prove the result by using (6.7.1), (6.7.2), and (A2). Since the posterior density of u_i given y_i is proportional to $W(u_i)\pi(u_i)$, where $W(u_i) = W(u_i; y_i) = {u_i}^{y_i}/(1 + u_i/\beta)^{y_i+\alpha}$, the difference between y_i and $\tilde{\lambda}_i$ is

$$y_i - \tilde{\lambda}_i = \int_0^{\infty} \frac{y_i - \alpha u_i/\beta}{1 + u_i/\beta} W(u_i)\pi(u_i)du_i \Big/ \int_0^{\infty} W(u_i)\pi(u_i)du_i, \tag{6.7.3}$$

which is finite by the propriety of the posterior. By making the change of variables $t = (u_i/\beta)/(1 + u_i/\beta)$, we have

$$y_i - \tilde{\lambda}_i = \int_0^1 \{y_i - (y_i + \alpha)t\}g(t)dt \Big/ \int_0^1 g(t)dt,$$

where $g(t) = g(t; y_i) = t^{y_i}(1-t)^{\alpha-2}\pi(\beta t/(1-t))$. Note that, by integration by parts and (6.7.2),

$$\begin{aligned}
(\alpha+1)\int_0^1 tg(t)dt &= \int_0^1 (\alpha+1)(1-t)^{\alpha}t^{y_i+1}(1-t)^{-2}\pi\Big(\beta\frac{t}{1-t}\Big)dt \\
&= \Big[-(1-t)^{\alpha+1}t^{y_i+1}(1-t)^{-2}\pi\Big(\beta\frac{t}{1-t}\Big)\Big]_{t=0}^{t=1} \\
&\quad + \int_0^1 \Big[(1-t)^{\alpha+1}t^{y_i+1}(1-t)^{-2}\pi\Big(\beta\frac{t}{1-t}\Big) \\
&\qquad \times \Big\{\frac{y_i+1}{t} + \frac{2}{1-t} + \frac{\partial}{\partial t}\log\pi\Big(\beta\frac{t}{1-t}\Big)\Big\}\Big]dt \\
&= (y_i+1)\int_0^1 (1-t)g(t)dt + 2\int_0^1 tg(t)dt \\
&\quad + \beta\int_0^1 \frac{t}{1-t}\Big\{\pi'\Big(\beta\frac{t}{1-t}\Big)/\pi\Big(\beta\frac{t}{1-t}\Big)\Big\}g(t)dt,
\end{aligned}$$

or

$$\int_0^1 \{y_i - (y_i+\alpha)t\}g(t)dt = -\int_0^1 g(t)dt - \beta\int_0^1 \frac{t}{1-t}\Big\{\pi'\Big(\beta\frac{t}{1-t}\Big)/\pi\Big(\beta\frac{t}{1-t}\Big)\Big\}g(t)dt.$$

Then, by making the change of variables $u_i = \beta t/(1-t)$, we obtain

$$\begin{aligned} y_i - \tilde{\lambda}_i &= -1 - \beta \left(\int_0^1 g(t)dt \right)^{-1} \int_0^1 \frac{t}{1-t} \Big\{ \pi'\Big(\beta \frac{t}{1-t}\Big) / \pi\Big(\beta \frac{t}{1-t}\Big) \Big\} g(t)dt \\ &= -1 - \int_0^\infty H(u_i) u_i \pi'(u_i) du_i \Big/ \int_0^\infty H(u_i)\pi(u_i)du_i, \end{aligned}$$

where the integrands are absolutely integrable by (6.7.1) and

$$H(u_i) = H(u_i;\beta) = \left(\frac{u_i/\beta}{1+u_i/\beta} \right)^{y_i} \frac{1}{(1+u_i/\beta)^\alpha}.$$

Now, suppose first that $\xi > -\infty$. Then, for any $M > 0$,

$$\begin{aligned} &|y_i - \tilde{\lambda}_i + 1 + \xi| && (6.7.4) \\ &\le \int_0^\infty \Big| \xi - \frac{u_i \pi'(u_i)}{\pi(u_i)} \Big| H(u_i)\pi(u_i)du_i \Big/ \int_0^\infty H(u_i)\pi(u_i)du_i && (6.7.5) \\ &= \frac{\int_0^\infty H(u_i)h_1(u_i)du_i}{\int_0^\infty H(u_i)h_0(u_i)du_i} \\ &\le \frac{\int_0^M H(u_i)h_1(u_i)du_i}{\int_{M+1}^\infty H(u_i)h_0(u_i)du_i} + \frac{\int_M^\infty H(u_i)h_1(u_i)du_i}{\int_0^\infty H(u_i)h_0(u_i)du_i}, \end{aligned}$$

where $h_k(u_i) = |\xi - u_i\pi'(u_i)/\pi(u_i)|^k \pi(u_i)$ for $k = 0, 1$. The first term in the fourth line converges to zero as $y_i \to \infty$ by Lemma 6.7.1. On the other hand,

$$\begin{aligned} &\limsup_{M\to\infty} \sup_{y_i \in \{0,1,2,\dots\}} \frac{\int_M^\infty H(u_i)h_1(u_i)du_i}{\int_0^\infty H(u_i)h_0(u_i)du_i} \\ &= \limsup_{M\to\infty} \sup_{y_i \in \{0,1,2,\dots\}} \frac{\int_M^\infty \big|\xi - u_i\pi'(u_i)/\pi(u_i)\big| H(u_i)h_0(u_i)du_i}{\int_M^\infty H(u_i)h_0(u_i)du_i} \\ &\le \limsup_{M\to\infty} \sup_{u_i \in (M,\infty)} \Big| \xi - \frac{u_i\pi'(u_i)}{\pi(u_i)} \Big| = \lim_{u_i\to\infty} \Big| \xi - \frac{u_i\pi'(u_i)}{\pi(u_i)} \Big| = 0. \end{aligned}$$

Thus,

$$\begin{aligned} \limsup_{y_i\to\infty} |y_i - \tilde{\lambda}_i + 1 + \xi| &\le \limsup_{M\to\infty} \limsup_{y_i\to\infty} \frac{\int_0^M H(u_i)h_1(u_i)du_i}{\int_{M+1}^\infty H(u_i)h_0(u_i)du_i} \\ &\quad + \limsup_{M\to\infty} \limsup_{y_i\to\infty} \frac{\int_M^\infty H(u_i)h_1(u_i)du_i}{\int_0^\infty H(u_i)h_0(u_i)du_i} \\ &\le 0 + 0 = 0. \end{aligned}$$

Next, suppose that $\xi = -\infty$. Then for any $M > 0$, there exists $N > 0$ such that $-u\pi'(u)/\pi(u) > M$ for all $u \ge N$. Therefore,

$$\begin{aligned} y_i - \tilde{\lambda}_i + 1 &= -\frac{\int_0^N H(u_i)u_i\pi'(u_i)du_i}{\int_0^\infty H(u_i)\pi(u_i)du_i} - \frac{\int_N^\infty H(u_i)u_i\pi'(u_i)du_i}{\int_0^\infty H(u_i)\pi(u_i)du_i} \\ &\ge -\frac{\int_0^N H(u_i)u_i\pi'(u_i)du_i}{\int_0^\infty H(u_i)\pi(u_i)du_i} + M\frac{\int_N^\infty H(u_i)\pi(u_i)du_i}{\int_0^\infty H(u_i)\pi(u_i)du_i} \to M \end{aligned}$$

as $y_i \to \infty$ by Lemma 6.7.1. Thus, since M is arbitrary, we conclude that

$$\lim_{y_i \to \infty} (\tilde{\lambda}_i - y_i) = -\infty = 1 + \xi.$$

This completes the proof. □

6.7.4 Related tail-robustness properties

We here discuss two related tail-robustness properties of the posterior mean $\tilde{\lambda}_i$. One variant is based on the ratio of the estimator and observation and given by

$$\lim_{y_i \to \infty} \frac{|\tilde{\lambda}_i - y_i|}{y_i} = 0, \tag{6.7.6}$$

which we name *weak tail-robustness*. It is obvious that the strong tail-robustness implies the weak one. The left-hand-side in (6.7.6) is the mean absolute percentage error (MAPE) loss function, which is frequently used in practice to evaluate the inferential/predictive performance of the models for count data. In this sense, the weakly tail-robust estimator $\tilde{\lambda}_i$ is asymptotically optimal in MAPE (Section 3.3.2, Berry et al. (2019)). Note that the Bayes estimator $(\alpha + y_i)/(1 + \beta/u_i)$ with fixed u_i does not satisfy the property (6.7.6).

We provide conditions for weak tail-robustness in the following proposition.

Proposition 6.7.1 *Suppose that $\pi(\cdot)$ is strictly positive. Then, under the model (6.2.1), we have*

$$\lim_{y_i \to \infty} \frac{|\tilde{\lambda}_i - y_i|}{y_i} = 0.$$

i.e., the Bayes estimator is weakly tail-robust.

Proof. From (6.7.3), we have

$$\frac{\tilde{\lambda}_i - y_i}{y_i} = -\frac{\beta \int_0^\infty (\beta + u_i)^{-1} W(u_i)\pi(u_i)du_i}{\int_0^\infty W(u_i)\pi(u_i)du_i} + \frac{\alpha}{y_i} \frac{\int_0^\infty u_i(\beta + u_i)^{-1} W(u_i)\pi(u_i)du_i}{\int_0^\infty W(u_i)\pi(u_i)du_i} \tag{6.7.7}$$

for $y_i \in \{1, 2, \dots\}$, where

$$W(u_i) = W(u_i; y_i) = \frac{{u_i}^{y_i}}{(1 + u_i/\beta)^{y_i + \alpha}}.$$

The second term on the right-hand side of (6.7.7) converges to zero as $y_i \to \infty$ since $u_i/(\beta + u_i) \leq 1$ for all $u_i \in (0, \infty)$. On the other hand, by Lemma 6.7.1,

$$\begin{aligned} &\frac{\int_0^\infty \beta(\beta + u_i)^{-1} W(u_i)\pi(u_i)du_i}{\int_0^\infty W(u_i)\pi(u_i)du_i} \\ &= \frac{\int_M^\infty \beta(\beta + u_i)^{-1} W(u_i)\pi(u_i)du_i}{\int_M^\infty W(u_i)\pi(u_i)du_i} \\ &\quad \times \frac{\int_0^\infty \beta(\beta + u_i)^{-1} W(u_i)\pi(u_i)/\int_M^\infty \beta(\beta + u_i)^{-1} W(u_i)\pi(u_i)du_i}{\int_0^\infty W(u_i)\pi(u_i)du_i/\int_M^\infty W(u_i)\pi(u_i)du_i} \\ &\sim \frac{\int_M^\infty \beta(\beta + u_i)^{-1} W(u_i)\pi(u_i)du_i}{\int_M^\infty W(u_i)\pi(u_i)du_i} \end{aligned} \tag{6.7.8}$$

as $y_i \to \infty$ for every $M > 0$. Furthermore, uniformly in y_i,

$$0 \leq \frac{\int_M^\infty \beta(\beta + u_i)^{-1} W(u_i)\pi(u_i)du_i}{\int_M^\infty W(u_i)\pi(u_i)du_i} \leq \frac{\beta}{\beta + M} \to 0 \quad \text{as} \quad M \to \infty.$$

Thus, we have proved the desired result. □

The key implication of this proposition is that, for fixed hyperparameters, the weak tail-robustness can be achieved for almost all priors for u_i. It suggests that the tail-robustness property in the main document and the weak tail-robustness property look similar but is substantially different properties.

Another concept of tail-robustness is

$$\lim_{y_i \to \infty} \frac{\tilde{\lambda}_i}{\alpha + y_i} = 1. \tag{6.7.9}$$

The denominator is a part of the Bayes estimator $(\alpha + y_i)/(1 + \beta/u_i)$. This definition requires that the coefficient $u_i/(\beta + u_i)$, viewed as a shrinkage factor, degenerates at 1 as $y_i \to \infty$ (Proposition 1, Datta and Dunson (2016)). It is trivial that the Bayes estimator with fixed u_i does not satisfy the property, but the weak tail-robustness leads to the tail-robustness of this type. Hence, by Proposition 6.7.1, the use of any strictly positive prior of u_i also leads to this tail-robustness.

6.7.5 Connection to the tail-robustness of three-parameter beta priors

The EH prior emerges in the course of examination of tail-robustness under the scaled-beta or three-parameter beta (TPB) distributions (Armagan et al. (2011)), known as a flexible class of priors for scale parameters. The density is given by

$$\pi(u_i; a_0, b_0, \phi_0, \gamma_0) \propto \frac{{u_i}^{a_0 - 1}}{(1 + \phi_0 u_i/\beta)^{\gamma_0}(1 + u_i/\beta)^{a_0 + b_0 - \gamma_0}}, \tag{6.7.10}$$

where a_0, b_0, ϕ_0, and γ_0 are all positive constants. For count data and Poisson likelihood, Datta and Dunson (2016) considered this prior with $a_0 = b_0 = 1/2$ and $\phi_0 = \beta$. Although the TPB prior (6.7.10) is flexible, it does not satisfy assumption (A3) in Corollary 6.2.1 and is not strongly tail-robust for any choice of hyperparameters. Under the prior (6.7.10), by Theorem 1, the asymptotic bias is $\lim_{y_i \to \infty}(\tilde{\lambda}_i - y_i) = -b_0 < 0$, negatively biased and dependent on its hyperparameter b_0. Similar to the inverse-gamma prior, the approximate tail-robustness for the TPB prior is justified by the limiting case of $b_0 \to 0$; with $\gamma_0 = \gamma$, $a_0 = 1 + \gamma$, and $\beta = \phi_0 = 1$ in (6.7.10), we obtain

$$\pi(u_i; \gamma) \propto \frac{u_i^\gamma}{(1 + u_i)^{1+\gamma}}, \quad \gamma > -1. \tag{6.7.11}$$

In return for the tail-robustness in the limit, however, it is inevitable for the prior in (6.7.11) to be improper. The EH prior can be viewed as the limit $\gamma \to \infty$ modified by the multiplied log-term for propriety.

6.7.6 Evaluation of the marginal of λ_i with EH prior

We evaluate the limit of the marginal density of λ_i implied by the EH prior.

$$\frac{p(\lambda_i;\alpha,\beta,\gamma)}{\pi_{\mathrm{EH}}(\lambda_i;\gamma)} = \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \frac{e^{-\beta/x}}{x^\alpha}\frac{1+\lambda_i}{1+\lambda_i x}\Big\{\frac{1+\log(1+\lambda_i)}{1+\log(1+\lambda_i x)}\Big\}^{1+\gamma}dx$$

To compute the limit at $\lambda_i = \infty$, note that

$$\lim_{\lambda_i\to\infty}\frac{1+\lambda_i}{1+\lambda_i x}\Big\{\frac{1+\log(1+\lambda_i)}{1+\log(1+\lambda_i x)}\Big\}^{1+\gamma} = \frac{1}{x}$$

for each $x > 0$. The result in the main text is verified by the dominated convergence theorem. To see this, evaluate the integrand for $\lambda_i \geq 1$ as

$$\begin{aligned}
\frac{1+\lambda_i}{1+\lambda_i x}\Big\{\frac{1+\log(1+\lambda_i)}{1+\log(1+\lambda_i x)}\Big\}^{1+\gamma} &\leq \frac{2}{x}\exp\Big((1+\gamma)\Big[\log\{1+\log(1+\lambda_i s)\}\Big]_{s=x}^{s=1}\Big) \\
&= \frac{2}{x}\exp\Big\{(1+\gamma)\int_x^1 \frac{1}{s}\frac{\lambda_i s}{1+\lambda_i s}\frac{1}{1+\log(1+\lambda_i s)}ds\Big\} \\
&\leq \begin{cases} \dfrac{2}{x} & \text{if } x \geq 1 \\ \dfrac{2}{x}\exp\Big\{(1+\gamma)\displaystyle\int_x^1 \frac{1}{s}ds\Big\} & \text{if } x < 1 \end{cases} \\
&\leq \frac{2}{x} + \frac{2}{x}\exp\Big\{(1+\gamma)\int_x^1 \frac{1}{s}ds\Big\} = \frac{2}{x}\Big(1+\frac{1}{x^{1+\gamma}}\Big)
\end{aligned}$$

in which we find the bounding function that is integrable as

$$\int_0^\infty \frac{e^{-\beta/x}}{x^\alpha}\frac{2}{x}\Big(1+\frac{1}{x^{1+\gamma}}\Big)dx < \infty$$

for large $\lambda_i > 1$.

6.7.7 Additional simulation results

We here provide additional simulation results under a larger sample size ($m = 400$), where the other settings are the same as ones in the main document. The results are shown in Tables 6.4 and 6.5. We can see that the results are not very different from Tables 6.2 and 6.3 in the main document.

We also assessed the computation time of the proposed methods, IG and EH, and two existing Bayesian methods, GH and PG. Using scenario (I) given in the main document with $\omega = 0.1$, we evaluated the computation time under $m \in \{200, 400\}$. For each m, 3000 posterior samples were generated after discarding the first 500 burn-in samples. The computation time is reported in Table 6.6, where the experiment was performed on a PC with 3.2 GHz 8-Core Intel Xeon W 8 Core Processor with approximately 32GB RAM. From Table 6.6, we can see that the computation time of the proposed methods, IG and EH, are quite comparable with that of PG, and are considerably smaller than that of GH. Moreover, as the number of (local) parameters in the four models linearly increase with m, their computation time would also linearly increase with m, which is partly supported by the results in Table 6.6.

Table 6.4: Averaged values of mean squared errors (MSE) and mean absolute percentage error (MAPE) in non-outlying (-n) and outlying (-o) areas under four scenarios with $m = 400$ and $\omega \in \{0.05, 0.1\}$.

Scenario	ω		IG	EH	GH	PG	KW	ML
(I)	0.05	MSE-n	0.24	0.27	0.42	0.25	0.25	0.40
		MSE-o	3.29	2.93	2.80	3.88	2.88	2.85
		MAPE-n	0.64	0.57	0.65	0.62	0.67	0.61
		MAPE-o	0.21	0.20	0.19	0.23	0.20	0.19
(I)	0.1	MSE-n	0.26	0.28	0.43	0.28	0.27	0.40
		MSE-o	3.05	2.85	2.80	3.02	2.51	2.84
		MAPE-n	0.65	0.59	0.66	0.63	0.68	0.61
		MAPE-o	0.20	0.19	0.19	0.20	0.19	0.19
(II)	0.05	MSE-n	0.21	0.26	0.43	0.23	0.22	0.40
		MSE-o	3.40	2.92	2.77	4.35	2.85	2.80
		MAPE-n	0.59	0.53	0.61	0.57	0.60	0.58
		MAPE-o	0.21	0.20	0.19	0.24	0.20	0.19
(II)	0.1	MSE-n	0.23	0.28	0.43	0.27	0.24	0.40
		MSE-o	3.09	2.83	2.80	3.17	2.46	2.83
		MAPE-n	0.59	0.54	0.62	0.58	0.61	0.58
		MAPE-o	0.20	0.19	0.19	0.20	0.18	0.19
(III)	0.05	MSE-n	0.19	0.25	0.43	0.21	0.17	0.40
		MSE-o	3.56	2.94	2.76	4.87	2.84	2.80
		MAPE-n	0.50	0.48	0.58	0.51	0.49	0.55
		MAPE-o	0.22	0.20	0.19	0.26	0.20	0.19
(III)	0.1	MSE-n	0.21	0.27	0.44	0.26	0.19	0.40
		MSE-o	3.14	2.83	2.80	3.34	2.41	2.82
		MAPE-n	0.52	0.49	0.58	0.53	0.49	0.55
		MAPE-o	0.21	0.19	0.19	0.21	0.18	0.19
(IV)	0.05	MSE-n	0.21	0.26	0.40	0.21	0.20	0.40
		MSE-o	2.38	1.97	2.00	2.67	2.37	2.07
		MAPE-n	2.38	1.65	1.36	2.19	2.14	0.63
		MAPE-o	0.25	0.22	0.22	0.26	0.24	0.22
(IV)	0.1	MSE-n	0.23	0.27	0.42	0.24	0.23	0.40
		MSE-o	2.09	1.91	1.99	2.10	1.90	2.04
		MAPE-n	2.42	1.77	1.45	2.19	2.30	0.63
		MAPE-o	0.23	0.22	0.22	0.23	0.20	0.22

Table 6.5: Coverage probabilities (CP) and average lengths (AL) of 95% credible intervals in non-outlying (n) and outlying (o) areas under four scenarios with $m = 400$ and $\omega \in \{0.05, 0.1\}$.

Scenario	ω		IG	EH	GH	PG	IG	EH	GH	PG
(I)	0.05	n	95.7	96.3	95.7	96.7	1.91	2.00	2.33	1.99
		o	88.4	91.0	94.5	80.7	5.61	5.72	6.25	4.81
	0.1	n	96.1	96.5	95.7	96.6	1.99	2.05	2.34	2.10
		o	90.6	92.0	94.7	88.3	5.76	5.78	6.26	5.19
(II)	0.05	n	95.8	96.4	95.5	96.9	1.88	2.02	2.36	1.98
		o	87.6	90.9	94.6	76.4	5.56	5.67	6.25	4.63
	0.1	n	96.2	96.6	95.5	96.8	1.97	2.07	2.37	2.12
		o	90.1	91.6	94.7	86.8	5.76	5.78	6.29	5.11
(III)	0.05	n	96.2	96.5	95.3	97.3	1.85	2.03	2.40	1.97
		o	85.9	90.4	94.8	70.9	5.49	5.64	6.24	4.45
	0.1	n	96.5	96.6	95.4	97.1	1.95	2.09	2.40	2.12
		o	89.8	91.7	94.8	85.5	5.73	5.75	6.28	5.00
(IV)	0.05	n	93.7	95.6	95.4	95.2	1.88	2.00	2.29	1.92
		o	84.2	90.8	94.5	78.3	4.31	4.73	5.30	3.79
	0.1	n	94.5	95.8	95.5	95.8	1.98	2.05	2.33	2.03
		o	88.2	91.7	94.6	86.5	4.51	4.79	5.33	4.13

Table 6.6: Computation time (seconds) of the four Bayesian methods with $m = 200$ and $m = 400$. In all the methods, 3000 posterior samples were generated after discarding the first 500 samples.

	m	IG	EH	GH	PG
Computation Time	200	2.00	5.49	19.24	1.75
	400	3.92	11.06	38.18	3.65

6.7.8 Metropolis-Hastings method for Poisson regression

The estimation of Poisson regression model in Section 6.5 requires the sampling of regression coefficients $\boldsymbol{\delta}$, in addition to λ_i and other parameters. The new step of sampling $\boldsymbol{\delta}$ is added to the existing MCMC algorithm in Appendix, as described here.

Consider the conditionally independent counts $y_1, \ldots, y_m$ that follow

$$y_i \sim \mathrm{Po}(\lambda_i \eta_i), \qquad \eta_i = \exp\{\boldsymbol{x}_i{}^\top \boldsymbol{\delta}\},$$

where λ_i is a random, individual effect, $\boldsymbol{x}_i$ is the p-vector of covariates and $\boldsymbol{\delta}$ is the coefficient vector. If the likelihood has a known offset term a_i as $y_i \sim \mathrm{Po}(a_i \lambda_i \eta_i)$, then read λ_i in the equation above as $a_i \lambda_i$. We are interested in the posterior analysis of $(\boldsymbol{\lambda}_{1:m}, \boldsymbol{\delta})$ (and the other parameters) by Gibbs sampler. For λ_i, the gamma prior is conditionally conjugate; if $\lambda_i \sim \mathrm{Ga}(\alpha, \beta/u_i)$, then the conditional posterior of λ_i is $\mathrm{Ga}(\alpha + y_i, \beta/u_i + \eta_i)$. With offset a_i, the conditional posterior is $\mathrm{Ga}(\alpha + y_i, \beta/u_i + a_i \eta_i)$. The sampling of the other parameters is not affected by the introduction of regression and offset terms. In this note, we explain the sampling of $\boldsymbol{\delta}$ by MCMC method.

The independent Metropolis-Hastings method can be tailored for the model with conditional posterior density that is analytically available or, at least, numerically evaluated. For the Poisson regression model, we assume the normal prior $\mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ for $\boldsymbol{\delta}$. Conditional on $\boldsymbol{\lambda}_{1:m}$ and current $\boldsymbol{\delta}^{\mathrm{old}}$, we generate the candidate $\boldsymbol{\delta}^{\mathrm{new}}$ from the proposal distribution, which is defined as the posterior distribution derived from the approximate likelihood,

$$\widehat{\boldsymbol{\delta}} \sim \mathrm{N}(\boldsymbol{\delta}, \widehat{\boldsymbol{\Sigma}}),$$

for some known $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\Sigma}}$. Then, denote $\eta_i^{\mathrm{new}} = \exp\{\boldsymbol{x}_i{}^\top \boldsymbol{\delta}^{\mathrm{new}}\}$ and $\eta_i^{\mathrm{old}} = \exp\{\boldsymbol{x}_i{}^\top \boldsymbol{\delta}^{\mathrm{old}}\}$, and accept $\boldsymbol{\delta}^{\mathrm{new}}$ with probability

$$\min \left\{ 1, \prod_{i=1}^{m} \frac{\mathrm{Po}(y_i|\lambda_i \eta_i^{\mathrm{new}}) \mathrm{N}(\widehat{\boldsymbol{\delta}}|\boldsymbol{\delta}^{\mathrm{old}}, \widehat{\boldsymbol{\Sigma}})}{\mathrm{Po}(y_i|\lambda_i \eta_i^{\mathrm{old}}) \mathrm{N}(\widehat{\boldsymbol{\delta}}|\boldsymbol{\delta}^{\mathrm{new}}, \widehat{\boldsymbol{\Sigma}})} \right\} \tag{6.7.12}$$

and set $\boldsymbol{\delta} = \boldsymbol{\delta}^{\mathrm{new}}$. Otherwise, set $\boldsymbol{\delta} = \boldsymbol{\delta}^{\mathrm{old}}$.

The approximate normal likelihood is obtained as the Taylor expansion of the log-likelihood around the mode. The log-likelihood of this model is

$$\begin{aligned} \ell(\boldsymbol{\delta}) &= \sum_{i=1}^{m} \log \left(\frac{\lambda_i^{y_i}}{y_i!} \right) + y_i \log(\eta_i) - \lambda_i \eta_i \\ &= \text{const.} + \sum_{i=1}^{m} y_i (\boldsymbol{x}_i{}^\top \boldsymbol{\delta}) - \lambda_i e^{\boldsymbol{x}_i{}^\top \boldsymbol{\delta}} \end{aligned}$$

The first and second derivatives are

$$\frac{\partial \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{m} y_i \boldsymbol{x}_i - \lambda_i e^{\boldsymbol{x}_i{}^\top \boldsymbol{\delta}} \boldsymbol{x}_i \qquad \text{and} \qquad \frac{\partial \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^\top} = -\sum_{i=1}^{m} \lambda_i e^{\boldsymbol{x}_i{}^\top \boldsymbol{\delta}} \boldsymbol{x}_i \boldsymbol{x}_i{}^\top$$

Then, we obtain $\widehat{\boldsymbol{\delta}}$ as the solution of the first order condition,

$$\frac{\partial \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \mathbf{0}^{(p)}, \qquad \text{i.e.,} \qquad \sum_{i=1}^{m} y_i \boldsymbol{x}_i = \sum_{i=1}^{m} \lambda_i e^{\boldsymbol{x}_i{}^\top \boldsymbol{\delta}} \boldsymbol{x}_i \tag{6.7.13}$$

where $\mathbf{0}^{(p)}$ is the p-vector of zeros. The precision is obtained by

$$\widehat{\boldsymbol{\Sigma}}^{-1} = -\frac{\partial \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^\top}\bigg|_{\boldsymbol{\delta}=\widehat{\boldsymbol{\delta}}} = \sum_{i=1}^{m} \lambda_i e^{\boldsymbol{x}_i{}^\top \widehat{\boldsymbol{\delta}}} \boldsymbol{x}_i \boldsymbol{x}_i{}^\top \tag{6.7.14}$$

The computation of $\widehat{\boldsymbol{\delta}}$ needs the numerical solver of the nonlinear equation above. It should be noted that we *do not have to solve this equation exactly*, for the solution $\widehat{\boldsymbol{\delta}}$ is used to construct the approximate, proposal distribution. The sampling from the proposal is justified by the acceptance-rejection step, no matter what the proposal distribution is used.

In summary, the sampling of $\boldsymbol{\delta}$ takes the following steps. Conditional on $\boldsymbol{\delta}^{\text{old}}$ and the other parameters,

(i) Compute $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\Sigma}}$ by Equations (6.7.13) and (6.7.14).

(ii) Generate $\boldsymbol{\delta}^{\text{new}}$ from the proposal distribution $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = (\widehat{\boldsymbol{\Sigma}}^{-1} + \boldsymbol{\Sigma}_0{}^{-1})^{-1}, \qquad \boldsymbol{\mu} = \boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\delta}} + \boldsymbol{\Sigma}_0{}^{-1}\boldsymbol{\mu}_0).$$

(iii) Set $\boldsymbol{\delta} = \boldsymbol{\delta}^{\text{new}}$ with probability given in Equation (6.7.12). Otherwise, set $\boldsymbol{\delta} = \boldsymbol{\delta}^{\text{old}}$.

Chapter 7

Shrinkage with Robustness: Log-Adjusted Priors for Sparse Signals

7.1 Introduction

Developing new classes of continuous prior distributions that realize the shrinkage effect of variable-selection type on location parameters has been an important research topic in the last few decades, especially in the context of the analysis of high-dimensional datasets to properly express one's prior belief on "few large signals among noises". As pointed out by Carvalho et al. (2009), we can express such belief explicitly via the parameterization of shrinkage effect in the Bayes estimator that shrinks the observed signals to zero or baseline. This parametrization opens the path to crafting the new class of continuous priors that mimic the discrete mixture for variable selection, namely, the spike-and-slab priors (Ishwaran and Rao (2015)), which is more desirable in the high-dimensional context than the existing shrinkage priors (e.g. Strawderman (1971); Berger (1980); Park and Casella (2008)). In addition to shrinking the negligible noises toward zero, the desirable prior here should also define the Bayes estimator that is robust to outlying large signals in the sense that such signals are kept unshrunk in the posterior analysis. This property is typically called tail-robustness (e.g. Carvalho et al. (2010)), and the aim of this research is to define a new class of shrinkage priors with strong tail-robustness.

The aforementioned parametrization describes both shrinkage effect and tail robustness implicitly assumed in the prior of interest. Suppose we observe $y_i \sim \mathrm{N}(\theta_i, 1)$ independently for $i = 1, \ldots, n$ and the prior is given by $\theta_i \sim \mathrm{N}(0, \tau u_i)$ (and $\tau = 1$, for simplicity) and $u_i \sim \pi(u_i)$. Then, the Bayes estimator of true signal θ_i is written as $(1 - E[\kappa_i|y_i])y_i$, where $\kappa_i = 1/(1 + u_i)$. It is this parameter, κ_i, that controls the amount of shrinkage in the Bayes estimator. In the presence of sparse signals, the standard choice of priors for κ_i's has been the beta distribution (Armagan et al. (2011); Pérez et al. (2017)), originated from the half-Cauchy distribution (Gelman (2006), or horseshoe prior; Carvalho et al. (2009, 2010)) given by $\pi(\kappa_i) \propto \kappa_i^{b-1}(1 - \kappa_i)^{a-1}$, $\kappa_i \in (0, 1)$, with positive a and b. The appropriate modeling of shrinkage and robustness is then translated into the choice of extremely small shape parameters (a, b). This preference on the choice of hyperparameters is, however, against the finding of Bai and Ghosh (2019); in order to guarantee the desirable posterior concentration for both small and large signals, a can be ex-

tremely small ($a = 1/n$) but b must be sufficiently large ($b \geq 1/2$), which clarifies the limitation of the class of beta distributions.

In this research, we consider the extension of beta-type shrinkage priors to strengthen the prior tail-robustness. Specifically, we propose the following modified version of the beta prior:

$$\pi(\kappa_i) \propto \kappa_i^{b-1}(1-\kappa_i)^{a-1}\left(1-\log\kappa_i\right)^{-(1+\gamma)}, \tag{7.1.1}$$

where $\gamma > 0$ is a newly introduced hyperparameter. The use of logarithm in the density slightly "slows down" the divergence of the density as $\kappa_i \downarrow 0$ and, in fact, makes the density kernel above integrable even if $b = 0$, as shown in Theorem 7.2.1. This distribution allows the stronger tail-robustness than the beta prior by setting $b = 0$, while remaining in the class of proper priors.

The use of logarithm term in the density function to define the new class of distributions has motivated many research on posterior inference. They include the analysis of ultra-sparse signals (Bhadra et al. (2017)), robust regression (Gagnon et al. (2020)), and admissibility (Maruyama and Strawderman (2020a)). The shrinkage with robustness– our research goal– has also been considered in Womack and Yang (2019) as the heavy-tailed extension of the horseshoe prior. A similar log-adjusted method was employed in Hamura et al. (2020a) for the analysis of high-dimensional counts. The key difference of our research from the listed literature is the focus on the robustness of Bayes estimator; we proved the superiority of the proposed prior to existing ones explicitly via improvement of the mean squared error for large y_i, as summarized in Theorem 7.2.2, and support this theoretical property by the extensive simulation study in Section 7.3.

Among the variants of probability distributions that involve the log-term, the novel feature of the prior of our interest is its potential of further generalization, by which one may modify the proper prior "as robust as possible". Although the prior in (7.1.1) becomes improper with $\gamma = 0$, we can multiply another log-term as

$$\pi(\kappa_i) \propto \kappa_i^{b-1}(1-\kappa_i)^{a-1}\left(1-\log\kappa_i\right)^{-1}\left\{1+\log(1-\log\kappa_i)\right\}^{-(1+\gamma)},$$

which is proper again even if $b = 0$ as long as $\gamma > 0$. Notably, we can repeatedly iterate this process of extension; if $\gamma = 0$ in the above equation and the density becomes improper, then the reciprocal of another log-term, $1 + \log\{1 + \log(1 - \log\kappa_i)\}$, can be multiplied to the density to regain the proper prior. It is expected, and verified later in Theorem 7.2.4, that such extension provides the stronger tail-robustness and makes the choice of γ less sensitive to the posterior analysis. Furthermore, as discussed in Sections 7.4 and 7.5.8, the limit of repeated extensions by log-terms is the discrete mixture of two point masses, i.e., the spike-and-slab prior, in the sense of convergence in distribution. This result on the limiting distribution could justify the proposed prior as the continuous alternative to the ideal, but inefficient or infeasible in computation, discrete priors of variable selection type.

The Bayes estimator under the proposed priors has no closed form, even if global scale τ is fixed, due to the intractable normalizing constant. Yet, the estimator can be evaluated fast by simulation. The proposed prior density admits the integral representation, or the augmentation by latent variables that follow gamma-shape Markov processes, by which the full conditional posteriors of those parameters and latent variables become normal, (inverse) gamma or generalized inverse Gaussian distributions. Sampling from those distributions is trivial, and the full posterior analysis becomes available by the simple but efficient Gibbs sampler.

The rest of this chapter is organized as follows. In Section 7.2, we define the log-adjusted shrinkage prior and its extension, and provide the theoretical properties, the improvement of the

mean squared errors of Bayes estimators, and the Gibbs sampler by augmentation. Simulation studies and data analysis follow in Section 7.3, with the extensive comparative analysis with the existing shrinkage priors. We conclude this chapter in Section 7.4 with the further discussion on the limiting distribution of repeated extensions by multiplying the iterated log-terms for both shrinkage and robustness.

All proofs and technical details are given in the Appendix.

7.2 Log-Adjusted Shrinkage Priors

7.2.1 The proposed prior and its properties

Suppose we observe an n-dimensional vector $(y_1, \ldots, y_n)$, that $y_i|\theta_i \sim \mathrm{N}(\theta_i, 1)$ independently for $i = 1, \ldots, n$. To estimate signals $(\theta_1, \ldots, \theta_n)$ that are potentially sparse, we adopt locally adaptive shrinkage priors known as global-local shrinkage priors (Polson and Scott (2012a, 2012b); Bhadra et al. (2016)) given by

$$\theta_i|\tau, u_i \sim \mathrm{N}(0, \tau u_i) \quad \text{and} \quad u_i \sim \pi(u_i), \quad \text{for } i = 1, \ldots, n, \tag{7.2.1}$$

where both τ and $(u_1, \ldots, u_n)$ are all positive. Here, τ is the global shrinkage parameter that shrinks all θ_i's toward zero uniformly, while u_i is the local scale parameters and customizes the shrinkage effect for each individual i. For simplicity, we assume $\tau = 1$ to focus our theoretical development on the priors for local scale parameters. We propose the following modified version of the scaled beta distribution:

$$\pi(u_i) = C(a, b, \gamma)^{-1} u_i^{a-1} (1 + u_i)^{-(a+b)} \left\{1 + \log(1 + u_i)\right\}^{-(1+\gamma)}, \tag{7.2.2}$$

where $C(a, b, \gamma)$ is a normalizing constant. Note that the class of distributions defined by density (7.2.2) includes the scaled beta distributions (Armagan et al. (2011)) as the density of $\gamma = -1$ and positive a and b. The hyperparameters, (a, b, γ), determine the functional form of the density around the origin and in the tails. Shape parameters a and b control shrinkage effect and tail robustness for Bayes estimators, respectively, and both parameters should be set to small values in order to achieve the desirable shrinkage and robustness properties. Specifically, we set $a = 1/n$, following Bai and Ghosh (2019), to realize the strong shrinkage effect on noises and set $b = 0$ to attain the strong tail robustness. Note again that setting $b = 0$ in the original scaled beta distribution leads to an improper prior, thereby it cannot be adopted as shrinkage priors in practice. The new parameter γ also affects the tail behavior of the density as b does, but it would have less impact on posterior analysis. We may either fix γ subjectively to a certain value, such as $\gamma = 1$, or take the fully Bayesian approach by considering the prior for γ as we discuss in the subsequent section. The new priors for θ_i under (7.2.1) with the log-adjusted scaled beta distribution (7.2.2) is named *log-adjusted shrinkage priors*. In what follows, we demonstrate properties of the proposed prior with general hyperparameters, (a, b, γ), but the priors of our interest and recommendation are those with $a = 1/n$ and $b = 0$.

We first provide important properties of the proposed shrinkage prior for θ_i in the theorem below.

Theorem 7.2.1 *The log-adjusted shrinkage prior for θ_i, $\pi(\theta_i)$, satisfies the following properties.*

1. *$\pi(\theta_i)$ is proper if $a > 0$, $b \geq 0$ and $\gamma > 0$.*

2. $\lim_{|\theta_i|\to 0} \pi(\theta_i) = \infty$ *for* $a \leq 1/2$.

3. $\pi(\theta_i) \propto |\theta_i|^{-2b-1} L(|\theta_i|)$ *under* $|\theta_i| \to \infty$, *where* $L(\cdot)$ *is a slowly varying function satisfying* $\lim_{M\to\infty} L(Mu)/L(M) = 1$ *for all* $u > 0$.

It is notable that the prior is proper even if $b = 0$ from the first property; this is obviously due to the additional log-term in (7.2.2). The second property is the same as that of the original beta prior, which indicates that the proposed prior has the density with the spike around the origin and holds strong shrinkage property. It also means that the additional log-term does not change the shrinkage property of the original beta-type prior. From the third property, our proposal of setting $b = 0$ results in the extremely heavy tailed prior distribution for θ_i, whose tail is heavier than even the Cauchy distribution. Such heavy-tailed properties are essential for strong-tail robustness, as shown in Theorem 7.2.2.

Figure 7.1 shows the examples of the log-adjusted beta distribution given in (7.1.1), in the scale of $\kappa_i = 1/(1 + u_i)$, for different choices of hyperparameter γ. When compared with beta density Beta$(1/2, 1/2)$, which is the half-Cauchy distribution in the scale of u_i and realizes the horseshoe prior, the log-adjusted shrinkage densities have steeper spike as $\kappa_i \to 0$, reflecting its tail property introduced by the additional log-term. The shrinkage effect is also affected by this additional term in the density, but the densities with moderate values of γ, such as $\gamma = 0.5$ or $\gamma = 1$, show the similar speed of divergence toward κ_i as the beta density does. These observations imply that, with the appropriate choice of hyperparaemters, the log-adjusted shrinkage prior can introduce the strong tail-robustness without losing the horseshoe-type shrinkage effect.

In order to clarify the importance of setting $b = 0$, we next examine the posterior tail-robustness under the proposed prior by computing the posterior mean squared error for large y_i. Denote the posterior mean squared error under a prior $\pi(\theta_i)$ by $\mathrm{MSE}_\pi(\theta_i|y_i) = E_\pi[(\theta_i - y_i)^2|y_i]$. We evaluate the mean squared error of the proposed class of priors in the following theorem based on the representation of posterior mean squared error by the marginal likelihood (e.g., see Polson (1991)).

Theorem 7.2.2 *Under the log-adjusted shrinkage priors for* θ_i *with hyperparameters* $a > 0, b \geq 0$ *and* $\gamma > 0$, *and under the beta-type prior (*$a > 0$, $b > 0$ *and* $\gamma = -1$*), it holds that*

$$\mathrm{MSE}_\pi(\theta_i|y_i) = 1 + \frac{2}{y_i^2}(1 + b)(1 + 2b) + o\left(\frac{1}{y_i^2}\right), \tag{7.2.3}$$

where ${y_i}^2 o(1/{y_i}^2) \to 0$ *as* $|y_i| \to \infty$.

We first note that the above approximation formula of mean squared error is independent of γ, thereby the same formula holds for the original beta-type shrinkage prior for θ_i. Moreover, Theorem 7.2.2 shows that the leading term of the posterior mean squared error for large signals is increasing in b, which clearly suggests that the best choice is $b = 0$. The proper log-adjusted shrinkage prior can attain the ideal mean squared error by setting $b = 0$, outperforming in the mean squared error the proper beta-type shrinkage prior for which we always have to set $b > 0$.

The posterior mean squared error for large y_i has also been calculated for other shrinkage priors. Theorem 7 of Bhadra et al. (2017) provided the posterior mean squared error of the horseshoe+ prior $\pi_{\mathrm{HS}+}$ as

$$\mathrm{MSE}_{\pi_{\mathrm{HS}+}}(\theta_i|y_i) = 1 + 3\left(\frac{2}{{y_i}^2}\right) + o\left(\frac{1}{{y_i}^2}\right),$$

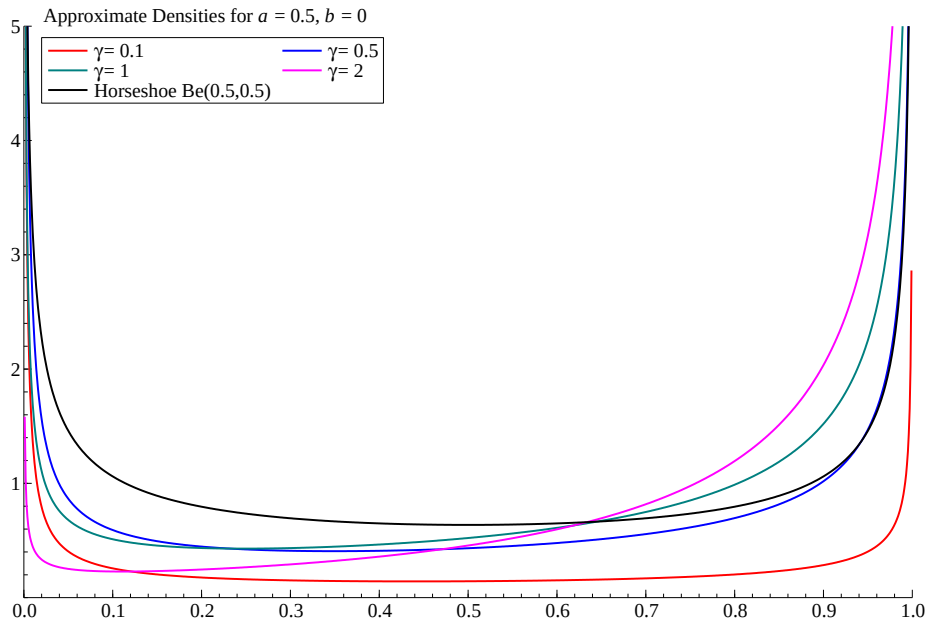

Figure 7.1: The prior density for the shrinkage factor κ_i given in (7.1.1) under the log-adjusted shrinkage prior with $a = 1/2$ and $b = 0$. The four densities represent the cases of $\gamma = 0.1$ (red), 0.5 (blue), 1 (green) and 2 (pink). The density of beta distribution Beta$(1/2, 1/2)$ (black) is equivalent to the half-Cauchy prior in the scale of u_i and realizes the horseshoe prior. The log-adjusted shrinkage priors have the steep increase toward $\kappa_i = 0$ for tail-robustness, while maintaining its spike around $\kappa_i = 1$ for strong shrinkage if we choose moderate values of γ, such as $\gamma = 0.5$ or $\gamma = 1$.

which is, as $|y_i| \to \infty$, inevitably larger than the posterior mean squared error of our proposed prior in (7.2.3) with $b = 0$.

7.2.2 Posterior computation

Although the Bayes estimator of θ_i for the log-adjusted shrinkage prior is not analytically available, there is an efficient yet simple Markov chain Monte Carlo algorithm for posterior computation. The full conditional posteriors of parameters, θ_i's, u_i's and τ, become well-known distributions after the appropriate augmentation by latent variables described below. The conditional posterior density of hyperparameter γ is complex due to the intractable normalizing constant $C(a, b, \gamma)$, but the sampling from its distribution is feasible by the accept-reject algorithm. The rest hyperparameters are fixed as $a = 1/n$ and $b = 0$.

The prior density of u_i in (7.2.2) has the following augmented expression:

$$\pi(u_i; \gamma) = \frac{1}{C(a, b, \gamma)} \int_0^\infty \int_0^\infty u_i^{a-1} \frac{v_i^\gamma e^{-v_i}}{\Gamma(1+\gamma)} \frac{w_i^{v_i+a+b-1} e^{-w_i(1+u_i)}}{\Gamma(v_i+a+b)} dw_i dv_i, \tag{7.2.4}$$

where w_i and v_i are latent variables for data augmentation. Given γ, the augmented posterior distribution is proportional to

$$\pi(\tau) \prod_{i=1}^n \mathrm{N}(y_i|\theta_i, 1)\mathrm{N}(\theta_i|0, \tau u_i) \frac{1}{C(a, b, \gamma)} u_i^{a-1} v_i^\gamma e^{-v_i} \frac{w_i^{v_i+a+b-1} e^{-w_i(1+u_i)}}{\Gamma(v_i+a+b)},$$

where $\pi(\tau)$ denotes a prior distribution of τ. We assign an inverse-gamma prior for τ and set $\pi(\tau) = \mathrm{IG}(\tau|c_0^\tau, d_0^\tau)$, which leads to conditional conjugacy. It is immediate from the expression above that all the full conditional distributions are of normal or (inverse) gamma distributions, so that we can efficiently carry out Gibbs sampling by generating posterior samples from those distributions. The procedure of Gibbs sampler is summarized as follows:

Algorithm 7.2.1 (Gibbs sampling algorithm) *Suppose γ is fixed. The Gibbs sampler algorithm, or the list of the full conditional distributions of the local and global parameters under the log-adjusted shrinkage prior, is summarized as follows:*

- *Generate τ from* $\mathrm{IG}(n/2 + c_0^\tau, \sum_{i=1}^n {\theta_i}^2/(2u_i) + d_0^\tau)$.
- *Generate θ_i from* $\mathrm{N}(y_i/\{1 + 1/(\tau u_i)\}, 1/\{1 + 1/(\tau u_i)\})$ *for* $i = 1, \ldots, n$.
- *Generate u_i from the generalized inverse Gaussian full conditional distribution* $p(u_i|\theta_i, w_i, \tau) \propto {u_i}^{-1/2+a-1} \exp[-\{2w_i u_i + ({\theta_i}^2/\tau)/u_i\}/2]$ *for* $i = 1, \ldots, n$.
- *Generate (v_i, w_i) by the following two steps:*
  - *Generate v_i from the conditional distribution marginalized over w_i, namely $p(v_i|\theta_i, \tau, y_i)$, which is* $\mathrm{Ga}(1 + \gamma, 1 + \log(1 + u_i))$, *for* $i = 1, \ldots, n$.
  - *Generate w_i from the full conditional distribution, $p(w_i|v_i, \theta_i, \tau, y_i)$, which is* $\mathrm{Ga}(v_i + a + b, 1 + u_i)$ *for* $i = 1, \ldots, n$.

We next consider incorporating the estimation of γ into the algorithm above. Let $C(\gamma) = C(a, b, \gamma)$ be the normalizing constant of $\pi(u_i)$ in (7.2.2), which is defined by

$$C(\gamma) = \int_0^\infty \frac{u^{a-1}(1+u)^{-(a+b)}}{\{1+\log(1+u)\}^{1+\gamma}} du = \int_0^1 \frac{(1-\kappa)^{a-1}\kappa^{b-1}}{(1-\log\kappa)^{1+\gamma}} d\kappa, \tag{7.2.5}$$

where $\kappa = 1/(1+u)$, $a = 1/n$ and $b = 0$. Due to this intractable normalizing constant in the prior, the direct sampling from the full conditional of γ is challenging and even infeasible. We circumvent this problem by constructing upper and lower bounds for the normalizing constant, which allows for the independent Metropolis-Hastings algorithm by providing the bounds of the acceptance probability with arbitrary accuracy. This approach is similar to the alternating series method (Devroye (1981, 2009)) and widely used in, for example, the sampling from the Polya-gamma distribution (Polson et al. (2013)). Our sampling procedures are briefly sketched in the following. Details of the bounds of the acceptance probability, $\overline{w}$ and $\underline{w}$, and some remarks about the sampling procedures are provided in Section 7.5.3.

We use the gamma prior $\mathrm{Ga}(a_0^\gamma, b_0^\gamma)$ for γ. The proposal distribution is $\mathrm{Ga}(a_1^\gamma, b_1^\gamma)$, whose parameters are given by

$$a_1^\gamma = a_0^\gamma + na \quad \text{and} \quad b_1^\gamma = b_0^\gamma + \sum_{i=1}^n \log\{1+\log(1+u_i)\}.$$

Denote the current state at an iteration of the Markov Chain Monte Carlo algorithm by γ, and the candidate drawn from the proposal by γ'. The acceptance probability $A(\gamma \to \gamma')$ is bounded below and above by $\underline{w}$ and $\overline{w}$. Both are functions of (a, γ, γ', K) and converge to $A(\gamma \to \gamma')$ as $K \to \infty$, where K controls the precision of the approximation and can be set as large as necessary. The procedure of the independent Metropolis-Hastings sampling is summarized as follows.

Algorithm 7.2.2 (Sampling from the full conditional of γ) *The steps for generating γ from its full conditional distribution is summarized as follows: given the current sample γ,*

(i) *Generate γ' from the proposal* $\mathrm{Ga}(a_1^\gamma, b_1^\gamma)$.

(ii) *Generate* $U \sim \mathrm{U}(0, 1)$.

(iii) *Given K, evaluate $\underline{w}$ and $\overline{w}$. Then,*

- *If $U < \underline{w}$, accept γ' as the sample of this iteration.*
- *If $U > \overline{w}$, reject γ' and keep γ as the sample of this iteration.*
- *Otherwise ($\underline{w} < U < \overline{w}$), increase K and redo step 3.*

7.2.3 Generalization using iterated logarithm

Following the motivation given in the introduction, the log-adjusted shrinkage prior is further extended to the more general class of distributions. As the (scaled) beta distributions is extended to the log-adjusted version by the multiplicative log-term, this generalization is naturally realized by the use of finitely iterated logarithmic functions.

For $z \geq 1$, let $f_1(z) \equiv f(z) \equiv 1 + \log(z)$. Then, the iterated logarithm is defined inductively by $f_{L+1}(z) \equiv f(f_L(z))$ for $L = 1, 2, \ldots$. Define the extended version of the modified scaled beta priors with parameter $\gamma > 0$ by

$$\pi(u_i; \gamma, L) \propto \frac{u_i^{a-1}}{(1+u_i)^{a+b}} \left\{ \prod_{k=1}^{L-1} \frac{1}{f_k(1+u_i)} \right\} \frac{1}{f_L(1+u_i)^{1+\gamma}}, \tag{7.2.6}$$

where $a > 0$ and $b \geq 0$ are constant (We again recommend $a = 1/n$ and $b = 0$). The corresponding prior for the shrinkage factor $\kappa_i = (1+u_i)^{-1}$ is given by

$$\pi(\kappa_i; \gamma, L) \propto \kappa_i^{b-1}(1-\kappa_i)^{a-1} \left\{ \prod_{k=1}^{L-1} \frac{1}{f_k(1/\kappa_i)} \right\} \frac{1}{f_L(1/\kappa_i)^{1+\gamma}}.$$

The prior for θ_i induced by this distribution as the scale mixture of normal is named *iteratively log-adjusted shrinkage prior*.

When $L = 1$, this prior is precisely the original log-adjusted shrinkage prior discussed in the previous subsections. We require that $b = 0$ for the improvement from the beta-type prior, but the priors are still proper if only $\gamma > 0$, as shown in the following theorem. As order L increases, the tails of the density for θ_i becomes heavier, while remaining in the class of proper priors, from which we expect the stronger tail-robustness of the Bayes estimators.

Theorem 7.2.3 *The following properties hold under the iterative log-adjustment.*

1. *The iteratively log-adjusted shrinkage prior for θ_i with finite L holds the same properties given in Theorem 7.2.1.*

2. *Suppose that $b = 0$. Then, for any $0 < \varepsilon < 1$, the prior probability of the log-adjusted beta distribution for κ_i falling in the interval $(0, \varepsilon)$ tends to 1 as $L \to \infty$; namely*

$$\lim_{L \to \infty} \int_0^{\varepsilon} \pi(\kappa_i; \gamma, L) d\kappa_i = 1.$$

The first property indicates that the iterative log-adjustments do not change the original properties of the proposed prior, including integrability, density spike around the origin and heavier tails. The second statement shows the convergence of the iterated log-adjusted shrinkage prior to the point mass on $\kappa_i = 0$ in distribution as $L \to \infty$. In the limit, the proposed prior does not shrink the outliers at all. However, losing the shrinkage effect at all is not desirable, and we fix L at some finite value so that the prior density keeps the steep spike around zero.

Although it is difficult to draw the density functions of the iteratively log-adjusted shrinkage priors as in Figure 7.1 for the intractable normalizing constant, the newly-multiplied log-terms can easily be evaluated and shown in Figure 7.2. It is clear in the top figure that function $f_L(1+u)$ is increasing in u, but converges to the constant function as $L \to \infty$, which are also verified in Section 7.5.4. This observation implies that the marginal effect of log-terms being multiplied to the prior is diminishing as L increases. The bottom figures displays the reciprocal of the iterative log-terms in the scale of κ_i, which are actually multiplied to the original log-adjusted shrinkage prior. The lower densities near $\kappa_i = 0$ moderates the spike and makes the density integrable, while the iterative log-term is unity around $\kappa_i = 1$ and affects the shrinkage effect less.

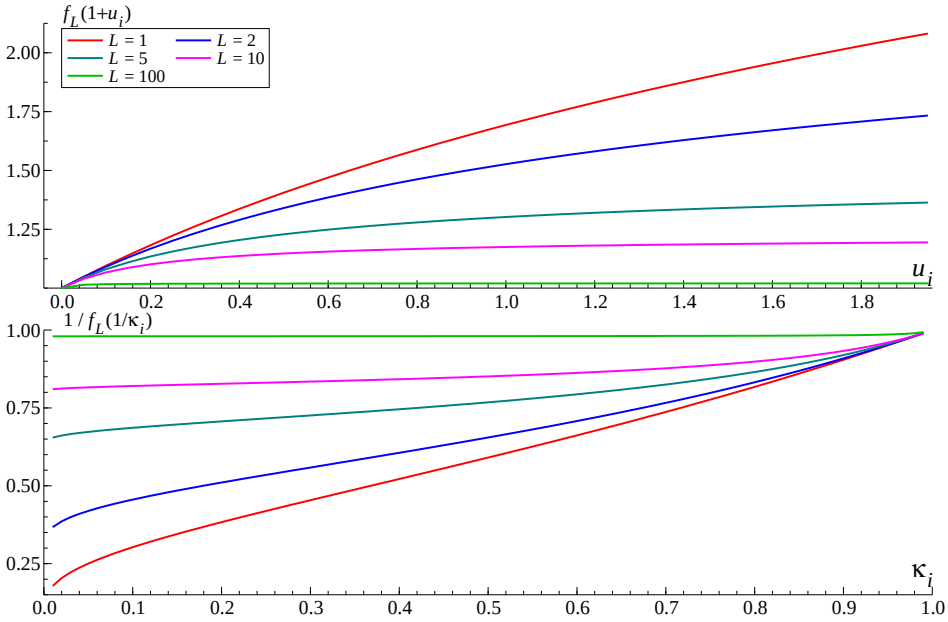

Figure 7.2: The functions $f_L(1+u_i)$ (top) and $1/f_L(1/\kappa_i)$ (bottom) with $L = 1$ (red), 2 (blue), 5 (green), 10 (pink) and 100 (light green). It is evident that the repeated application of operation f makes the function closer to constant. In the bottom figure, the decrease of functions as $\kappa_i \to 0$ moderates the divergence of the prior density around zero and contributes to the integrability of the iteratively log-adjusted shrinkage priors.

The posterior mean squared error under the iteratively log-adjusted shrinkage prior can also be computed in the similar way as in the proof of Theorem 7.2.2.

Theorem 7.2.4 *The posterior mean squared error under the iteratively log-adjusted shrinkage prior satisfies, for $b = 0$,*

$$\mathrm{MSE}_\pi(\theta_i|y_i) = 1 + \frac{1}{{y_i}^2/2}\Big\{1 - \frac{3a/2}{{y_i}^2/2} + \sum_{k=1}^{L-1} \frac{1}{f_k(1+{y_i}^2/2)} \cdots \frac{1}{f_1(1+{y_i}^2/2)} + (1+\gamma)\frac{1}{f_L(1+{y_i}^2/2)} \cdots \frac{1}{f_1(1+{y_i}^2/2)}\Big\} + o\Big(\frac{1}{{y_i}^2}\Big), \tag{7.2.7}$$

where ${y_i}^2 o(1/{y_i}^2) \to 0$ as $|y_i| \to \infty$.

Theorem 7.2.4 derives the higher-order terms of y_i that is ignored in the mean squared error under the original log-adjusted shrinkage prior in Theorem 7.2.2, whose leading term is simply $1 + 2/y_i^2$. We summarize our findings on the derived mean squared error in the following three points. First, this result reveals that the effect of hyperparameters a and γ is limited to the higher-order terms, which is consistent with the result of Theorem 7.2.2. In addition, the choice of hyperparameter γ is less sensitive to the Bayes estimator if L is large. Secondly, there is no difference in the mean squared errors under the original and iteratively log-adjusted shrinkage priors in the order of $2/y_i^2$, while both priors are still superior to the beta-type shrinkage priors in the mean squared error in the tail. Finally, increasing the order L has no negative effect on the point estimation so long as the posterior mean squared error under large signals is concerned. In fact, it is difficult to understand whether one increment of L increases or decreases the mean squared error in this expression; it is determined together with the values of y_i and γ. We revisit this issue partially in the simulation studies in Section 7.3.

The posterior computation with the iteratively log-adjusted shrinkage prior is also straightforward. The parameter augmentation given in (7.2.4) can be generalized for the new density of u_i in (7.2.6) as, ignoring the normalizing constant,

$$\pi(u_i;\gamma,L) \propto \int_{(0,\infty)^{L+1}} u_i^{a-1} e^{-t_{iL}u_i} \mathrm{GS}_L(\boldsymbol{t}_{i,0:L}|\gamma) d\boldsymbol{t}_{i,0:L}, \tag{7.2.8}$$

where $\boldsymbol{t}_{i,0:L} = (t_{i0}, t_{i1}, \ldots, t_{iL})$ and $\mathrm{GS}_L(\boldsymbol{t}_{i,0:L}|\gamma)$ is the joint density of a non-stationary Markov process defined by

$$\mathrm{GS}_L(\boldsymbol{t}_{i,0:L}|\gamma) = \mathrm{Ga}(t_{i0}|1+\gamma,1)\mathrm{Ga}(t_{i1}|t_{i0}+1,1) \times \cdots \times \mathrm{Ga}(t_{i,L-1}|t_{i,L-2}+1,1)\mathrm{Ga}(t_{iL}|t_{i,L-1}+a+b,1).$$

The density of u_i is the shape mixture of density kernel of gamma distribution by a gamma-shape Markov process. The integral expression above defines latent variables t_{il}'s and gives the following tractable full conditional distributions for Gibbs sampler.

Algorithm 7.2.3 (Gibbs sampler for local parameters under ILAS prior) *The sampling steps for local parameters u_i and $\boldsymbol{t}_{0:L}$ are summarized as follows:*

- *The full conditional distribution of u_i is* $\mathrm{GIG}(-1/2+a, 2t_{iL}, \theta_i^2/\tau)$.

- *The full conditional distribution of* $\boldsymbol{t}_{0:L}$ *has the compositional form,*

$$\begin{aligned}&\mathrm{Ga}(t_{i,0}|1+\gamma, f_L(1+u_i))\mathrm{Ga}(t_{i,1}|t_{i,0}+1, f_{L-1}(1+u_i)) \times \cdots \\ &\quad \times \mathrm{Ga}(t_{i,L-1}|t_{i,L-2}+1, f_1(1+u_i))\mathrm{Ga}(t_{i,L}|t_{i,L-1}+a+b, 1+u_i),\end{aligned}$$

  thereby the random samples can be sequentially generated.

The above procedure can be incorporated into Algorithm 7.2.1, which enables us to efficiently generate posterior samples of θ_i. It is worth noting that $t_{i,k}$, $k \leq L-1$, are not used to generate samples of $\{u_i, \theta_i, \tau\}$.

The shrinkage priors with logarithm terms in their densities have been studied in various ways. We considered the prior distributions proposed in Bhadra et al. (2017) and Womack and Yang (2019) and confirmed that their prior densities could be extended in a similar way by repeatedly multiplying the additional terms to the density function. However, such iterative operation is extremely complex, compared with the simple recursive construction of the additional terms in this research, $f_{L+1}(z) = 1 + \log f_L(z)$, that defines the log-adjusted shrinkage priors.

7.3 Numerical Study

7.3.1 Simulation study

We illustrate finite-sample performance of the Bayes estimators under the proposed priors and other shrinkage priors proposed in the recent research in various situations of true sparse signals. We generated $n = 200$ observations from $y_i \sim \mathrm{N}(\theta_i, 1)$, where θ_i is a true signal. We adopted the following two scenarios for θ_i:

$$\begin{aligned}&\text{(I)} \quad \theta_i \sim \frac{\omega}{2}\delta(c) + \frac{\omega}{2}\delta\left(-\frac{c}{2}\right) + (1-\omega)\,\delta(0), \\ &\text{(II)} \quad \theta_i \sim \frac{\omega}{2}\mathrm{N}(c,1) + \frac{\omega}{2}\mathrm{N}\left(-\frac{c}{2}, 1\right) + (1-\omega)\,\delta(0),\end{aligned}$$

where $\delta(x)$ denotes the one-point distribution on x. Weight ω controls the sparsity level in the signals θ_i; smaller value of ω leads to more sparsity. c is the locations of non-null signals. We considered six settings of ω and c as the combinations of $\omega \in \{0.1, 0.2, 0.3\}$ and $c \in \{6, 9\}$.

For the simulated dataset, we applied three types of proposed priors: the log-adjusted shrinkage prior with $a = 1/n, b = 0$ and $\gamma = 1$ (denoted by LAS), an adaptive version of the log-adjusted shrinkage prior with a fully Bayesian approach for γ (denoted by aLAS), and the iteratively log-adjusted shrinkage prior with $a = 1/n, b = 0, \gamma = 1$ and $L = 3$, denoted by ILAS. As competitors, we also applied the Horseshoe prior (HS; Carvalho et al. (2010)), normal-beta prime prior (NBP; Bai and Ghosh (2019)), Dirichlet-Laplace prior (DL; Bhattacharya et al. (2015)), Horseshoe+ prior (HS+; Bhadra et al. (2017)). To implement the posterior analysis with the Horseshoe+ and Dirichlet-Laplace priors, we employed R package "NormalBetaPrime" (Bai and Ghosh (2019, 2020)) with default settings, such as the use of uniform prior on $(1/n, 1)$ for the global scale parameter. For the other models, we adopted $\tau \sim \mathrm{C}^+(0, 1/n)$. In applying all the priors, we generated 1000 posterior samples after discarding the first 1000 samples as burn-in period, and computed posterior means of θ_i. The squared error losses of the posterior means $\hat{\theta}_i$, give by $\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2$, were calculated and averaged over 500 replications of simulations.

The results are reported in Table 7.1. It shows that the methods are comparable when ω is small and the true signals are very sparse. On the other hand, as ω increases, the proposed

priors get appealing compared with the other methods. This result is consistent with Theorems 7.2.2 and 7.2.4 and reflects the fact that the proposed priors have heavier tails to accommodate large signals. It is also observed that the proposed three methods are almost equally successful and it is difficult to discuss their superiority. Focusing the comparison on the performance of LAS and aLAS, the benefit from estimating the adjustment parameter γ could be limited possibly because of the trade-off between flexibility of data-adaptive selection of γ and inflation of uncertainty arising from estimating γ. Finally, LAS and ILAS perform quite similarly in every setting, which could be related to the fact that these two priors differ only in the form of the slowly varying part. It can also be explained by their mean squared errors in the tails that are exactly the same up to the order of ${y_i}^2$.

Table 7.1: Comparison of averaged values of squared error losses of the posterior mean estimates of θ_i under the fixed log-adjusted shrinkage (LAS) prior, the adaptive LAS prior (aLAS), the iteratively log-adjusted shrinkage prior (ILAS) of order three (EH-IL), the Horseshoe prior (HS), the normal-beta prime prior (NBP), the Dirichlet-Laplace prior (DL), and the Horseshoe+ prior (HS+). The lowest averaged squared error loss for each setting (in rows) is in bold.

c	omega	LAS	aLAS	ILAS	HS	NBP	DL	HS+
	0.1	**52.5**	55.0	52.5	55.9	55.7	62.7	54.2
6	0.2	90.6	**89.9**	90.7	96.6	107.6	120.5	102.7
	0.3	124.3	**121.3**	124.2	126.4	159.5	177.9	150.5
	0.1	43.3	47.9	43.0	49.4	41.7	50.9	**41.4**
9	0.2	**68.6**	71.5	**68.6**	87.7	76.2	95.5	73.7
	0.3	**92.5**	94.3	**92.5**	117.8	110.4	137.5	105.3
	0.1	47.9	51.1	**47.6**	50.9	49.3	56.2	48.2
6	0.2	**84.3**	84.7	84.4	92.2	98.4	112.4	94.1
	0.3	113.5	**111.8**	113.7	120.5	143.1	164.8	135.7
	0.1	43.9	48.6	43.7	49.2	41.7	49.2	**41.6**
9	0.2	**71.6**	74.5	**71.6**	88.6	78.3	93.6	76.0
	0.3	**96.0**	97.7	96.3	119.0	113.2	136.2	108.6

7.3.2 Example: Prostate cancer data

We demonstrate real-data application of the proposed priors using a popular prostate cancer dataset in Singh et al. (2002). In this dataset, there are gene expression values for $n = 6033$ genes for $m = 102$ subjects, with $m_1 = 50$ normal control subjects and $m_2 = 52$ prostate cancer patients. The goal of this analysis is to identify genes that are significantly different between the two groups. We first conduct t-test for each gene to compute the test statistics $t_1, \ldots, t_n$, and then transform them to z-scores through $z_i = \Phi^{-1}(F^t_{m-2}(t_i))$, where $\Phi(\cdot)$ is the standard normal distribution function and $F^t_k(\cdot)$ is the distribution function of t-distribution with k degrees of freedom. For the resulting z-scores, $z_1, \ldots, z_n$, we applied the following model:

$$z_i = \theta_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, 1), \quad i = 1, \ldots, n.$$

We again compare the same seven priors for θ_i as in the previous subsection. Based on 5000 posterior samples after discarding the first 5000 samples, we computed posterior means of θ_i. In Table 7.2, we presented top 10 genes selected by Efron (2010) and their estimated effect size θ_i on prostate cancer. The absolute value of each effective size estimate is largest for aLAS, because of its strong tail-robustness. However, the estimates of all the seven methods do not differ drastically from one another.

Table 7.2: The z-scores and the effect size estimates based on posterior means for the top 10 genes selected by Efron under the seven priors.

Gene	z-score	LAS	aLAS	ILAS	HS	NBP	DL	HS+
610	5.29	4.87	5.00	4.58	4.89	4.90	4.58	4.88
1720	4.83	4.29	4.52	4.11	4.33	4.36	4.09	4.40
332	4.47	3.98	4.13	3.68	3.86	3.88	3.62	3.89
364	-4.42	-3.85	-4.09	-3.59	-3.80	-3.90	-3.64	-3.76
914	4.40	3.78	3.98	3.63	3.79	3.85	3.58	3.30
3940	-4.33	-3.78	-4.00	-3.46	-3.71	-3.76	-3.50	-3.61
4546	-4.29	-3.70	-3.91	-3.37	-3.40	-3.64	-3.36	-3.56
1068	4.25	3.61	3.86	3.35	3.45	3.67	3.34	3.66
579	4.19	3.61	3.81	3.33	3.33	3.54	3.27	3.55
4331	-4.14	-3.61	-3.80	-3.28	-3.27	-3.54	-3.14	-3.42

7.4 Discussion

In this research, the repeated multiplication of the log-terms to the density is proven successful in defining the new class of distributions that are continuous, proper and extremely heavy-tailed. Although the focus of this research is on tail-robustness, it is also natural to consider the idea of log-adjustment to define the stronger shrinkage effect. To be precise, *the doubly log-adjusted shrinkage prior*, whose density in the scale of κ_i is given by,

$$\pi(\kappa_i;\alpha,\beta,L) \propto {\kappa_i}^{-1}(1-\kappa_i)^{-1} \times \left\{\prod_{l=1}^{L-1} f_l\left(\frac{1}{\kappa_i}\right)^{-1} f_l\left(\frac{1}{1-\kappa_i}\right)^{-1}\right\} f_L\left(\frac{1}{\kappa_i}\right)^{-(1+\alpha)} f_L\left(\frac{1}{1-\kappa_i}\right)^{-(1+\beta)}$$

is of great interest. We proved that, as iteration L increases, for any bounded sequence of hyperparameters (α_L, β_L), the prior, $\pi(\kappa_i;\alpha_L,\beta_L,L)$, converges in distribution to the point masses on $\{\kappa_i = 0\}$ and $\{\kappa_i = 1\}$, i.e., the spike-and-slab prior. For the details of the proof, see Section 7.5.8. The resemblance to the degenerate prior shown in this result could justify the use of iteratively log-adjusted priors as the continuous alternative of the degenerate variable-selection priors. Although the finite-sample properties of Bayes estimators under the prior above is not developed here, the posterior inference with this prior is feasible by the same augmentation we proved for the iteratively log-adjusted priors. We believe that the priors with iterated logarithm is the promising future research in exploring the class of shrinkage priors with logarithms.

7.5 Appendix

7.5.1 Proof of Theorem 7.2.1

This proof can be obtained as the special case of the proof of Theorem 7.2.3 with $L = 1$ given in Section 7.5.5.

7.5.2 Proof of Theorem 7.2.2

We first provide a useful lemma. For details, see, for example, the discussion at the end of Section 1.2 of Seneta (1976).

Lemma 7.5.1 *Let $L(u)$ be a strictly positive and continuously differentiable function of $u > 0$. Suppose that*

$$\lim_{u\to\infty} \frac{uL'(u)}{L(u)} = 0.$$

Then the function $L(u)$ is slowly varying as $u \to \infty$, that is, $\lim_{M\to\infty} L(Mv)/L(M) = 1$ for all $v > 0$.

We will suppress the subscript i and write u, θ, and y for u_i, θ_i, and y_i, respectively, for notational simplicity. Let $p(\theta)$ and $m(y)$ denote the marginal densities of θ and y under the log-adjusted shrinkage prior $\pi(u) \propto u^{a-1}(1+u)^{-a-b}\{1+\log(1+u)\}^{-(1+\gamma)}$. We define $S(u)$ as

$$S(u) = \Big(\frac{u}{1+u}\Big)^{a+b}\big\{1+\log(1+u)\big\}^{-(1+\gamma)}, \tag{7.5.1}$$

so that $\pi(u) = C^{-1}u^{-b-1}S(u)$ with normalizing constant $C = \int_0^\infty u^{-b-1}S(u)du$. From Lemma 7.5.1, it can be shown that $S(\cdot)$ is a slowly varying function.

We first note that the posterior mean squared error can be written as

$$\mathrm{MSE}_\pi(\theta|y) = 1 + \frac{1}{m(y)}\frac{\partial^2 m(y)}{\partial y^2},$$

since the second order derivative of $m(y)$ can be expressed as

$$\begin{aligned}\frac{\partial^2 m(y)}{\partial y^2} &= \int_{-\infty}^{\infty}\Big[\frac{\partial^2}{\partial y^2}\frac{1}{\sqrt{2\pi}}\exp\Big\{-\frac{(y-\theta)^2}{2}\Big\}\Big]p(\theta)d\theta \\ &= \int_{-\infty}^{\infty}\{-1+(\theta-y)^2\}\frac{1}{\sqrt{2\pi}}\exp\Big\{-\frac{(y-\theta)^2}{2}\Big\}p(\theta)d\theta.\end{aligned}$$

On the other hand, since $y|u \sim \mathrm{N}(0, 1+u)$, we have that

$$m(y) = \frac{1}{\sqrt{2\pi}}\int_0^\infty \frac{1}{\sqrt{1+u}}\exp\Big(-\frac{y^2/2}{1+u}\Big)\pi(u)du$$

and hence that

$$\begin{aligned}\frac{\partial^2 m(y)}{\partial y^2} = &-\frac{1}{\sqrt{2\pi}}\int_0^\infty \frac{1}{(1+u)^{3/2}}\exp\Big(-\frac{y^2/2}{1+u}\Big)\pi(u)du \\ &+\frac{y^2}{\sqrt{2\pi}}\int_0^\infty \frac{1}{(1+u)^{5/2}}\exp\Big(-\frac{y^2/2}{1+u}\Big)\pi(u)du.\end{aligned}$$

Therefore, by making the change of variables $u = (y^2/2)v$, it follows that

$$\frac{1}{m(y)}\frac{\partial^2 m(y)}{\partial y^2} = \frac{2}{y^2}\frac{2I(y,5/2) - I(y,3/2)}{I(y,1/2)}, \tag{7.5.2}$$

where

$$I(y,k) = \int_0^\infty \Big\{\frac{y^2/2}{1+(y^2/2)v}\Big\}^k v^{-b-1}\exp\Big\{-\frac{y^2/2}{1+(y^2/2)v}\Big\}S\Big(\frac{y^2}{2}v\Big)dv$$

for $k \in \{1/2, 3/2, 5/2\}$. We here use the following asymptotic evaluation of the integral $I(y,k)$:

$$\lim_{|y|\to\infty} I(y,k)/S(y^2/2) = \Gamma(b+k), \tag{7.5.3}$$

for which the proof is given later. Using this result, (7.5.2) can be approximated as

$$\begin{aligned}\frac{1}{m(y)}\frac{\partial^2 m(y)}{\partial y^2}\Big/\frac{2}{y^2} &= \frac{2\Gamma(b+5/2)\{1+o(1)\} - \Gamma(b+3/2)\{1+o(1)\}}{\Gamma(b+1/2)\{1+o(1)\}} \\ &\sim (1+b)(1+2b)\end{aligned}$$

as $|y| \to \infty$, which is the desired result.

Finally, we give the proof of (7.5.3). Let $M = y^2/2$ and define $h_M(v,k)$ by

$$h_M(v,k) = \Big(\frac{M}{1+Mv}\Big)^k v^{-b-1}\exp\Big(-\frac{M}{1+Mv}\Big)\frac{S(Mv)}{S(M)}.$$

Then it holds that $I(y,k)/S(y^2/2) = \int_0^\infty h_M(v,k)dv$. Note that

$$\frac{S(Mv)}{S(M)} = v^{a+b}\Big(\frac{1+M}{1+Mv}\Big)^{a+b}\Big\{\frac{1+\log(1+M)}{1+\log(1+Mv)}\Big\}^{1+\gamma}.$$

Then, for any $M \geq 1$ and $v \geq 1$, we have

$$h_M(v,k) \leq v^{-k-b-1}v^{a+b}\Big(\frac{1+M}{1+Mv}\Big)^{a+b}\Big\{\frac{1+\log(1+M)}{1+\log(1+Mv)}\Big\}^{1+\gamma} \leq 2^{a+b}v^{-k-b-1}.$$

Next, for any $M \geq 1$ and $v \leq 1$,

$$\begin{aligned}\frac{1+\log(1+M)}{1+\log(1+Mv)} &= \exp\Big\{\int_v^1 \frac{M}{1+Mt}\frac{1}{1+\log(1+Mt)}dt\Big\} \\ &\leq \exp\Big(\int_v^1 \frac{1}{1/M+t}dt\Big) = \frac{1/M+1}{1/M+v} \leq \frac{2}{1/M+v}.\end{aligned}$$

Then, it follows that, for $M \geq 1$ and $v \leq 1$,

$$\begin{aligned}h_M(v,k) &= \frac{e^{-1/(1/M+v)}}{(1/M+v)^k}v^{a-1}\Big(\frac{1/M+1}{1/M+v}\Big)^{a+b}\Big\{\frac{1+\log(1+M)}{1+\log(1+Mv)}\Big\}^{1+\gamma} \\ &\leq \frac{e^{-1/(1/M+v)}}{(1/M+v)^{k+a+b}}v^{a-1}2^{a+b}\frac{2^{1+\gamma}}{(1/M+v)^{1+\gamma}} \\ &\leq \Big(\sup_{x\in(0,\infty)}\frac{e^{-1/x}}{x^{k+a+b+1+\gamma}}\Big)2^{a+b+1+\gamma}v^{a-1} < \infty\end{aligned}$$

noting that the function $e^{-1/x}/x^{k+a+b+1+\gamma}$ is bounded in $(0,\infty)$. Therefore, from the dominated convergence theorem, we have

$$\lim_{M\to\infty}\int_0^\infty h_M(v,k) = \int_0^\infty v^{-k-b-1}\exp(-1/v)dv = \Gamma(b+k),$$

which proves (7.5.3).

7.5.3 Details on sampling from γ given in Algorithm 7.2.2

We describe and justify the algorithm of the independent Metropolis-Hastings method for sampling hyperparameter γ by evaluating the upper and lower bounds of the intractable normalizing constant. The normalizing constant of the log-adjusted shrinkage prior, which is dependent on γ, is given by the integral

$$C(\gamma) = C\Big(a=\frac{1}{n}, b=0, \gamma\Big) = \int_0^1 \frac{\kappa^{-1}(1-\kappa)^{1/n-1}}{(1-\log\kappa)^{1+\gamma}}d\kappa = \int_0^\infty g(x;\gamma)dx, \tag{7.5.4}$$

where $g(x;\gamma) = (1-e^{-x})^{1/n-1}(1+x)^{-1-\gamma}$ for $x>0$; the last integral is obtained by the change of variables $\kappa = e^{-x}$. This is bounded above and below by, with any $K>0$ and $N=K^3$,

$$\begin{aligned} U(\gamma,K) &= \frac{(1-e^{-1/K})^{1/n}}{(1/n)e^{-1/K}} + \frac{(1-e^{-K})^{1/n-1}}{\gamma(1+K)^\gamma} \\ &\quad + \sum_{j=1}^N \frac{K^2-1}{KN} g\Big(\frac{1+(j-1)(K^2-1)/N}{K};\gamma\Big) \\ L(\gamma,K) &= \frac{(1-e^{-1/K})^{1/n}}{(1/n)(1+1/K)^\gamma} + \frac{1}{\gamma(1+K)^\gamma} + \sum_{j=1}^N \frac{K^2-1}{KN} g\Big(\frac{1+j(K^2-1)/N}{K};\gamma\Big) \end{aligned}$$

i.e., $L(\gamma,K) \le C(\gamma) \le U(\gamma,K)$ for any (γ,K). In addition, these bounds can be as tight as desired if one increases K; we prove $L(\gamma,K)\to C(\gamma)$ and $U(\gamma,K)\to C(\gamma)$ as $K\to\infty$ in Lemma 7.5.2 at the end of this section. These bounds are utilized in implementing the independent Metropolis-Hasting algorithm, where the acceptance probability is dependent on the intractable normalizing constant and cannot be directly computed, but their upper and lower bounds are available with arbitrary accuracy.

The prior for γ is the gamma distribution, $\gamma\sim \mathrm{Ga}(a_0^\gamma, b_0^\gamma)$. Each likelihood $\pi(u_i;\gamma)$ can be approximated by $\gamma^a\exp\{-\gamma\log(1+\log(1+u_i))\}$ with $a>0$ so that the gamma prior becomes conjugate. The proposal distribution is the posterior distribution with the approximate likelihoods, or $\gamma\sim\mathrm{Ga}(\gamma|a_1^\gamma, b_1^\gamma)$, where

$$a_1^\gamma = a_0^\gamma + na, \quad \text{and} \quad b_1^\gamma = b_0^\gamma + \sum_{i=1}^n \log(1+\log(1+u_i)),$$

and we set $a=1/n$. Denote the current state by γ, and the candidate drawn from the proposal by γ'. The acceptance probability is

$$A(\gamma\to\gamma') = \Big\{\frac{C(\gamma)}{C(\gamma')}\Big\}^n \left(\frac{\gamma}{\gamma'}\right)^{n(1/n)},$$

which is not evaluated directly due to the intractable constant $C(\gamma)/C(\gamma')$. We bound these constants to obtain the upper and lower bounds of the acceptance probability. The bounds of the acceptance probability are defined by

$$\underline{w}(\gamma,\gamma',K)=\Big\{\frac{L(\gamma,K)}{U(\gamma',K)}\Big\}^n\frac{\gamma}{\gamma'}\qquad\text{and}\qquad\overline{w}(\gamma,\gamma',K)=\Big\{\frac{U(\gamma,K)}{L(\gamma',K)}\Big\}^n\frac{\gamma}{\gamma'},$$

and satisfy

$$\underline{w}(\gamma,\gamma',K)\le A(\gamma\to\gamma')\le\overline{w}(\gamma,\gamma',K).$$

The definitions of $\underline{w}(\gamma,\gamma',K)$ and $\overline{w}(\gamma,\gamma',K)$ above are used in the sampling algorithm given in Algorithm 7.2.2.

The bounds, $U(\gamma,K)$ and $L(\gamma,K)$, are obtained by a straightforward application of the Riemann approximation, and their properties are verified by the following lemma.

Lemma 7.5.2 *Let $g(\cdot)$, $\underline{g}_K(\cdot)$, and $\overline{g}_K(\cdot)$, $K=1,2,\dots$, be integrable functions defined on $(0,\infty)$ satisfying $0\le\underline{g}_K(x)\le g(x)\le\overline{g}_K(x)<\infty$ for all $x\in(0,\infty)$. Assume that $g(\cdot)$ is nonincreasing on $(0,\infty)$. Let $0<x_0^{(K)}<\cdots<x_{l_K}^{(K)}<\infty$ for $K=1,2,\dots$ and assume that $\lim_{K\to\infty}x_0^{(K)}=0$ and that $\lim_{K\to\infty}x_{l_K}^{(K)}=\infty$. Suppose that $\lim_{K\to\infty}\int_0^{x_0^{(K)}}\overline{g}_K(x)dx=\lim_{K\to\infty}\int_{x_{l_K}^{(K)}}^{\infty}\overline{g}_K(x)dx=0$ and that $\lim_{K\to\infty}\sum_{j=1}^{l_K}(x_j^{(K)}-x_{j-1}^{(K)})\{g(x_{j-1}^{(K)})-g(x_j^{(K)})\}=0$. Then*

$$\begin{aligned}0&\le\int_0^{x_0^{(K)}}\underline{g}_K(x)dx+\sum_{j=1}^{l_K}(x_j^{(K)}-x_{j-1}^{(K)})g(x_j^{(K)})+\int_{x_{l_K}^{(K)}}^{\infty}\underline{g}_K(x)dx\\&\le\int_0^{\infty}g(x)dx\\&\le\int_0^{x_0^{(K)}}\overline{g}_K(x)dx+\sum_{j=1}^{l_K}(x_j^{(K)}-x_{j-1}^{(K)})g(x_{j-1}^{(K)})+\int_{x_{l_K}^{(K)}}^{\infty}\overline{g}_K(x)dx<\infty\end{aligned}\tag{7.5.5}$$

for all $K=1,2,\dots$ and

$$\begin{aligned}\int_0^{\infty}g(x)dx&=\lim_{K\to\infty}\Big\{\int_0^{x_0^{(K)}}\underline{g}_K(x)dx+\sum_{j=1}^{l_K}(x_j^{(K)}-x_{j-1}^{(K)})g(x_j^{(K)})+\int_{x_{l_K}^{(K)}}^{\infty}\underline{g}_K(x)dx\Big\}\\&=\lim_{K\to\infty}\Big\{\int_0^{x_0^{(K)}}\overline{g}_K(x)dx+\sum_{j=1}^{l_K}(x_j^{(K)}-x_{j-1}^{(K)})g(x_{j-1}^{(K)})+\int_{x_{l_K}^{(K)}}^{\infty}\overline{g}_K(x)dx\Big\}.\end{aligned}\tag{7.5.6}$$

Proof. The inequalities in (7.5.5) are trivial. We obtain (7.5.6) since

$$\begin{aligned}0&\le\int_0^{x_0^{(K)}}\overline{g}_K(x)dx+\sum_{j=1}^{l_K}(x_j^{(K)}-x_{j-1}^{(K)})g(x_{j-1}^{(K)})\\&\quad+\int_{x_{l_K}^{(K)}}^{\infty}\overline{g}_K(x)dx-\sum_{j=1}^{l_K}(x_j^{(K)}-x_{j-1}^{(K)})g(x_j^{(K)})\\&=\sum_{j=1}^{l_K}(x_j^{(K)}-x_{j-1}^{(K)})\{g(x_{j-1}^{(K)})-g(x_j^{(K)})\}+\int_0^{x_0^{(K)}}\overline{g}_K(x)dx+\int_{x_{l_K}^{(K)}}^{\infty}\overline{g}_K(x)dx\to0\end{aligned}$$

as $K \to \infty$ by assumption. □

In our problem, where function g is given in (7.5.4), the condition in the lemma,

$$\lim_{K\to\infty} \sum_{j=1}^{l_K} (x_j^{(K)} - x_{j-1}^{(K)})\{g(x_{j-1}^{(K)}) - g(x_j^{(K)})\} = 0,$$

is satisfied. To see this, observe that $\lim_{K\to\infty}\{\max_{1\le j\le l_K}(x_j^{(K)} - x_{j-1}^{(K)})\}g(x_0^{(K)}) = 0$, and that the grid becomes sufficiently fine when $K \to \infty$.

7.5.4 Properties of iterated logarithmic functions

We here give some properties of iterated logarithmic functions related to Figure 7.2 in the following lemma.

Lemma 7.5.3 *For $x > 1$,*

(i) $f_L(x) > 1$.

(ii) $f_L(x)$ *is increasing in* x.

(iii) $f_{L+1}(x) < f_L(x)$ *(decreasing in* L *at each point* x*).*

(iv) $\lim_{L\to\infty} f_L(x) = 1$.

Proof. The first and second properties follow immediately from the definition of f_L. The third property is verified by the inequality $z - 1 > \log z$ for $z > 1$. To prove the last property, fix $x > 1$ and write $a_L = f_L(x)$. Then this sequence is decreasing and bounded below by 1. Therefore, it has a limit $a = \lim_{L\to\infty} a_L$ in $[1, \infty)$. Now, by the definition of f_{L+1}, we have $a_{L+1} = 1 + \log a_L$. Letting $L \to \infty$, we have $a = 1 + \log a$, which shows that $a = 1$. □

7.5.5 Proof of Theorem 7.2.3

As in the proof of Theorem 7.2.2, we suppress the subscript i. Here we write $\pi(u)$ and $\pi(\kappa)$ for $\pi(u; \gamma, L)$ and $\pi(\kappa; \gamma, L)$ and use $p(\theta)$ to denote the marginal density of θ under this prior. Let

$$S(u) = \Big(\frac{u}{1+u}\Big)^{a+b}\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1+u)}\Big\}\frac{1}{\{f_L(1+u)\}^{1+\gamma}}$$

and let $C = \int_0^\infty u^{-b-1}S(u)du$, so that $\pi(u) = C^{-1}u^{-b-1}S(u)$. Then $S(u)$ is a slowly varying function from Lemma 7.5.1. Note that the above definition of $S(u)$ is not identical to that in Section 7.5.2. Also, let

$$\pi_0(\kappa) = \pi_0(\kappa; L) = \frac{\partial}{\partial\kappa}\Big[\frac{1}{\{f_L(1/\kappa)\}^\gamma}\Big] = \frac{\gamma}{\kappa}\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1/\kappa)}\Big\}\frac{1}{\{f_L(1/\kappa)\}^{1+\gamma}}.$$

Then we have

$$\int_0^1 \pi_0(\kappa)d\kappa = 1.$$

Integrability (the first property in Theorem 7.2.1)

The inequality, $\kappa^b(1-\kappa)^{a-1}\pi_0(\kappa) \le (1-\kappa)^{a-1}\pi_0(\kappa)$, shows the integrability of $\pi(\kappa)$ in $(0,1/2)$,

$$\int_0^{1/2}(1-\kappa)^{a-1}\pi_0(\kappa)d\kappa \le \Big\{\sup_{\kappa\in(0,1/2)}(1-\kappa)^{a-1}\Big\}\int_0^{1/2}\pi_0(\kappa)d\kappa < \infty$$

and the integrability in $(1/2,1)$,

$$\int_{1/2}^{1}(1-\kappa)^{a-1}\pi_0(\kappa)d\kappa \le \int_{1/2}^{1}(1-\kappa)^{a-1}2\gamma d\kappa < \infty.$$

Spike around the origin (the second property in Theorem 7.2.1)

By the monotone convergence theorem,

$$\begin{aligned} C\sqrt{2\pi}p(\theta) &= \int_0^\infty \frac{u^{-1/2+a-1}}{(1+u)^{a+b}}\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1+u)}\Big\}\frac{e^{-\theta^2/2u}}{\{f_L(1+u)\}^{1+\gamma}}du \\ &\ge \int_0^1 \frac{u^{-1/2+a-1}}{2^{a+b}}\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(2)}\Big\}\frac{e^{-\theta^2/2u}}{\{f_L(2)\}^{1+\gamma}}du \to \infty \end{aligned} \tag{7.5.7}$$

as $\theta \to 0$ for $a \le 1/2$.

Slowly varying density (the third property in Theorem 7.2.1)

By the change of variables $u = (\theta^2/2)v$, it follows that

$$\begin{aligned} p(\theta) &= \frac{C^{-1}}{\sqrt{2\pi}}\int_0^\infty u^{-1/2-b-1}e^{-(\theta^2/2)/u}S(u)du \\ &= \frac{C^{-1}}{\sqrt{2\pi}}\Big(\frac{\theta^2}{2}\Big)^{-1/2-b}\int_0^\infty v^{-1/2-b-1}e^{-1/v}S\Big(\frac{\theta^2}{2}v\Big)dv. \end{aligned}$$

Now for $\theta^2/2 \ge 1$, we have

$$\begin{aligned} \frac{S((\theta^2/2)v)}{S(\theta^2/2)} &= v^{a+b}\Big\{\frac{1+\theta^2/2}{1+(\theta^2/2)v}\Big\}^{a+b} \\ &\quad\times\Big\{\prod_{k=1}^{L-1}\frac{f_k(1+\theta^2/2)}{f_k(1+(\theta^2/2)v)}\Big\}\Big\{\frac{f_L(1+\theta^2/2)}{f_L(1+(\theta^2/2)v)}\Big\}^{1+\gamma} \\ &\le 2^{a+b}\max\{1, 1/v^{L+\gamma}\} \end{aligned}$$

since

$$\frac{1+\theta^2/2}{1+(\theta^2/2)v} \le \frac{1}{v}\frac{1+\theta^2/2}{\theta^2/2} \le \frac{2}{v}$$

and since for all $k = 1,\ldots,L$,

$$\frac{f_k(1+\theta^2/2)}{f_k(1+(\theta^2/2)v)} \le 1$$

when $v \geq 1$ while

$$\begin{aligned}\frac{f_k(1+\theta^2/2)}{f_k(1+(\theta^2/2)v)} &= \exp\Big\{\int_{t=v}^{t=1} \frac{\partial}{\partial t}\log f_k(1+(\theta^2/2)t)dt\Big\} \\ &= \exp\Big\{\int_{t=v}^{t=1} \frac{1}{f_k(1+(\theta^2/2)t)\cdots f_1(1+(\theta^2/2)t)}\frac{\theta^2/2}{1+(\theta^2/2)t}dt\Big\} \\ &\leq \exp\Big(\int_{t=v}^{t=1}\frac{dt}{t}\Big) = 1/v\end{aligned}$$

when $v < 1$. Thus, by the dominated convergence theorem, we obtain

$$\frac{p(\theta)}{S(\theta^2/2)} \sim \frac{C^{-1}}{\sqrt{2\pi}}\Big(\frac{\theta^2}{2}\Big)^{-1/2-b}\Gamma\Big(\frac{1}{2}+b\Big)$$

as $|\theta| \to \infty$, where $S(\theta^2/2)$ is a slowly varying function of $|\theta|$. This completes the proof of the first statement of Theorem 7.2.3.

Convergence to the degenerate distribution (the second property in Theorem 7.2.3)

To prove the second statement of Theorem 7.2.3, we have that

$$\lim_{L\to\infty}\int_0^\varepsilon \pi_0(\kappa;L)d\kappa = \lim_{L\to\infty}\Big[\frac{1}{\{f_L(1/\kappa)\}^\gamma}\Big]_0^\varepsilon = \lim_{L\to\infty}\frac{1}{\{f_L(1/\varepsilon)\}^\gamma} = 1 \tag{7.5.8}$$

for all $0 < \varepsilon < 1$. Hence,

$$0 = \limsup_{L\to\infty}\int_\kappa^1 \pi_0(\tilde{\kappa};L)d\tilde{\kappa} \geq \limsup_{L\to\infty}\int_\kappa^1 \kappa\pi_0(\kappa;L)d\tilde{\kappa} = \limsup_{L\to\infty}\{\kappa(1-\kappa)\pi_0(\kappa;L)\} \geq 0,$$

or, equivalently, $\lim_{L\to\infty}\pi_0(\kappa;L) = 0$, for all $\kappa \in (0,1)$. Now, set $b = 0$. Then, for any $0 < \varepsilon < 1$, the probability of $(\varepsilon, 1)$ under the prior $\pi(\kappa;L) = \pi(\kappa;\gamma,L) \propto (1-\kappa)^{a-1}\pi_0(\kappa;L)$ is

$$\int_\varepsilon^1 \pi(\kappa;L)d\kappa = \frac{\displaystyle\int_\varepsilon^1 (1-\kappa)^{a-1}\pi_0(\kappa;L)d\kappa}{\displaystyle\int_0^1 (1-\kappa)^{a-1}\pi_0(\kappa;L)d\kappa} \leq \frac{\displaystyle\int_\varepsilon^1 (1-\kappa)^{a-1}\pi_0(\kappa;L)d\kappa}{\displaystyle\Big\{\inf_{\kappa\in(0,1/2)}(1-\kappa)^{a-1}\Big\}\int_0^{1/2}\pi_0(\kappa;L)d\kappa}.$$

The numerator on the right side converges to zero as $L \to \infty$ by the dominated convergence theorem since $\pi_0(\kappa;L) \leq \gamma/\varepsilon$ for all $\varepsilon < \kappa < 1$ and $\pi_0(\kappa;L) \to 0$. Also, it follows from (7.5.8) that $\int_0^{1/2}\pi_0(\kappa;L)d\kappa \to 1$ as $L \to \infty$. Thus, $\lim_{L\to\infty}\int_0^\varepsilon \pi(\kappa;L)d\kappa = \lim_{L\to\infty}\{1 - \int_\varepsilon^1 \pi(\kappa;L)d\kappa\} = 1$, which is the desired result.

7.5.6 Proof of Theorem 7.2.4

As in the previous section, we suppress the subscript i. Let $\pi(u)$, $S(u)$, and C be as in Section 7.5.5.

The formal proof of Theorem 7.2.4 is completely analogous to that of Theorem 7.2.2 since

$$-\frac{3a/2}{y^2/2} + \sum_{k-1}^{L-1}\frac{1}{f_k(1+y^2/2)\cdots f_1(1+y^2/2)} + \frac{1+\gamma}{f_L(1+y^2/2)\cdots f_1(1+y^2/2)} = o(1)$$

as $|y| \to \infty$. In the following, we informally derive the expression (7.2.7) using integration by parts.

First, $\pi(u)$ is approximated by $\tilde{\pi}(u) = \pi(u)e^{-\varepsilon/u}$ for some small $\varepsilon > 0$ in the sense that

$$\mathrm{MSE}_{\pi}(\theta|y) - 1 \approx \mathrm{MSE}_{\tilde{\pi}}(\theta|y) - 1. \tag{7.5.9}$$

Denote by $\tilde{m}(y)$ the marginal density of y under the prior $\tilde{C}^{-1}\tilde{\pi}(u)$, where $\tilde{C} = \int_0^{\infty} \tilde{\pi}(u)du$. Then, as in the proof of Theorem 7.2.2, we have that

$$\mathrm{MSE}_{\tilde{\pi}}(\theta|y) = 1 + \frac{\tilde{m}''(y)}{\tilde{m}(y)}$$

and that

$$\tilde{m}(y) = \frac{\tilde{C}^{-1}}{\sqrt{2\pi}} \int_0^{\infty} \frac{1}{\sqrt{1+u}} \exp\Big(- \frac{y^2/2}{1+u}\Big)\tilde{\pi}(u)du.$$

Now, by integration by parts,

$$\begin{aligned}
\tilde{m}''(y) &= \frac{\tilde{C}^{-1}}{\sqrt{2\pi}} \int_0^{\infty} \Big\{ \frac{y^2}{(1+u)^2} - \frac{1}{1+u} \Big\} \frac{1}{\sqrt{1+u}} \exp\Big(- \frac{y^2/2}{1+u}\Big)\tilde{\pi}(u)du \\
&= \frac{\tilde{C}^{-1}}{\sqrt{2\pi}} \Big\{ \Big[\frac{2}{\sqrt{1+u}} \exp\Big(- \frac{y^2/2}{1+u}\Big)\tilde{\pi}(u)\Big]_0^{\infty} \\
&\quad - \int_0^{\infty} \frac{2}{\sqrt{1+u}} \exp\Big(- \frac{y^2/2}{1+u}\Big)\tilde{\pi}'(u)du \Big\} \\
&= -2\frac{\tilde{C}^{-1}}{\sqrt{2\pi}} \int_0^{\infty} \frac{1}{\sqrt{1+u}} \exp\Big(- \frac{y^2/2}{1+u}\Big)\tilde{\pi}'(u)du.
\end{aligned}$$

Therefore, by the change of variables $u = (y^2/2)v$, we obtain

$$\begin{aligned}
\mathrm{MSE}_{\tilde{\pi}}(\theta|y) &= 1 - 2\frac{\displaystyle\int_0^{\infty} \frac{1}{\sqrt{1+u}} \exp\Big(- \frac{y^2/2}{1+u}\Big)\tilde{\pi}'(u)du}{\displaystyle\int_0^{\infty} \frac{1}{\sqrt{1+u}} \exp\Big(- \frac{y^2/2}{1+u}\Big)\tilde{\pi}(u)du} \\
&= 1 - 2\frac{\displaystyle\int_0^{\infty} \sqrt{\frac{y^2/2}{1+(y^2/2)v}} \exp\Big\{ - \frac{y^2/2}{1+(y^2/2)v}\Big\}\tilde{\pi}'\Big(\frac{y^2}{2}v\Big)dv}{\displaystyle\int_0^{\infty} \sqrt{\frac{y^2/2}{1+(y^2/2)v}} \exp\Big\{ - \frac{y^2/2}{1+(y^2/2)v}\Big\}\tilde{\pi}\Big(\frac{y^2}{2}v\Big)dv}.
\end{aligned} \tag{7.5.10}$$

Now, if $|y|$ is sufficiently large, it follows that

$$\sqrt{\frac{y^2/2}{1+(y^2/2)v}} \exp\Big\{ - \frac{y^2/2}{1+(y^2/2)v}\Big\} \approx \frac{1}{\sqrt{v}} \exp\Big(- \frac{1}{v}\Big) \tag{7.5.11}$$

and that

$$\tilde{\pi}\Big(\frac{y^2}{2}v\Big) \approx \pi\Big(\frac{y^2}{2}v\Big) \approx C^{-1}\Big(\frac{y^2}{2}\Big)^{-1} v^{-1} S\Big(\frac{y^2}{2}\Big) \tag{7.5.12}$$

since $S(u)$ is a slowly varying function. Furthermore, since

$$\frac{\tilde{\pi}'(u)}{\tilde{\pi}(u)} - \frac{\varepsilon}{u^2} = \frac{\pi'(u)}{\pi(u)} = -\frac{1}{u} + \frac{a}{u(1+u)} - \sum_{k=1}^{L-1} \frac{1}{f_k(1+u)} \cdots \frac{1}{f_1(1+u)} \frac{1}{1+u} - (1+\gamma) \frac{1}{f_L(1+u)} \cdots \frac{1}{f_1(1+u)} \frac{1}{1+u}$$

and since $f_k(1+u)$, $k = 1, \ldots, L$, are slowly varying functions, it also follows that

$$\frac{\tilde{\pi}'((y^2/2)v)}{\tilde{\pi}((y^2/2)v)} \approx -\frac{1}{y^2/2}\frac{1}{v} + \frac{a}{(y^2/2)^2}\frac{1}{v^2} - \sum_{k=1}^{L-1} \frac{1}{f_k(1+y^2/2)} \cdots \frac{1}{f_1(1+y^2/2)} \frac{1}{y^2/2}\frac{1}{v} - (1+\gamma) \frac{1}{f_L(1+y^2/2)} \cdots \frac{1}{f_1(1+y^2/2)} \frac{1}{y^2/2}\frac{1}{v} \tag{7.5.13}$$

for sufficiently large $|y|$. Substituting (7.5.11), (7.5.12), and (7.5.13) into (7.5.10) yields

$$\begin{aligned} \mathrm{MSE}_{\tilde{\pi}}(\theta|y) &\approx 1 - 2\frac{\int_0^\infty v^{-1/2} e^{-1/v} v^{-1} \{\tilde{\pi}'((y^2/2)v)/\tilde{\pi}((y^2/2)v)\} dv}{\int_0^\infty v^{-1/2} e^{-1/v} v^{-1} dv} \\ &\approx 1 + 2\Big\{ \frac{1}{y^2/2}\frac{1}{2} - \frac{a}{(y^2/2)^2}\frac{1}{2}\frac{3}{2} \\ &\quad + \sum_{k=1}^{L-1} \frac{1}{f_k(1+y^2/2)} \cdots \frac{1}{f_1(1+y^2/2)} \frac{1}{y^2/2}\frac{1}{2}\frac{1}{2} \\ &\quad + (1+\gamma) \frac{1}{f_L(1+y^2/2)} \cdots \frac{1}{f_1(1+y^2/2)} \frac{1}{y^2/2}\frac{1}{2}\frac{1}{2} \Big\} \end{aligned} \tag{7.5.14}$$

since $\Gamma(1/2+k) = \int_0^\infty v^{-1/2-k-1} e^{-1/v} dv$ for $k = 0, 1, 2$. Finally, combining (7.5.9) and (7.5.14) yields (7.2.7).

7.5.7 Derivation of the augmentation (7.2.8) and Gibbs sampling given in Algorithm 7.2.3

Let $u > 0$, $c_0 > 0$, and $c_1, \ldots, c_L \geq 0$. We first note that, for any positive $(t_0, t_1, \ldots, t_L)$,

$$\begin{aligned} &\frac{1}{(1+u)^{c_L}} \prod_{k=1}^{L} \frac{1}{\{f_k(1+u)\}^{c_{L-k}}} \\ &\quad = \frac{\mathrm{Ga}(t_0|c_0, 1)}{\mathrm{Ga}(t_0|c_0, f_L(1+u))} \frac{\mathrm{Ga}(t_1|t_0+c_1, 1)}{\mathrm{Ga}(t_1|t_0+c_1, f_{L-1}(1+u))} \times \cdots \\ &\quad \times \frac{\mathrm{Ga}(t_{L-1}|t_{L-2}+c_{L-1}, 1)}{\mathrm{Ga}(t_{L-1}|t_{L-2}+c_{L-1}, f_1(1+u))} \frac{\mathrm{Ga}(t_L|t_{L-1}+c_L, 1)}{\mathrm{Ga}(t_L|t_{L-1}+c_L, 1+u)} e^{-t_L u}. \end{aligned} \tag{7.5.15}$$

The denominator in the right is in fact the probability density of $\boldsymbol{t}_{0:L} \in (0,\infty)^{L+1}$;

$$\int_{(0,\infty)^{L+1}} \mathrm{Ga}(t_0|c_0, f_L(1+u)) \cdots \mathrm{Ga}(t_L|t_{L-1}+c_L, 1+u) d\boldsymbol{t}_{0:L} = 1.$$

From (7.5.15), we obtain

$$\begin{aligned}
&\frac{1}{(1+u)^{c_L}}\prod_{k=1}^{L}\frac{1}{\{f_k(1+u)\}^{c_{L-k}}} \\
&= \int_{(0,\infty)^{L+1}} \Big\{ \mathrm{Ga}(t_0|c_0,1)\mathrm{Ga}(t_1|t_0+c_1,1)\times\cdots \\
&\quad \times \mathrm{Ga}(t_{L-1}|t_{L-2}+c_{L-1},1)\mathrm{Ga}(t_L|t_{L-1}+c_L,1)e^{-t_L u}\Big\} d\boldsymbol{t}_{0:L} \qquad (7.5.16)
\end{aligned}$$

Expression (7.5.16) gives the augmentation(7.2.8) by setting $c_0 = 1+\gamma$, $c_1 = \cdots = c_{L-1} = 1$, and $c_L = a+b$. Moreover, the full conditional distributions of the latent variables in Algorithm 7.2.3 are obtained from (7.2.8) and (7.5.15).

7.5.8 Properties of doubly log-adjusted shrinkage priors in Section 7.4

In this section, we prove the results stated in the discussion of Section 7.4. For $\alpha, \beta > 0$, consider the density of the doubly log-adjusted prior given by

$$\pi(\kappa;\alpha,\beta,L) \propto \pi_0(\kappa;\alpha,L)\pi_0(1-\kappa;\beta,L), \quad \kappa\in(0,1),$$

where

$$\pi_0(\kappa;\gamma,L) = \frac{\partial}{\partial\kappa}\Big[\frac{1}{\{f_L(1/\kappa)\}^{\gamma}}\Big] = \frac{\gamma}{\kappa}\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1/\kappa)}\Big\}\frac{1}{\{f_L(1/\kappa)\}^{1+\gamma}}$$

as defined in Section 7.5.5 (or, this is the iteratively log-adjusted shrinkage prior with $a=1$ and $b=0$ in (7.2.6)), and let $F(\kappa;\alpha,\beta,L)$ denote the corresponding distribution function. For $0<\varepsilon<1$, let

$$R(\varepsilon;\alpha,\beta,L) = \frac{F(\varepsilon;\alpha,\beta,L)}{1-F(1-\varepsilon;\alpha,\beta,L)}$$

be the ratio of the prior probability of $\kappa\in(0,\varepsilon)$ to that of $\kappa\in(1-\varepsilon,1)$. Proposition 7.5.1 summarizes the fundamental properties of $\pi(\kappa;\alpha,\beta,L)$.

Proposition 7.5.1 *The prior $\pi(\kappa;\alpha,\beta,L)$ satisfies the following properties.*

1. $\int_0^1 \pi(\kappa;\alpha,\beta,L)d\kappa < \infty$.
2. *(a) If $0<\varepsilon\le 1/2$, then $R(\varepsilon;\alpha,\beta,L)$ is increasing in β and decreasing in α.*
   (b) $\lim_{\beta\to 0} R(\varepsilon;\alpha,\beta,L) = 0$ and $\lim_{\alpha\to 0} R(\varepsilon;\alpha,\beta,L) = \infty$.
   (c) $R(\varepsilon;\alpha,\beta,L) \gtreqless 1$ if and only if $\alpha \lesseqgtr \beta$.
3. *For two arbitrary bounded sequences of positive real numbers, α_L and β_L, $L=1,2,\ldots$, we have $\lim_{L\to\infty}\{F(1-\varepsilon;\alpha_L,\beta_L,L) - F(\varepsilon;\alpha_L,\beta_L,L)\} = 0$ if $0<\varepsilon<1/2$.*

4. The prior density of $u = (1-\kappa)/\kappa$ can be expressed as

$$\begin{aligned}\pi(u;\alpha,\beta,L) \propto \frac{1}{u}\int_{(0,\infty)^{L+1}} &\Big\{\mathrm{Ga}(r_0|1+\alpha,1)\mathrm{Ga}(r_1|r_0+1,1)\times\cdots\\ &\times \mathrm{Ga}(r_{L-1}|r_{L-2}+1,1)\mathrm{Ga}(r_L|r_{L-1},1)e^{-r_L u}\Big\}d\boldsymbol{r}_{0:L}\\ &\times\int_{(0,\infty)^{L+1}} \Big\{\mathrm{Ga}(s_0|1+\beta,1)\mathrm{Ga}(s_1|s_0+1,1)\times\cdots\\ &\times \mathrm{Ga}(s_{L-1}|s_{L-2}+1,1)\mathrm{Ga}(s_L|s_{L-1},1)e^{-s_L/u}\Big\}d\boldsymbol{s}_{0:L}.\end{aligned}$$

Proof. Let $\tilde{\pi}_0(\kappa;\gamma,L) = \pi_0(\kappa;\gamma,L)/\gamma$. Then property 1 follows immediately by

$$\begin{aligned}&\int_0^1 \tilde{\pi}_0(\kappa;\alpha,L)\tilde{\pi}_0(1-\kappa;\beta,L)d\kappa\\ &\le \int_0^1 \tilde{\pi}_0(\kappa;\min\{\alpha,\beta\},L)\tilde{\pi}_0(1-\kappa;\min\{\alpha,\beta\},L)d\kappa\\ &= 2\int_0^{1/2} \tilde{\pi}_0(\kappa;\min\{\alpha,\beta\},L)\tilde{\pi}_0(1-\kappa;\min\{\alpha,\beta\},L)d\kappa\\ &\le 4\int_0^{1/2} \tilde{\pi}_0(\kappa;\min\{\alpha,\beta\},L)d\kappa < \infty.\end{aligned}$$

For property 2, we start from the proof of 2-(a) and 2-(b) with the focus on β. Note that

$$R(\varepsilon;\alpha,\beta,L) = \frac{\int_0^\varepsilon \tilde{\pi}_0(\kappa;\alpha,L)\tilde{\pi}_0(1-\kappa;\beta,L)d\kappa}{\int_{1-\varepsilon}^1 \tilde{\pi}_0(\kappa;\alpha,L)\tilde{\pi}_0(1-\kappa;\beta,L)d\kappa} = \frac{\int_0^\varepsilon \tilde{\pi}_0(\kappa;\alpha,L)\tilde{\pi}_0(1-\kappa;\beta,L)d\kappa}{\int_0^\varepsilon \tilde{\pi}_0(\kappa;\beta,L)\tilde{\pi}_0(1-\kappa;\alpha,L)d\kappa}.$$

Then we have

$$\begin{aligned}&\Big\{\int_0^\varepsilon \tilde{\pi}_0(\kappa;\beta,L)\tilde{\pi}_0(1-\kappa;\alpha,L)d\kappa\Big\}^2\frac{\partial}{\partial\beta}R(\varepsilon;\alpha,\beta,L)\\ &= \int_0^\varepsilon \tilde{\pi}_0(\kappa;\alpha,L)\tilde{\pi}_0(1-\kappa;\beta,L)\{-\log f_L(1/(1-\kappa))\}d\kappa\int_0^\varepsilon \tilde{\pi}_0(\kappa;\beta,L)\tilde{\pi}_0(1-\kappa;\alpha,L)d\kappa\\ &\quad-\int_0^\varepsilon \tilde{\pi}_0(\kappa;\alpha,L)\tilde{\pi}_0(1-\kappa;\beta,L)d\kappa\int_0^\varepsilon \tilde{\pi}_0(\kappa;\beta,L)\tilde{\pi}_0(1-\kappa;\alpha,L)\{-\log f_L(1/\kappa)\}d\kappa\\ &> [\{-\log f_L(1/(1-\varepsilon))\}-\{-\log f_L(1/\varepsilon)\}]\\ &\quad\times\int_0^\varepsilon \tilde{\pi}_0(\kappa;\alpha,L)\tilde{\pi}_0(1-\kappa;\beta,L)d\kappa\int_0^\varepsilon \tilde{\pi}_0(\kappa;\beta,L)\tilde{\pi}_0(1-\kappa;\alpha,L)d\kappa.\end{aligned}$$

The right-hand side is nonnegative for $0<\varepsilon\le 1/2$. Thus, $R(\varepsilon;\alpha,\beta,L)$ is increasing in β for $0<\varepsilon\le 1/2$. In addition,

$$\begin{aligned}0 \le R(\varepsilon;\alpha,\beta,L) &= \frac{\int_0^\varepsilon \pi_0(\kappa;\alpha,L)\pi_0(1-\kappa;\beta,L)d\kappa}{\int_0^\varepsilon \pi_0(\kappa;\beta,L)\pi_0(1-\kappa;\alpha,L)d\kappa}\\ &\le \frac{\beta/\alpha}{1-\varepsilon}\Big\{\prod_{k=1}^{L-1} f_k(1/(1-\varepsilon))\Big\}\{f_L(1/(1-\varepsilon))\}^{1+\alpha}\frac{\int_0^\varepsilon \pi_0(\kappa;\alpha,L)d\kappa}{\int_0^\varepsilon \pi_0(\kappa;\beta,L)d\kappa}\\ &= \Big\{\prod_{k=1}^{L-1} f_k(1/(1-\varepsilon))\Big\}\{f_L(1/(1-\varepsilon))\}^{1+\alpha}\frac{\int_0^\varepsilon \pi_0(\kappa;\alpha,L)d\kappa}{\alpha(1-\varepsilon)}\frac{\beta}{\{f_L(1/\varepsilon)\}^{-\beta}} \to 0\end{aligned}$$

as $\beta \to 0$.

To prove 2-(a) and 2-(b) for α, note that $R(\varepsilon;\alpha,\beta,L) = 1/R(\varepsilon;\beta,\alpha,L)$ for any $\alpha,\beta > 0$ and any $0 < \varepsilon < 1$. From this fact, it is immediate that $R(\varepsilon;\alpha,\beta,L) \to \infty$ as $\alpha \to 0$ and that $R(\varepsilon;\alpha,\beta,L)$ is decreasing in α for $0 < \varepsilon \le 1/2$.

To prove property 2-(c), we first assume $0 < \varepsilon \le 1/2$. Because $R(\varepsilon;\alpha,\alpha;L) = 1$ for any $\varepsilon \in (0,1)$ by definition and it is increasing in the second α, we have $1 = R(\varepsilon;\alpha,\alpha;L) < R(\varepsilon;\alpha,\beta,L)$ for $\alpha < \beta$. Similarly, we have $R(\varepsilon;\alpha,\beta,L) < 1$ for $\alpha > \beta$. To extend this result to $1/2 < \varepsilon < 1$, we first confirm that $R(1-\varepsilon,\alpha,\beta,L) > 1$ for $\alpha < \beta$, which implies $\int_0^{1-\varepsilon} \pi_0(\kappa;\alpha,\beta,L)d\kappa > \int_0^{1-\varepsilon} \pi_0(\kappa;\beta,\alpha,L)d\kappa$. Then, observe that

$$\begin{aligned} R(\varepsilon;\alpha,\beta,L) &= \frac{\int_0^{\varepsilon} \pi_0(\kappa;\alpha,\beta,L)d\kappa}{\int_0^{\varepsilon} \pi_0(\kappa;\beta,\alpha,L)d\kappa} \\ &= \frac{1-\int_{\varepsilon}^{1} \pi_0(\kappa;\alpha,\beta,L)d\kappa}{1-\int_{\varepsilon}^{1} \pi_0(\kappa;\beta,\alpha,L)d\kappa} = \frac{1-\int_0^{1-\varepsilon} \pi_0(\kappa;\beta,\alpha,L)d\kappa}{1-\int_0^{1-\varepsilon} \pi_0(\kappa;\alpha,\beta,L)d\kappa} > 1. \end{aligned}$$

The same argument applies to $\alpha > \beta$. Thus, we conclude for any $0 < \varepsilon < 1$ that $R(\varepsilon;\alpha,\beta,L) \gtreqless 1$ if and only if $\alpha \lesseqgtr \beta$, which completes the proof of 2-(c).

For the proof of property 3, let $F_0(\kappa;\gamma,L)$ denote the distribution function of the prior $\pi_0(\kappa;\gamma,L)$. Select $M > 0$ large enough so that $0 < \alpha_L < M$ for all $L \ge 1$. Then we have

$$1 \ge F_0(\varepsilon;\alpha_L,L) = \int_0^{\varepsilon} \pi_0(\kappa;\alpha_L,L)d\kappa = \{f_L(1/\varepsilon)\}^{-\alpha_L} \ge \{f_L(1/\varepsilon)\}^{-M} \to 1 \tag{7.5.17}$$

as $L \to \infty$. Next,

$$F(1-\varepsilon;\alpha_L,\beta_L,L) - F(\varepsilon;\alpha_L,\beta_L,L) = \int_{\varepsilon}^{1-\varepsilon} \pi(\kappa;\alpha_L,\beta_L,L)d\kappa = \frac{I(\varepsilon;\alpha_L,\beta_L,L)}{I(0;\alpha_L,\beta_L,L)},$$

where

$$I(\omega;\alpha,\beta,L) = \int_{\omega}^{1-\omega} \pi_0(\kappa;\alpha,L)\pi_0(1-\kappa;\beta,L)d\kappa$$

for $\omega \in [0,1]$ and $\alpha,\beta > 0$. Now suppose $0 < \varepsilon < 1/2$ and let U be a uniform random variable on the interval $(\varepsilon, 1-\varepsilon)$. Then it follows from the covariance inequality that

$$\begin{aligned} \frac{I(\varepsilon;\alpha_L,\beta_L,L)}{1-2\varepsilon} &= E[\pi_0(U;\alpha_L,L)\pi_0(1-U;\beta_L,L)] \\ &= E\Big[\frac{\alpha_L}{1-U}\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1/U)}\Big\}\frac{1}{\{f_L(1/U)\}^{1+\alpha_L}} \\ &\quad \times \frac{\beta_L}{U}\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1/(1-U))}\Big\}\frac{1}{\{f_L(1/(1-U))\}^{1+\beta_L}}\Big] \\ &\le E\Big[\frac{\alpha_L}{1-U}\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1/U)}\Big\}\frac{1}{\{f_L(1/U)\}^{1+\alpha_L}}\Big] \\ &\quad \times E\Big[\frac{\beta_L}{U}\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1/(1-U))}\Big\}\frac{1}{\{f_L(1/(1-U))\}^{1+\beta_L}}\Big] \\ &= \widetilde{I}(\varepsilon;\alpha_L,L)\widetilde{I}(\varepsilon;\beta_L,L), \end{aligned}$$

where

$$\widetilde{I}(\varepsilon;\gamma,L) = E\Big[\frac{\gamma}{1-U}\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1/U)}\Big\}\frac{1}{\{f_K(1/U)\}^{1+\gamma}}\Big]$$
$$= E\Big[\frac{\gamma}{U}\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1/(1-U))}\Big\}\frac{1}{\{f_L(1/(1-U))\}^{1+\gamma}}\Big]$$

for $\gamma > 0$. Furthermore,

$$(1-2\varepsilon)\widetilde{I}(\varepsilon;\gamma,L) = \int_\varepsilon^{1-\varepsilon}\frac{\kappa}{1-\kappa}\pi_0(\kappa;\gamma,L)d\kappa \le \frac{1-\varepsilon}{\varepsilon}\{F_0(1-\varepsilon;\gamma,L) - F_0(\varepsilon;\gamma,L)\}$$

for all $\gamma > 0$. On the other hand, letting

$$h(\kappa;\gamma,L) = \gamma\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1/\kappa)}\Big\}\frac{1}{\{f_L(1/\kappa)\}^{1+\gamma}}$$

for $\kappa \in (0,1)$ and $\gamma > 0$. Noting that h is increasing in κ, we obtain

$$\begin{aligned}
I(0;\alpha_L,\beta_L,L) &\ge \int_0^\varepsilon \pi_0(\kappa;\alpha_L,L)\beta_L\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1/(1-\varepsilon))}\Big\}\frac{1}{\{f_L(1/(1-\varepsilon))\}^{1+\beta_L}}d\kappa \\
&= F_0(\varepsilon;\alpha_L,L)h(1-\varepsilon;\beta_L,L) = \frac{F_0(\varepsilon;\alpha_L,L)}{1-2\varepsilon}\int_\varepsilon^{1-\varepsilon} h(1-\varepsilon;\beta_L,L)d\kappa \\
&\ge \frac{F_0(\varepsilon;\alpha_L,L)}{1-2\varepsilon}\int_\varepsilon^{1-\varepsilon}\frac{\varepsilon}{\kappa}h(\kappa;\beta_L,L)d\kappa = \frac{F_0(\varepsilon;\alpha_L,L)}{1-2\varepsilon}\int_\varepsilon^{1-\varepsilon}\varepsilon\pi_0(\kappa;\beta_L,L)d\kappa \\
&= \frac{\varepsilon}{1-2\varepsilon}F_0(\varepsilon;\alpha_L,L)\{F_0(1-\varepsilon;\beta_L,L) - F_0(\varepsilon;\beta_L,L)\}.
\end{aligned}$$

Thus, we conclude by (7.5.17) that

$$\begin{aligned}
&F(1-\varepsilon;\alpha_L,\beta_L,L) - F(\varepsilon;\alpha_L,\beta_L,L) \\
&\le \frac{1}{1-2\varepsilon}\Big(\frac{1-\varepsilon}{\varepsilon}\Big)^2\frac{1-2\varepsilon}{\varepsilon}\frac{\{F_0(1-\varepsilon;\alpha_L,L) - F_0(\varepsilon;\alpha_L,L)\}\{F_0(1-\varepsilon;\beta_L,L) - F_0(\varepsilon;\beta_L,L)\}}{F_0(\varepsilon;\alpha_L,L)\{F_0(1-\varepsilon;\beta_L,L) - F_0(\varepsilon;\beta_L,L)\}} \\
&= \frac{(1-\varepsilon)^2}{\varepsilon^3}\frac{F_0(1-\varepsilon;\alpha_L,L) - F_0(\varepsilon;\alpha_L,L)}{F_0(\varepsilon;\alpha_L,L)} \to 0
\end{aligned}$$

as $L \to \infty$.

For part 4, note that the unnormalized density of $u = (1-\kappa)/\kappa$ based on $\pi_0(\kappa;\alpha,L)\pi_0(1-\kappa;\beta,L)$ is

$$\frac{\alpha\beta}{u}\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1+u)}\Big\}\frac{1}{\{f_L(1+u)\}^{1+\alpha}}\Big\{\prod_{k=1}^{L-1}\frac{1}{f_k(1+1/u)}\Big\}\frac{1}{\{f_L(1+1/u)\}^{1+\beta}}.$$

Then, apply the integral representation in (7.5.16) to the two products of functions of $1+u$ and $1+1/u$ with $c_L = 0$, $c_{L-1} = \cdots = c_1 = 1$ and $c_0 \in \{1+\alpha, 1+\beta\}$ to obtain the desired result. □

Chapter 8

Log-Regularly Varying Scale Mixture of Normals for Robust Regression

8.1 Introduction

The robustness to outliers in linear regression models has been well-studied for its importance, and the research on theory and methodology for robust statistics has been accumulated in the past years. Yet, the modeling of error distributions in practice to accommodate outliers has not advanced significantly from Student's t-distribution. In modern applied statistics, where data are enriched by massive observations, the more extreme outliers are expected to arrive, and the more likely, and significantly, the inference of regression coefficients and scale parameter is affected by such outliers. Our research aims to contribute to the development of novel error distributions for outlier-robustness which we believe are still in demand.

In the full posterior inference, the concept of robustness is not limited to the point estimation, but targets the whole posterior distributions of parameters of interest. Also known as outlier-proneness or outlier-rejection, the posterior robustness defines the property of posterior distributions that the difference of posteriors with and without outliers diminishes as the values of outliers become extreme (O'Hagan (1979)). The series of research on posterior robustness has revealed both the (sufficient) conditions for error distributions to achieve the robustness, and the specific model that meets such conditions; see the detailed review by O'Hagan and Pericchi (2012). The recent studies introduced the concept of regularly varying density functions (Andrade and O'Hagan (2006, 2011)), which was later extended to log-regularly varying functions (Desgagné (2015); Desgagné and Gagnon (2019)), and provided the robustness conditions for the partial and whole posteriors of interest to be unaffected by outliers. As an error distribution whose density function is log-regularly varying, Gagnon et al. (2020) proposed log-Pareto truncated normal (LPTN) distribution, which replaced the thin-tails of normal distribution by those of heavily-tailed log-Pareto distribution. Despite its desirable property of robustness, the class of LPTN distributions has hyperparameters that are difficult to tune and/or estimate, such as the truncation point of Gaussian tail, that could result in the efficiency loss in practice. Another issue in such distribution is the difficulty in posterior computation; unlike t-distribution, direct sampling from the conditional posteriors is infeasible, and one has to rely on Metroplis-Hastings algorithm, which may result in the increased computational cost.

We, in contrast, explore a different class of error distributions that have received less at-

tention in the literature. Following Box and Tiao (1968), we model the error distribution by the finite mixture of two components; one has thinner tails such as normal distributions, and the other is extremely heavily-tailed to accommodate potential outliers. While remaining in the general class of scale mixture of normals (West (1984)), this simple, intuitive approach to the modeling of outliers contrasts the literature listed above, where the error is modeled by a single, continuous distribution. The structure of finite mixture helps controlling the effect of outliers on the posteriors of parameters of interest, while allowing the conditional conjugacy for posterior computation. For these theoretical and practical utilities, the finite mixture models have been routinely practiced in applied statistics (see, for example, Carter and Kohn (1994), West (1997), Frühwirth-Schnatter (2006) and Tak et al. (2019)). In this research, we specifically focus on this class of error distributions in proving the posterior robustness.

For the heavily-tailed distribution that comprises the finite mixture, Student's t-distribution is still regarded thin-tailed for its outlier sensitivity. We propose the use of distributions that has been utilized in the robust inference for high-dimensional count data (Hamura et al. (2020a)) for their extremely-heavy tails. This is another scale mixture of normals by the gamma distribution with the hierarchical structure on shape parameters, which allows the posterior inference by a simple but efficient Gibbs sampler. The tails of such distributions are heavier than those of Cauchy distribution; this tail property is consistent with those of other heavily-tailed distributions considered for posterior robustness, including LPTN distributions.

The finite mixture of the thinly-tailed and heavily-tailed distributions used as the error distribution in linear models, which we name the extremely heavily-tailed error (EHE) distribution, is proved to achieve the whole posterior robustness. The wider class of error distributions including the EHE distributions is considered, but the error distribution that attains the posterior robustness is shown to be the proposed EHE distribution only. The posterior robustness realized by the EHE distributions is extensively compared with the other alternatives in simulation study, showing its competence in point and interval estimations.

Another notable feature of the EHE distributions is that the posterior robustness is guaranteed for the variety of priors on regression coefficients and scale parameter. The assumptions for the posterior robustness do not exclude the unbounded prior densities for regression coefficients. Such prior distributions include the shrinkage priors for high-dimensional regression, e.g., horseshoe priors (Carvalho et al. (2009, 2010)). We illustrate the utility of the robustness with the shrinkage prior for regression coefficients in the empirical studies for Boston housing dataset that is suspected to be contaminated with possible outliers. Likewise, in another example of the famous diabetes data, we confirm that the loss of efficiency by the introduction of heavily-tailed distribution is minimal even in the absence of outliers.

The rest of this chapter is organized as follows. In Section 8.2, we introduce the new error distribution and describe its use in linear regression models. We also provide theoretical robustness properties regarding the posterior distribution. The algorithm for posterior computation is provided in Section 8.3 with the discussion on its computational efficiency. In Section 8.4, we carry out simulation studies to compare the proposed method with existing ones. In Section 8.5, we illustrate the proposed method using two famous datasets. Finally, we conclude with further discussions in Section 8.6.

8.2 A New Error Distribution for Robust Bayesian Regression

8.2.1 Extremely heavy-tailed error distribution

Let y_i be a response variable and $\boldsymbol{x}_i$ be an associated p-dimensional vector of covariates, for $i = 1, \ldots, n$. We consider a linear regression model, $y_i = \boldsymbol{x}_i{}^\top \boldsymbol{\beta} + \sigma\varepsilon_i$, where $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients and σ is an unknown scale parameter. The error terms, $\varepsilon_1, \ldots, \varepsilon_n$, are directly linked to the posterior robustness; it is well-known that modeling those errors simply by Gaussian distributions makes the posterior inference very sensitive to outliers.

To achieve the posterior robustness, we introduce a local random variable u_i and assume that the error distribution is conditionally Gaussian, as $\varepsilon_i | u_i \sim \mathrm{N}(0, u_i)$. Under this setting, when an outlier arrives, then the higher value of local variable explains such outlier and keeps the posterior distribution of $(\boldsymbol{\beta}, \sigma)$ unchanged. A typical choice of the distribution of u_i is the inverse-gamma distribution, which leads to the marginal distribution of ε_i being the t-distribution. However, as shown in Gagnon et al. (2020) and our main theorem, this choice does not hold desirable robustness properties of the posterior distribution even when the distribution of ε_i is Cauchy distribution.

The model for the local scale variable u_i studied in this research is given by the mixture of two components as follows;

$$u_i = \begin{cases} u_{1i} & \text{if } z_i = 0 \\ u_{2i} & \text{if } z_i = 1 \end{cases}$$

where $\Pr[z_i = 1] = 1 - \Pr[z_i = 0] = s$ with mixing probability $s \in [0, 1]$. These variables independently follow different distributions defined below:

$$u_{1i} \sim \mathrm{Ga}(a, a), \qquad u_{2i} \sim H(\cdot; \gamma) \tag{8.2.1}$$

with fixed value a and unknown parameter $\gamma > 0$. The second, newly-introduced H-distribution is defined by the proper density,

$$H(u; \gamma) = \frac{\gamma}{1+u} \frac{1}{\{1 + \log(1+u)\}^{1+\gamma}}, \quad u > 0,$$

Preparing two distributions in modeling of the variance structure in the form (8.2.1) is based on the same modeling philosophy of Box and Tiao (1968); the first component generates non-outlying errors and the second component is supposed to absorb outlying errors. For non-outlying part, we set $a > 0$ to be a large value such that the variance of $\mathrm{Ga}(a, a)$ is very small; the point mass on unity is included in our model as the limit of $a \to \infty$. In what follows, we adopt $a = 10^8$ as a default choice. In contrast, as the model for outlying errors, the second component $H(\cdot; \gamma)$ is extremely heavily-tailed since $H(u; \gamma) \approx u^{-1} (\log u)^{-1-\gamma}$ as $u \to \infty$, which is known as log-regularly varying density (Desgagné (2015)). This property is inherited to the error distribution and plays an important role in the robustness properties of the posterior distribution.

Under the formulation (8.2.1), the marginal distribution of ε_i is obtained as

$$f_{\mathrm{EH}}(\varepsilon_i) = (1-s) \int_0^\infty \mathrm{N}(\varepsilon_i; 0, u_{1i}) \mathrm{Ga}(u_{1i}; a, a) du_{1i} + s \int_0^\infty \mathrm{N}(\varepsilon_i; 0, u_{2i}) H(u_{2i}; \gamma) du_{2i}, \tag{8.2.2}$$

where $\mathrm{N}(\varepsilon_i; 0, u)$ is the normal density with mean zero and variance u. Both components are the scale mixtures of normals, and the first component is the normal-gamma distribution in

general (Griffin and Brown (2010)), but in our application, it is essentially the standard normal distribution for $a > 0$ is set to a large value. The second component does not admit any closed-form expression. To handle with this component in posterior computation, as we see later in Section 8.3.1, we utilize the augmentation of H-distribution by a couple of gamma-distributed state variables. By this augmentation, the posterior inference for this model is straightforward.

A notable property of the new error distribution is its extremely heavy tails shown in the following proposition, with the proof left in the Appendix.

Proposition 8.2.1 *The density (8.2.2) satisfies*

$$f_{\mathrm{EH}}(x) \approx |x|^{-1}(\log|x|)^{-1-\gamma}$$

for large $|x|$ if $s > 0$.

The above proposition indicates that the density of the EHE distribution is a family of log-regularly varying functions. In addition, the tails of the EHE density are heavier than those of Cauchy distribution; $f_{\mathrm{C}}(x) \approx |x|^{-2}$. This property follows that the EHE distribution directly inherits the heavy tails of the mixing H-distribution in the second component of the density (8.2.2). In what follows, we call the new error distribution (8.2.2) *extremely heavily-tailed error (EHE) distribution.*

The density in (8.2.2) is shown in Figure 8.1 for $s = 0.05, 0.1$ and 0.2, with the standard normal density. It is observed that the shape of the EHE distribution is very similar to one of the standard normal distribution around the origin, whereas the tail gets heavier as s increases. Figure 8.2 shows the cumulative distribution functions (CDFs) of H-distributions and the EHE distributions to emphasize their tail property. The tails of the proposed EHE distributions are heavier than those of Cauchy distribution, as seen in the right panel. This fact is also confirmed via the comparison of CDFs of H- and inverse-gamma distributions in the left panel. Owing to these properties of the EHE density, we can achieve robustness properties for the entire posterior distribution as shown in Theorem 8.2.1.

8.2.2 Robustness properties

We here consider theoretical robustness properties of the posterior distribution based on the proposed EHE distribution. To this end, we consider a wider class of error distributions which includes the proposed distribution as a special case, defined by replacing $H(u;\gamma)$ in (8.2.2) with

$$H(u;\gamma,\delta) = C(\delta,\gamma)\frac{1}{(1+u)^{1+\delta}}\frac{1}{\{1+\log(1+u)\}^{1+\gamma}}, \quad u > 0, \tag{8.2.3}$$

where $C(\delta,\gamma)$ is a normalizing constant, and $\delta \geq 0$ is an additional shape parameter. Note that the distribution in (8.2.3) reduces to the proposed distribution in (8.2.2) under $\delta = 0$. This parameter is also related to the decay of the density tail of (8.2.3), that is, $H(u;\gamma,b) \approx u^{-\delta-1}(\log u)^{-1-\gamma}$. Hence, the tail gets heavier as δ decreases, and the EHE distribution with $\delta = 0$, in fact, has the heaviest tail in this class of distributions. Among this general class in (8.2.3), we show later in Theorem 8.2.1 that only the proposed error distribution with $\delta = 0$ attains the robustness property. This theorem also clarifies the difference from t-distributions with degree of freedom ν, the density tails of which is $u^{-\nu-1}$ and lighter than those of the proposed distribution even when $\nu = 1$ (Cauchy tail).

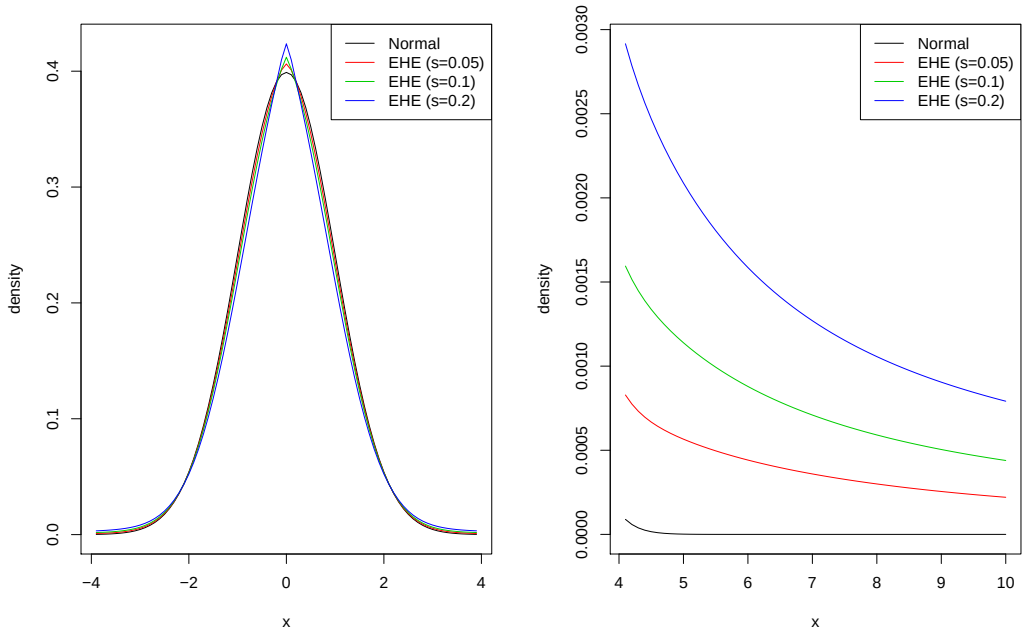

Figure 8.1: Densities of the proposed error distribution with $a = 10^8$, $\gamma = 1$ and $s \in \{0.05, 0.1, 0.2\}$ and the standard normal error distribution. The intractable integral of the second component is computed by the Monte Carlo integration.

For simplicity, we fix γ in what follows, but the same property holds if the support of γ is compact. Let $\mathcal{D}$ be the set of the observed data. To discuss the posterior robustness, we target the *unnormalized* posterior distribution of $(\boldsymbol{\beta}, \sigma)$ under the general error distribution with (8.2.3),

$$\pi_\delta(\boldsymbol{\beta}, \sigma|\mathcal{D}) = \int \prod_{i=1}^{n} \sigma^{-1} f_{\mathrm{EH}}\{\sigma^{-1}(y_i - \boldsymbol{x}_i{}^\top \boldsymbol{\beta}); s, \gamma, \delta\} \pi(\boldsymbol{\Phi}) ds, \tag{8.2.4}$$

where $\boldsymbol{\Phi} = \{\boldsymbol{\beta}, \sigma^2, s\}$ and $\pi(\boldsymbol{\Phi})$ is a joint prior distribution of $\boldsymbol{\Phi}$. Next, to analyze the effect of outliers explicitly, we assume that each outlier goes to infinity at its own specific rate. More precisely, the observed value of responses is parametrized by ω as $y_i = y_i(\omega)$ for some i's, and $|y_i(\omega)| \to \infty$ as $\omega \to \infty$. Let $\mathcal{D}^*$ be the set of non-outlying observations; y_i is independent of ω for $i \in \mathcal{D}^*$. The posterior robustness is defined as the diminishing difference of posteriors conditional on $\mathcal{D}$ and $\mathcal{D}^*$ as $\omega \to \infty$. The formal statement of posterior robustness for our model is given below. For the detailed proof, see the Appendix.

Theorem 8.2.1 *For the unnormalized posterior density given in (8.2.4), it holds that*

$$\pi_\delta(\boldsymbol{\beta}, \sigma|\mathcal{D}) \to \pi_\delta(\boldsymbol{\beta}, \sigma|\mathcal{D}^*) \quad \text{as} \quad \omega \to \infty, \tag{8.2.5}$$

for any $(\boldsymbol{\beta}, \sigma) \in K$ if and only if $\delta = 0$, where K is a compact set.

We note again that the general error distribution with $\delta = 0$ is exactly the proposed EHE distribution, so that the above theorem indicates that the desirable robustness property is achieved

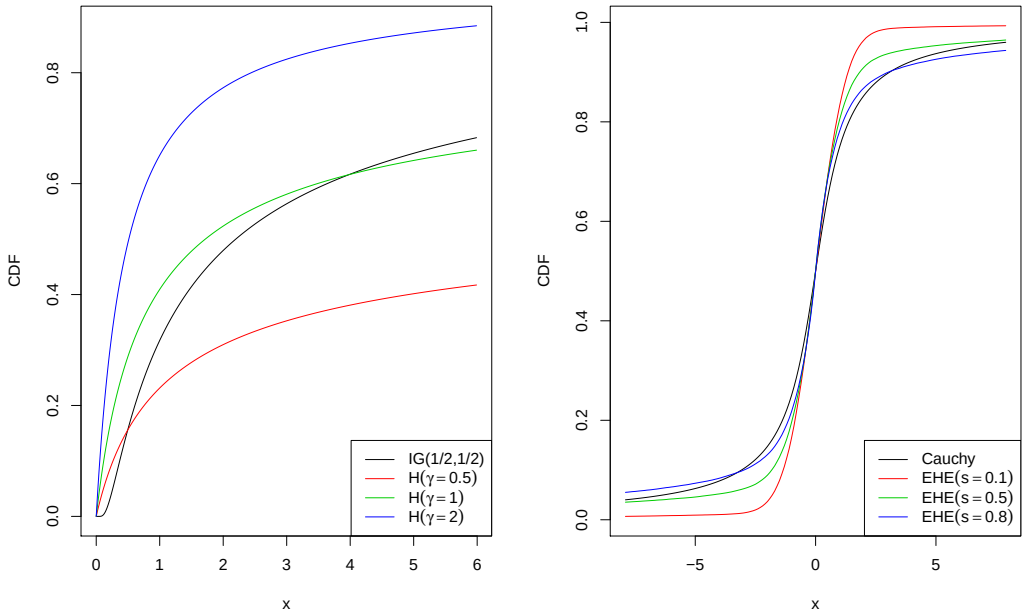

Figure 8.2: Left: Cumulative distribution functions of scale distributions, $H(u;\gamma)$ for $\gamma \in \{0.5, 1.0, 2.0\}$, and the inverse gamma distribution with shape and scale 0.5. Right: The empirical cumulative distributions of the EHE distributions with $\gamma = 1$ and $s = 0.1, 0.5, 0.8$ computed by the Monte Carlo integration, compared with the distribution function of Cauchy distribution.

only under the proposed EHE distribution among the general class of error distributions with the mixing distribution in (8.2.3).

As clarified in the proof of Theorem 8.2.1, the ratio of the two unnormalized posteriors converges to the function of σ and δ if $\delta > 0$. The same asymptotic ratio is obtained for t-distribution with degree-of-freedom δ. In other words, the posterior robustness cannot be attained by the finite mixture with t-distribution.

The main theorem shows the uniform convergence of the posterior distribution with outliers to one without outliers on a compact set. Although this result is proved with almost no assumption other than the model structure, we can also prove other variations of posterior robustness seen in other literature with appropriate conditions. Examples include the convergence with normalized constant and convergence in distribution by introducing additional assumptions on the models and priors. The explicit benefit of the version of posterior robustness in our theorem is the minimal set of assumption required for the priors on $\boldsymbol{\beta}$ and σ^2, and the posterior robustness is valid for any proper priors, even if the density is unbounded. In fact, unbounded density functions are common in some advanced but widely adopted shrinkage priors, such as the horseshoe priors (Carvalho et al. (2010)). Thus, the theoretical framework of this research guarantees the posterior robustness for the boarder and important class of statistical problems, including the high-dimensional regression by shrinkage as an important example.

8.3 Posterior Computation

8.3.1 Gibbs sampler by augmentation

An important property of the proposed EHE distribution (8.2.2) is its computational tractability, that is, we can easily construct a simple Gibbs sampling for posterior inference. Note that the error distribution contains two unknown parameters, s and γ, and we adopt conditionally conjugate priors given by $s \sim \text{Beta}(a_s, b_s)$ and $\gamma \sim \text{Ga}(a_\gamma, b_\gamma)$. The conditionally conjugate priors can also be found for main parameters, $\boldsymbol{\beta}$ and σ^2, and we use $\boldsymbol{\beta} \sim \text{N}(\boldsymbol{A}_{\boldsymbol{\beta}}, \boldsymbol{B}_{\boldsymbol{\beta}})$ and $\sigma^{-2} \sim \text{Ga}(a_\sigma, b_\sigma)$. The multivariate normal prior for $\boldsymbol{\beta}$ can be replaced with the scale mixture of normals, such as shrinkage priors, which is discussed later in Section 8.3.3.

To derive the tractable conditional posteriors, we need to keep the likelihood conditionally Gaussian with scale u_i. For this purpose, we need to rely on a set of latent variables, z_i, u_{1i} and u_{2i}, to obtain a hierarchical expression of u_i. Now, the scale parameter is written as $u_i = (1 - z_i)u_{1i} + z_i u_{2i}$, where z_i, u_{1i} and u_{2i} are mutually independent and distributed as $z_i \sim \text{Ber}(s)$, $u_{1i} \sim \text{Ga}(a, a)$ and $u_{2i} \sim H(u_{2i}; \gamma)$ as in (8.2.1). The conditional conjugacy for (β, σ^2) follows immediately from the Gaussian likelihoods, and the conditional posteriors are normal and inverse gamma, given u_i.

The full conditional distributions of the other parameters and latent variables in the EHE distribution are not any well-known distribution, but we can utilize the following integral expression of density $H(u_{2i}; \gamma)$ as

$$H(u_{2i}; \gamma) = \iint_{(0,\infty)^2} \text{Ga}(u_{2i}; 1, v_i)\text{Ga}(v_i; w_i, 1)\text{Ga}(w_i; \gamma, 1)dv_i dw_i,$$

namely, the random variable u_{2i} following the density $H(u_{2i}; \gamma)$ admits the mixture representation: $u_{2i}|(v_i, w_i) \sim \text{Ga}(1, v_i)$, $v_i|w_i \sim \text{Ga}(w_i, 1)$ and $w_i \sim \text{Ga}(\gamma, 1)$, which enables us to easily generate samples from the full conditional distribution of $(u_{2i}|v_i, w_i)$ and $(v_i, w_i|u_{2i})$.

The latent state (v_i, w_i) is useful in deriving the conditional posterior of u_{2i}, and one can derive the Gibbs sampler with latent (v_i, w_i) as the part of the Markov chain, although (v_i, w_i) is totally redundant in posterior sampling of the other parameters. We, instead, marginalize (v_i, w_i) out when sampling γ, s, u_{1i}'s and z_i's from their conditional posteriors. This modification of the original Gibbs sampler simplifies the sampling procedure, and even facilitates the mixing, while targeting the same stationary distribution (Partially collapsed Gibbs sampler, van Dyk and Park (2019)). The algorithm for posterior sampling is summarized as follows.

Summary of the posterior sampling

- Sample $\boldsymbol{\beta}$ from the full conditional distribution $\text{N}(\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$, where

$$\widetilde{\boldsymbol{B}}^{-1} = \boldsymbol{B}_{\boldsymbol{\beta}}{}^{-1} + \sigma^{-2}\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X}, \quad \widetilde{\boldsymbol{A}} = \boldsymbol{B}_{\boldsymbol{\beta}}{}^{-1}\boldsymbol{A}_{\boldsymbol{\beta}} + \sigma^{-2}\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{Y}$$

  with $\boldsymbol{D} = \text{diag}(u_1^{-1}, \ldots, u_n^{-1})$.

- Sample σ^{-2} from $\text{Ga}(\tilde{a}_\sigma, \tilde{b}_\sigma)$, where

$$\tilde{a}_\sigma = a_\sigma + n/2, \quad \tilde{b}_\sigma = b_\sigma + \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i{}^\top \boldsymbol{\beta})^2/2u_i$$

- Sample z_i from Bernoulli distribution; the probabilities of $z_i = 0$ and $z_i = 1$ are proportional to $(1-s)\mathrm{N}(y_i; \boldsymbol{x}_i{}^\top\boldsymbol{\beta}, \sigma^2 u_{1i})$ and $s\mathrm{N}(y_i; \boldsymbol{x}_i{}^\top\boldsymbol{\beta}, \sigma^2 u_{2i})$, respectively.

- The full conditional distributions of s and γ are given by $\mathrm{Beta}(\tilde{a}_s, \tilde{b}_s)$ and $\mathrm{Ga}(\tilde{a}_\gamma, \tilde{b}_\gamma)$, respectively, where $\tilde{a}_s = a_s + \sum_{i=1}^n z_i$ and $\tilde{b}_s = b_s + n - \sum_{i=1}^n z_i$, $\tilde{a}_\gamma = a_\gamma + n$ and $\tilde{b}_\gamma = b_\gamma + \sum_{i=1}^n \log\{1 + \log(1+u_{2i})\}$.

- For each i, independently, sample u_{1i} from $\mathrm{GIG}(a + 1/2, 2a, (y_i - \boldsymbol{x}_i{}^\top\boldsymbol{\beta})^2/\sigma^2)$ if $z_i = 0$ or from $\mathrm{Ga}(a, a)$ if $z_i = 1$, where $\mathrm{GIG}(p, a, b)$ denotes the generalized Gaussian inverse distribution with the density of the form, $f(x) \propto x^{p-1}\exp\{-(ax + b/x)/2\}$.

- For each i, independently, sample (v_i, w_i) first in a compositional way; sample w_i from $\mathrm{Ga}(1+\gamma, 1+\log(1+u_{2i}))$ and $(v_i|w_i)$ as $\mathrm{Ga}(1+w_i, 1+u_{2i})$. Then, sample u_{2i} from $\mathrm{GIG}(1/2, 2v_i, (y_i - \boldsymbol{x}_i{}^\top\boldsymbol{\beta})^2/\sigma^2)$ if $z_i = 1$ or from $\mathrm{Ga}(1, v_i)$ if $z_i = 0$.

8.3.2 Efficiency in computation

A possible reason that the finite mixture has attracted less attention in the past research on posterior robustness is, as mentioned in Desgagné and Gagnon (2019), the increased number of latent state variables introduced by augmentation, and the concern for the potential inefficiency in posterior computation. It is the same concern seen in Bayesian variable selection (George and McCulloch (1993)); the finite mixture model for the prior on regression coefficients results in the necessity of stochastic search in the high-dimensional model space, hence causes the slow convergence of Markov chains and the costly computation. It is clear in the above algorithm, however, that the use of finite mixture as error distributions is completely different from the variable selection in terms of the model structure and free from such computational problem. Unlike the variable selection, the membership of each i to either of the two components in our model is independent of one another, which facilitates the stochastic search in 2^n possible combination of the model space. This fact also shows that the sampling of $(z_i, u_{1i}, u_{2i}, v_i, w_i)$ can be done completely in parallel across i's, hence our algorithm is scaled and computational feasible for the dataset with extremely large n.

We, again, emphasize that the use of the finite mixture is designed for controlling the effect of outliers on the other parameters of interest, and we focus on the inference for regression coefficients and scale parameter, not on the outlier detection. Although this view has already been clarified, and supported, by the posterior robustness in Theorem 8.2.1, we further discuss the utility of finite mixture approach by the extensive comparison with other models by the simulation study in Section 8.4.

8.3.3 Robust Bayesian variable selection with shrinkage priors

When the dimension of x_i is moderate or large, it is desirable to select a suitable subset of x_i to achieve efficient estimation. This procedure of variable selection would also be seriously affected by the possible outliers, by which we may fail to select suitable subsets of covariates. For a robust Bayesian variable selection procedure, we introduce shrinkage priors for regression coefficients. Here we rewrite the regression model to explicitly express an intercept term as $y_i = \alpha + \boldsymbol{x}_i{}^\top\boldsymbol{\beta} + \varepsilon_i$, and consider a normal prior $\alpha \sim \mathrm{N}(0, A_\alpha)$ with fixed hyperparameter $A_\alpha > 0$. For the regression coefficients $\boldsymbol{\beta}$, we consider a class of independent priors expressed as

a scale mixture of normals given by

$$\pi(\boldsymbol{\beta}) = \prod_{k=1}^{p} \int_0^{\infty} \mathrm{N}(\boldsymbol{\beta}_k; 0, \sigma^2\tau^2\xi_k) g(\xi_k) d\xi_k, \tag{8.3.1}$$

where $g(\cdot)$ is a mixing distribution, and τ^2 is an unknown global parameter that controls the strength of the shrinkage effects. Examples of the mixing distribution $g(\cdot)$ includes the exponential distribution leading to the Laplace prior of $\boldsymbol{\beta}$ (Bayesian Lasso, Park and Casella (2008)), and the half-Cauchy distribution for $\xi_k^{1/2}$ which results in the horseshoe prior (Carvalho et al. (2009, 2010)). The robustness property of the resulting posterior distributions is guaranteed for those shrinkage priors; Theorem 8.2.1 does not require any conditions other than the prior propriety.

In terms of posterior computation, the key property is that the conditional distribution of β_k given ξ_k under (8.3.1) is a normal distribution, so the sampler given in Section 8.3.1 is still valid with minor modification. Specifically, the sampling steps from the full conditional distributions of α, $\boldsymbol{\beta}$, σ^2 and $\xi_1, \ldots, \xi_p$ are modified or added as follows:

- Sample α from $\mathrm{N}(\widetilde{A}_\alpha{}^{-1}\widetilde{B}_\alpha, \widetilde{A}_\alpha{}^{-1})$, where

$$\widetilde{A}_\alpha = A_\alpha + \sigma^{-2}\sum_{i=1}^{n} u_i{}^{-1}, \quad \widetilde{B}_\alpha = \sigma^{-2}\sum_{i=1}^{n} u_i{}^{-1}(y_i - \boldsymbol{x}_i{}^\top\boldsymbol{\beta}).$$

- Sample $\boldsymbol{\beta}$ from $\mathrm{N}(\widetilde{\boldsymbol{A}}_{\boldsymbol{\beta}}{}^{-1}\boldsymbol{X}^\top\boldsymbol{D}\widetilde{\boldsymbol{Y}}, \sigma^2\widetilde{\boldsymbol{A}}_{\boldsymbol{\beta}}{}^{-1})$, where

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{Y} - \alpha\boldsymbol{j}^{(n)}, \quad \widetilde{\boldsymbol{A}}_{\boldsymbol{\beta}} = \boldsymbol{\Lambda}^{-1} + \boldsymbol{X}^\top\boldsymbol{D}\boldsymbol{X}, \quad \text{with} \quad \boldsymbol{\Lambda} = \tau^2\mathrm{diag}(\xi_1, \ldots, \xi_p).$$

- Sample σ^{-2} from $\mathrm{Ga}(\tilde{a}_\sigma, \tilde{b}_\sigma)$, where

$$\tilde{a}_\sigma = a_\sigma + (n+p)/2, \quad \tilde{b}_\sigma = b_\sigma + \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i{}^\top\boldsymbol{\beta})^2/2u_i + \boldsymbol{\beta}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta}.$$

- Sample ξ_k for each k and τ^2 from their full conditionals. Their densities are proportional to $\mathrm{N}(\beta_k; 0, \sigma^2\tau^2\xi_k)g(\xi_k)$ and $\pi(\tau^2)\prod_{k=1}^{p}\mathrm{N}(\beta_k; 0, \sigma^2\tau^2\xi_k)$, respectively, where $\pi(\tau^2)$ is a prior density for τ^2.

The full conditional distributions of α and $\boldsymbol{\beta}$ are familiar forms owing to the normal mixture representation of the EHE distribution as well as the shrinkage priors. The sampling of ξ_k and τ^2 depends on the choice of shrinkage priors, but the existing algorithms in the literature can be directly imported to our method.

In Section 8.5, we adopt the horseshoe prior for regression coefficients with the EHE distribution for the error terms. We here provide the details of sampling algorithm under the horseshoe model. The horseshoe prior assumes that $\sqrt{\xi_k} \sim \mathrm{C}^+(0,1)$ independently for $k = 1, \ldots, p$ and $\tau \sim \mathrm{C}^+(0,1)$, where $\mathrm{C}^+(0,1)$ is the standard half-Cauchy distribution with probability density function given by $p(x) = 2/\pi(1+x^2)$ for $x > 0$. Note that they admit hierarchical expressions given by $\xi_k|\lambda_k \sim \mathrm{IG}(1/2, 1/\lambda_k)$ and $\lambda_k \sim \mathrm{IG}(1/2, 1/2)$ for ξ_k, and $\tau^2|\nu \sim \mathrm{IG}(1/2, 1/\nu)$ and

$\nu \sim \mathrm{IG}(1/2, 1/2)$ for τ^2. Then, the full conditional distributions of ξ_k and τ^2 as well as the latent parameters λ_k and ν are given by

$$\xi_k|- \sim \mathrm{IG}\left(1, \frac{1}{\lambda_k} + \frac{\beta_k^2}{2\tau^2\sigma^2}\right), \qquad \lambda_k|- \sim \mathrm{IG}\left(1, 1 + \frac{1}{\xi_k}\right)$$
$$\tau^2|- \sim \mathrm{IG}\left(\frac{p+1}{2}, \frac{1}{\nu} + \frac{1}{2\sigma^2}\sum_{k=1}^{p}\frac{\beta_k^2}{\xi_k}\right), \qquad \nu|- \sim \mathrm{IG}\left(1, 1 + \frac{1}{\tau^2}\right).$$

8.4 Simulation Studies

We here carry out simulation studies to investigate the performance of the proposed method together with existing methods. We generated $n = 300$ observations from the linear regression model with $p = 20$ covariates, given by

$$y_i = \beta_0 + \sum_{k=1}^{p}\beta_k x_{ik} + \sigma\varepsilon_i, \quad i = 1, \ldots, n,$$

where $\beta_0 = 0.5, \beta_1 = \beta_4 = 0.3, \beta_7 = \beta_{10} = 2, \sigma = 0.5$ and the other coefficients are set to 0. Here the vector of covariates $(x_{i1}, \ldots, x_{ip})$ were generated from a multivariate normal distribution with zero mean vector and variance-covariance matrix having (k, ℓ)-element equal to $(0.2)^{|k-\ell|}$ for $k, \ell \in \{1, \ldots, p\}$. Regarding the error term, we adopted the following contamination structure:

$$\varepsilon_i \sim (1-\omega)\mathrm{N}(0, 1) + \omega\mathrm{N}(\mu, 1), \quad i = 1, \ldots, n,$$

where ω is the contamination ratio and μ is the location of outliers. We considered all the combinations of $\omega \in \{0, 0.05, 0.1\}$ and $\mu \in \{5, 10, 15, 20\}$, which leads to 9 scenarios in total since $\omega = 0$ with arbitrary μ leads to the same structures of ε_i, namely no contamination.

For the simulation dataset, we applied the robust regression methods with the EHE distribution, the LPTN distribution (Gagnon et al. (2020)), and t-distribution with ν degrees of freedom. When using the EHE distribution, we adopted a simple method with setting $\gamma = 1$ (denoted by EH), and the adaptive version with γ estimated (aEH) by assigning prior distribution. In the LPTN distribution, we need to specify the tuning parameter $\rho \in (2\Phi(1) - 1, 1) \approx (0.6827, 1)$, and adopted two cases, $\rho = 0.9$ and $\rho = 0.7$, denoted by LP1 and LP2, respectively. Regarding the t-distribution, we considered $\nu = 1$ corresponding to Cauchy distribution (denoted by C), $\nu = 3$ (T3) and an adaptive version by assigning a discrete prior for ν (denoted by aT). We also employed the standard normal distribution (denoted by N). We implemented all the methods in Bayesian ways by assigning prior distributions: $\beta_k \sim \mathrm{N}(0, 1000)$ and $\sigma^{-2} \sim \mathrm{Ga}(1, 1)$. Under the EHE distribution, t-distributions and normal distribution, we generated the posterior samples of β_k by Gibbs sampler. On the other hand, we generated posterior samples under the LPTN distribution by the random-walk Metropolis-Hastings algorithm as adopted in Gagnon et al. (2020), in which the step sizes were set to 0.05. For each model, we generated 3000 posterior samples after discarding the first 1000 posterior samples.

Based on the posterior samples, we computed posterior means as well as 95% credible intervals of β_k for $k = 1, \ldots, p$. The performance of the point and interval estimation was assessed using square root of mean squared errors (RMSE), coverage probabilities (CP) and average length (AL) based on 500 replications of the simulation, and these values were averaged over

$\beta_0, \ldots, \beta_p$. In addition, we evaluated the efficiency of the sampling schemes by computing the average of inefficient factors (IF) of the posterior samples.

In Table 8.1, we reported the values of these performance measures in 9 scenarios. When $\omega = 0$ (no outlier), as predicted, the normal distribution provides the most efficient result in all measures while the other methods are slightly inefficient. However, the proposed method (EH and aEH in the table) performs almost in the same way as the normal distribution. This is an empirical evidence that the efficiency loss of the EHE distribution is very limited owing to the normal component in the mixture. In the other robust methods, MSEs are slightly higher than the that of the normal distribution and CPs are smaller than the nominal level.

In the other scenarios, where outliers are incorporated in the data generating process, the performance of the normal distribution breaks down, and the robustness property is highlighted in the performance measures of the other models. In particular, the EHE distribution with fixed γ (EH) performs quite stably in both point and interval estimation. The adaptive version (aEH) also works reasonably well, but the performances is slightly worse at the cost of estimation of γ, thereby the estimation of γ may not be beneficial. The LPTN model with $\rho = 0.9$ (LP1) shows reasonable performance, but its CPs tend to be smaller than the nominal level. The other LPTN model with $\rho = 0.7$ (LP2) greatly worsens the accuracy of point estimation, implying the sensitivity of the choice of hyperparameter ρ to the posteriors. The other models (C, T and aT) also suffer from the larger MSE values, which might relate to the lack of posterior robustness under the t-distribution family. The results of interval estimation severely depend on the degree-of-freedom parameter, as the Cauchy and t_3-distributions produce too narrow/wide credible intervals.

In terms of computational efficiency, it is remarkable that the IF values of the EHE methods are small and comparable with those of the t-distribution methods, which shows the efficiency of the proposed Gibbs sampling algorithm. On the other hand, the IFs of the LPTN models are very large due to the use of Metropolis-Hastings algorithm. To obtain the reliable posterior analysis under the LPTN models, one needs to increase the number of iterations drastically, or to spend more effort tuning the step-size parameter. The performance of LPTNs is improved under the simpler settings of less predictors, $p = 10$, but the overall result of comparison of 8 models remains almost the same. See the Appendix for this additional experiment.

8.5 Real Data Examples

The posterior robustness of the proposed EHE distribution is demonstrated via the analysis of two real datasets: Boston housing data and diabetes data. The goal of statistical analysis here is the variable selection with $p = 29$ and $p = 64$ predictors in the presence of outliers. Our robustness scheme is a prominent part of such analysis by allowing the use of unbounded prior densities for strong shrinkage effect– specifically the horseshoe priors we discussed in Section 8.3.3– while protecting the posteriors from the potential outliers. The former dataset is suspected to be contaminated by some outliers, where the difference of the proposed EHE distribution and the traditional t-distribution is emphasized. In contrast, the latter dataset is free from extreme outliers, by which we discuss the possible efficiency loss caused by the use of EHE distributions.

In our examples, we consider robust Bayesian inference using the proposed method with taking account of variable selection, since the number of covariates is not small in two cases. Specifically, we employed the horseshoe prior as described in Section 8.3.3. For comparison, we

Table 8.1: Average values of RMSEs, CPs, ALs and IFs of the proposed extremely-heavy tailed distribution with γ fixed (EH) and estimated (aEH), log-Pareto normal distribution with $\rho = 0.9$ (LP1) and $\rho = 0.7$ (LP2), Cauchy distribution (C), t-distribution with 3 degrees of freedom (T3) and estimated degrees of freedom (aT), based on 500 replications in 9 combinations of $(100\omega, \mu)$. All values except for IFs are multiplied by 100.

	$(100\omega, \mu)$	EH	aEH	LP1	LP2	C	T3	aT	N
	(0, –)	6.25	6.26	6.61	7.92	7.76	6.70	6.48	6.25
	(5, 5)	6.99	7.60	7.07	8.22	8.04	7.17	7.42	10.68
	(10, 5)	9.09	8.63	8.82	9.46	8.32	8.27	9.63	15.73
	(5, 10)	6.53	6.77	6.76	8.03	7.85	6.85	7.14	18.56
RMSE	(10, 10)	7.03	7.54	7.08	8.27	7.98	7.30	9.73	29.20
	(5, 15)	6.58	6.74	6.79	8.15	7.88	6.84	7.00	26.76
	(10, 15)	6.99	7.26	7.02	8.32	7.90	7.09	10.07	43.70
	(5, 20)	6.50	6.63	6.70	8.02	7.78	6.75	6.90	35.56
	(10, 20)	6.94	7.12	6.96	8.29	7.79	6.94	10.19	58.22
	(0, –)	95.0	95.0	89.6	72.6	88.3	93.3	94.4	95.1
	(5, 5)	94.9	92.7	92.1	78.2	89.5	94.5	95.7	91.5
	(10, 5)	93.3	91.9	91.6	80.1	90.5	93.8	94.4	90.1
	(5, 10)	95.0	94.3	92.1	77.4	90.0	95.6	97.8	90.6
CP	(10, 10)	94.8	93.5	93.4	78.7	92.0	97.1	98.2	90.6
	(5, 15)	95.1	94.6	92.2	76.2	90.0	95.6	98.4	90.6
	(10, 15)	94.7	93.8	93.2	78.6	92.3	97.7	99.2	90.3
	(5, 20)	95.0	94.7	92.0	76.2	90.5	95.9	98.7	90.3
	(10, 20)	94.6	94.1	93.3	78.0	92.5	98.0	99.6	90.3
	(0, –)	24.6	24.6	23.0	18.5	24.6	24.6	25.0	24.6
	(5, 5)	27.6	27.5	26.1	21.7	26.2	27.7	30.4	36.3
	(10, 5)	31.7	30.6	31.1	24.9	28.1	31.9	37.2	44.2
	(5, 10)	25.8	26.0	25.1	20.6	26.1	27.8	33.9	58.6
AL	(10, 10)	27.3	27.8	27.4	22.1	28.0	32.6	49.1	77.3
	(5, 15)	25.8	25.9	25.1	20.3	26.1	27.9	35.9	83.1
	(10, 15)	27.1	27.3	26.9	22.1	27.9	32.8	60.1	113.3
	(5, 20)	25.6	25.7	24.8	20.2	26.0	27.7	37.2	109.2
	(10, 20)	27.0	27.1	26.7	21.7	27.9	32.9	69.4	149.4
	(0, –)	1.01	1.44	45.25	54.19	4.65	2.11	1.86	0.98
	(5, 5)	2.23	5.03	42.73	52.94	4.30	1.96	1.80	0.99
	(10, 5)	3.73	5.36	40.53	51.92	3.98	1.86	1.82	0.98
	(5, 10)	1.99	3.46	43.56	53.41	4.26	1.90	1.79	0.98
IF	(10, 10)	3.10	5.35	41.73	52.69	3.86	1.70	1.93	0.98
	(5, 15)	1.98	3.13	43.58	53.52	4.23	1.88	1.76	0.98
	(10, 15)	3.13	4.62	42.30	52.80	3.84	1.66	2.07	0.98
	(5, 20)	1.97	2.93	43.84	53.50	4.21	1.88	1.75	0.98
	(10, 20)	3.11	4.23	42.45	52.84	3.80	1.65	2.18	0.98

also applied Bayesian regression with the normal and t-error distribution, where the degrees of freedom is also estimated, while using the horseshoe prior for regression coefficients. In these three model, we assign the same prior distribution as in Section 8.4. Note that the horseshoe prior can be easily incorporated into the regression models with both normal and t-distribution, and efficient Gibbs sampling methods can be used. On the other hand, it is not straightforward to incorporate such priors into the robust method with the LPTN distribution, thereby we omitted it from the comparison. In all the methods, we generated 5000 posterior samples after discarding the first 2000 posterior samples as burn-in.

8.5.1 Boston housing data

We first consider the famous Boston housing dataset (Harrison and Rubinfeld (1978)). The response variable is the corrected median value of owner-occupied homes (in 1,000 USD). The covariates in the original datasets consist of 14 continuous-valued variables about the information of houses, such as longitude and latitude, and 1 binary covariate. After standardizing the continuous covariates, we also create squared values of those, which results in $p = 29$ covariates in our models. The sample size is $n = 506$.

To see the presence of outliers, we first applied a simple linear regression model to the dataset with Gaussian error distribution and compute standardized residuals, which are shown in the left panel of Figure 8.3. Large residuals in the figure imply the possible outliers in the dataset, which thereby affects the inference of regression coefficients and makes the analysis by the standard Gaussian regression model implausible.

In the proposed error distribution, the effect of possible outliers is reflected on the posterior of s, i.e., mixture proportion of the extremely heavy-tailed distribution. The trace plot of posterior samples of s under the EHE model is presented in the right panel of Figure 8.3. Since all the sampled values are bounded away from 0, it suggests that a certain proportion of the heavy-tailed distribution to take account of the outliers shown in the left panel. Other than the default prior $s \sim \text{Beta}(1, 1)$, we also applied slightly more informative priors, $\text{Beta}(1, 5)$ and $\text{Beta}(1, 9)$, based on the prior belief that s should be small, but the results were almost the same for all the parameters.

The posterior means and 95% credible intervals of the regression coefficients based on the three methods are shown in Figure 8.4. It shows that the results of the normal error model are quite different from those of t- and H-distributions. The difference of estimates becomes visually clear especially for the significant covariates– if we define the significance in the sense that the 95% credible intervals do not contain zero– as the result of proneness/sensitivity to the representative outliers observed in Figure 8.3. Comparing the models with the t- and H-distribution, they select the same set of covariates by significance, but the lengths of posterior credible intervals in the EHE model are shorter than those in the t-distribution model. In fact, the average interval lengths in the EHE and the t-distribution models are 1.01 and 1.13, showing the efficiency of the EHE model. This finding is consistent with the simulation results in Section 8.4.

8.5.2 Diabetes data

We next consider another famous dataset known as Diabetes data (Efron et al. (2004)). The data contains information of 442 individuals and 10 covariates regarding individual information

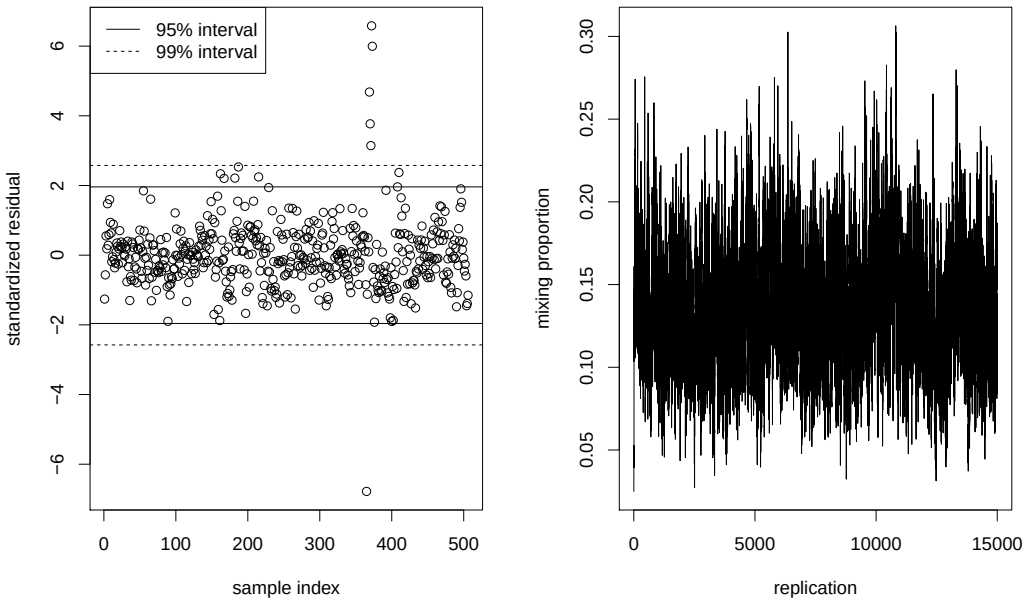

Figure 8.3: Standardized residuals (left) and trace plot of s (mixing proportion) in the proposed EHE distribution (right), obtained form the Boston housing data.

(age and sex) and several medical measures. We consider the same formulation of linear model as in Efron et al. (2004); the set of predictors consists of the original 10 variables, 10 main effects, 45 interactions, and 9 squared values, which results in $p = 64$ predictors in the model.

Similarly to the analysis of Boston housing data, we check the standardized residuals computed under the standard linear regression model, which was presented in the left panel of Figure 8.5. Few outliers are confirmed in the dataset as most of residuals are included in the 99% interval, which strongly supports the standard normal assumption in this example. In the main analysis by a regression models with horseshoe prior and three error distributions of normal, t- and EHE distributions, we generated 5000 posterior samples after discarding the first 2000 posterior samples as burn-in.

The right panel of Figure 8.5 shows the trace plot of posterior samples of s. All the sampled values are very close to zero, as expected from the residual plot in the left panel of Figure 8.5. For the small weight s is inferred from the data, the heavy-tailed component of the finite mixture is regarded "redundant" for this dataset. The same sensitivity analysis on the choice of priors for s is done as in the previous section, but we find no significant change to the results.

To see the possible inefficiency of using the EHE models for the dataset without outliers, the posterior means and 95% credible intervals of the regression coefficients are reported in Figure 8.6. The results of the three models are comparable; the predictors selected by significance are almost the same under the three models. The only notable difference is that the credible intervals produced by the t-distribution model is slightly larger than those of the other two methods. This indicates the loss of efficiency in using the t-distribution method under no outliers, as also confirmed in the simulation results in Section 8.4. In contrast, the difference

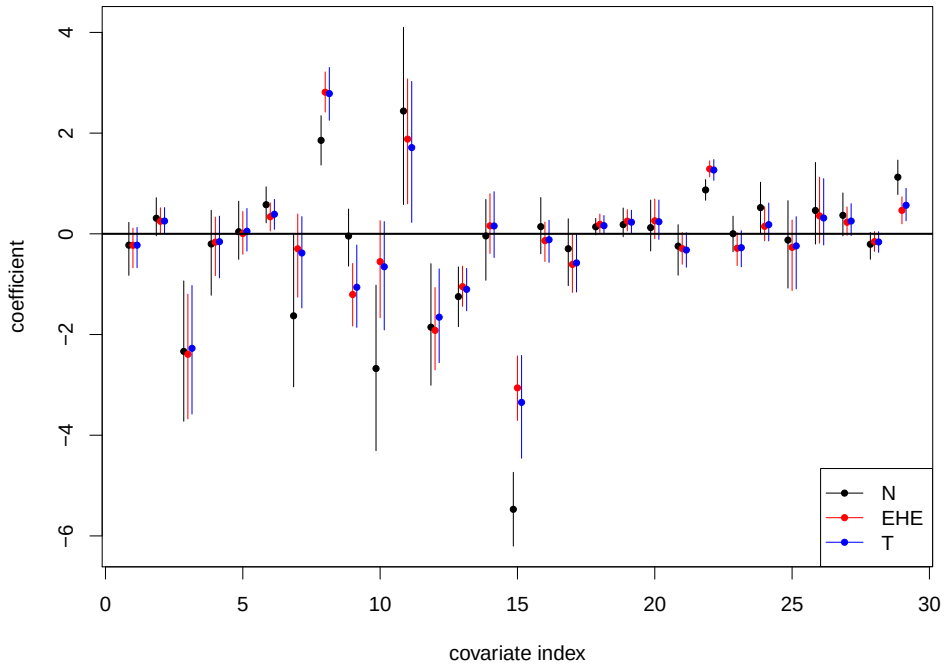

Figure 8.4: Posterior means and 95% credible intervals of the regression coefficients in the normal regression with normal distribution error (N), the proposed EHE distribution, and the t-distribution (T) with estimated degrees of freedom, applied to the Boston housing data.

in the credible intervals of the Gaussian and EHE models is hardly visible in the figure. We conclude from this finding that the choice of the EHE model is a safe option; even if no outlier exists, the efficiency loss in estimation is minimal.

8.6 Discussions

While we focused on the inference for the regression coefficients and scale parameter in this research, it is also of great interest to employ the predictive analysis based on the proposed model. Because H-distribution, as well as many log-regularly varying distributions, is too heavily-tailed to have finite moments, the posterior predictive mean under the EHE models do not exist. In predictive analysis, one needs to consider the posterior predictive medians or other alternatives for the point prediction. In uncertainty quantification, the second component of the EHE distribution could have a significant impact on the posterior predictive credible intervals for its heavy tails. In practice, it is important to monitor the posterior of mixing weight s to interpret the predictive analysis.

The use of the proposed method is not limited to the linear regression models, but can be immediately applied to other Gaussian models such as graphical models or state space models. Even under these highly-structured models, we are able to develop an efficient posterior compu-

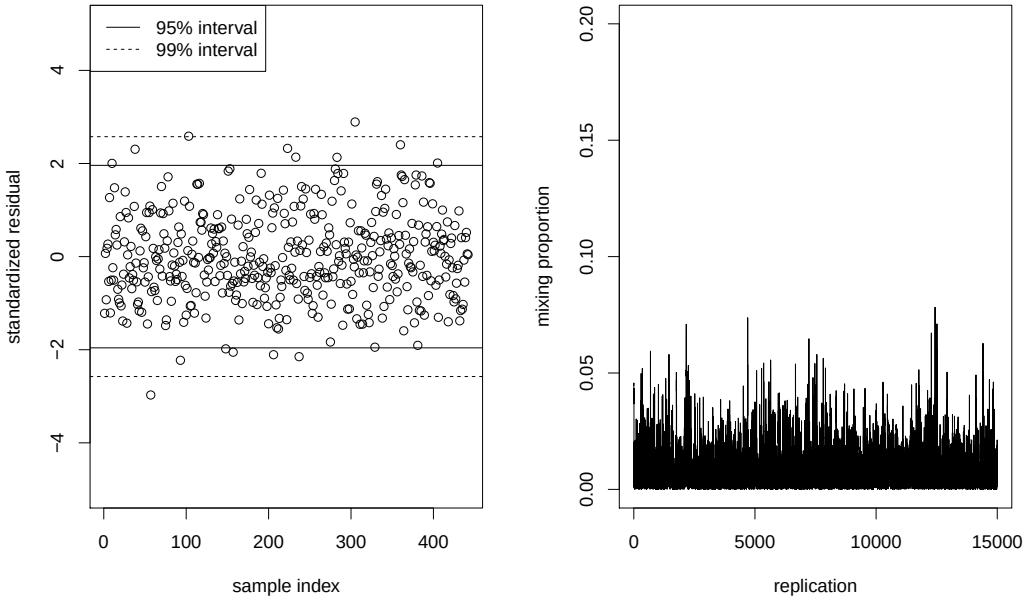

Figure 8.5: Standardized residuals (left) and trace plot of s (mixing proportion) in the proposed EHE distribution (right), obtained form the Diabetes data.

tation algorithm by utilizing the hierarchical representation of the proposed error distribution. The similar theoretical robustness properties may also be confirmed for those models.

8.7 Appendix

8.7.1 Lemmas

We provide two lemmas used in the proofs of Proposition 8.2.1 and Theorem 8.2.1.

Lemma 8.7.1 *Let $\alpha(\cdot)$ and $\beta(\cdot)$ be continuous, positive, and integrable functions defined on $(0, \infty)$. Suppose that $\lim_{u\to\infty} \beta(u)/\alpha(u) = \rho \in [0, \infty]$. Then*

$$\lim_{z\to\infty} \int_0^\infty \mathrm{N}(z|0,u)\beta(u)du \Big/ \int_0^\infty \mathrm{N}(z|0,u)\alpha(u)du = \rho.$$

Proof. We can assume that $\rho < \infty$; if $\rho = \infty$, then we can exchange the definitions of $\alpha(\cdot)$ and $\beta(\cdot)$, and this reduces to the case of $\rho = 0$. Let $\gamma(\cdot)$ be either $\alpha(\cdot)$ or $\beta(\cdot)$. We can also assume without loss of generality that $u^{-1/2}\alpha(u)$ and $u^{-1/2}\beta(u)$ are integrable. To see this, observe that, for any $\eta > 0$, there exist $\varepsilon > 0$ satisfying

$$0 \le \frac{\int_0^\varepsilon \mathrm{N}(1|0,u)\gamma(u)du}{\int_0^\infty \mathrm{N}(1|0,u)\gamma(u)du} < \eta/2$$

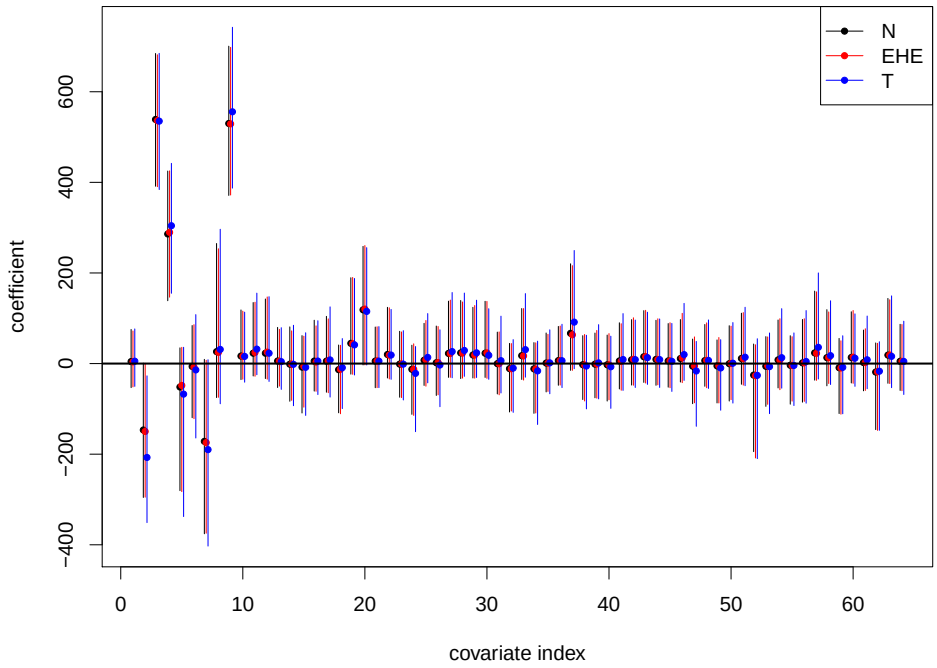

Figure 8.6: Posterior means and 95% credible intervals of the regression coefficients in the normal regression with normal distribution error (N), the proposed EHE distribution, and the t-distribution (T) with estimated degrees of freedom, applied to the Diabetes data.

and, for these η and ε, there also exists $\delta > 0$ such that $0 \leq 1 - e^{-\delta/\varepsilon} < \eta/2$. Hence, for all $z \geq 1$, the covariance inequality implies

$$\begin{aligned}
\frac{\int_0^\varepsilon \mathrm{N}(z|0,u)\gamma(u)du}{\int_0^\infty \mathrm{N}(z|0,u)\gamma(u)du} &= E[\chi_{(0,\varepsilon)}(U_z)] \\
&\leq \frac{E[\exp\{(z^2-1)/(2U_z)\}\chi_{(0,\varepsilon)}(U_z)]}{E[\exp\{(z^2-1)/(2U_z)\}]} \\
&= \frac{\int_0^\varepsilon \mathrm{N}(1|0,u)\gamma(u)du}{\int_0^\infty \mathrm{N}(1|0,u)\gamma(u)du}
\end{aligned}$$

where $\chi_{(0,\varepsilon)}(x)$ is the indicator function ($\chi_{(0,\varepsilon)}(x) = 1$ if $x \in (0,\varepsilon)$ and 0 otherwise) and the density of random variable U_z is proportional to $\mathrm{N}(z|0,u)\gamma(u)$. Finally, we have

$$\begin{aligned}
\left|\frac{\int_0^\infty \mathrm{N}(z|0,u)\gamma(u)e^{-\delta/u}du}{\int_0^\infty \mathrm{N}(z|0,u)\gamma(u)du} - 1\right| &\leq \frac{\int_0^\varepsilon \mathrm{N}(z|0,u)\gamma(u)du}{\int_0^\infty \mathrm{N}(z|0,u)\gamma(u)du} + \frac{\int_\varepsilon^\infty \mathrm{N}(z|0,u)\gamma(u)(1-e^{-\delta/u})du}{\int_\varepsilon^\infty \mathrm{N}(z|0,u)\gamma(u)du} \\
&\leq \frac{\int_0^\varepsilon \mathrm{N}(1|0,u)\gamma(u)du}{\int_0^\infty \mathrm{N}(1|0,u)\gamma(u)du} + 1 - e^{-\delta/\varepsilon} \\
&< \eta,
\end{aligned}$$

which shows the difference of $\gamma(u)$ and $e^{-\delta/u}\gamma(u)$ is ignorable in $u \to \infty$. This result verifies that, if $u^{-1/2}\gamma(u)$ is not integrable, then we can replace $\gamma(u)$ by $e^{-\delta/u}\gamma(u)$.

Again, assume $\rho < \infty$ and both $u^{-1/2}\alpha(u)$ and $u^{-1/2}\beta(u)$ are integrable. Let $M > 0$. Then we have

$$\begin{aligned}\Big|\frac{\int_0^\infty \mathrm{N}(z|0,u)\gamma(u)du}{\int_M^\infty \mathrm{N}(z|0,u)\gamma(u)du} - 1\Big| &\le \frac{\int_0^M \mathrm{N}(z|0,u)\gamma(u)du}{\int_{M+1}^\infty \mathrm{N}(z|0,u)\gamma(u)du} \\ &\le \Big\{\frac{e^{1/(M+1)}}{e^{1/M}}\Big\}^{z^2/2} \frac{\int_0^M u^{-1/2}\gamma(u)du}{\int_{M+1}^\infty u^{-1/2}\gamma(u)du} \\ &\to 0\end{aligned}$$

as $z \to \infty$ since $u^{-1/2}\gamma(u)$ is assumed to be integrable on $(0,\infty)$. Therefore,

$$\frac{\int_0^\infty \mathrm{N}(z|0,u)\beta(u)du}{\int_0^\infty \mathrm{N}(z|0,u)\alpha(u)du} \approx \frac{\int_M^\infty \mathrm{N}(z|0,u)\beta(u)du}{\int_M^\infty \mathrm{N}(z|0,u)\alpha(u)du} \tag{8.7.1}$$

as $z \to \infty$. Furthermore, uniformly in z,

$$\begin{aligned}\Big|\frac{\int_M^\infty \mathrm{N}(z|0,u)\beta(u)du}{\int_M^\infty \mathrm{N}(z|0,u)\alpha(u)du} - \rho\Big| &\le \frac{\int_M^\infty |\beta(u)/\alpha(u) - \rho|\mathrm{N}(z|0,u)\alpha(u)du}{\int_M^\infty \mathrm{N}(z|0,u)\alpha(u)du} \\ &\le \sup_{u>M}\Big|\frac{\beta(u)}{\alpha(u)} - \rho\Big| \\ &\to 0 \qquad (8.7.2)\end{aligned}$$

as $M \to \infty$ by assumption. Combining (8.7.1) and (8.7.2) gives the desired result. □

Lemma 8.7.2 *Let $M, v > 0$. Then we have*

$$(a) \qquad \frac{1+\log(1+M)}{1+\log(1+Mv)} \le \max\{1, v^{-1}\},$$

$$(b) \qquad \lim_{M\to\infty} \frac{1+\log(1+M)}{1+\log(1+Mv)} = 1.$$

Proof. The inequality in part (a) is trivial when $v \ge 1$; the left-hand-side is bounded by 1. For the case of $v < 1$, first observe that

$$\begin{aligned}\frac{1+\log(1+M)}{1+\log(1+Mv)} &= \exp\Big(\int_v^1 \Big[\frac{\partial}{\partial t}\log\{1+\log(1+Mt)\}\Big]dt\Big) \\ &= \exp\Big\{\int_v^1 \frac{1}{1+\log(1+Mt)}\frac{M}{1+Mt}dt\Big\}\end{aligned}$$

for all $v > 0$. Then it is immediate from this expression that

$$\frac{1+\log(1+M)}{1+\log(1+Mv)} \le \exp\Big(\int_v^1 \frac{1}{t}dt\Big) = v^{-1}$$

for $v < 1$. For part (b), we use the same expression to obtain

$$\begin{aligned}
&\lim_{M\to\infty} \frac{1+\log(1+M)}{1+\log(1+M/v)} \\
&= \exp\Big\{ \lim_{M\to\infty} \int_v^1 \frac{1}{1+\log(1+Mt)} \frac{M}{1+Mt} dt \Big\} \\
&= 1
\end{aligned}$$

by the dominated convergence theorem. □

8.7.2 Proof of Proposition 8.2.1

Here we prove Proposition 8.2.1. We show that

$$\lim_{|x|\to\infty} \frac{f_{\mathrm{EH}}(x)}{|x|^{-1}(\log|x|)^{-1-\gamma}} = A$$

for some constant $A > 0$. Since

$$\lim_{|x|\to\infty} \frac{\int_0^\infty \mathrm{N}(x;0,u)\mathrm{Ga}(u;a,a)du}{\int_0^\infty \mathrm{N}(x;0,u)H(u;\gamma)du} = 0$$

by Lemma 8.7.1, we can assume $s = 1$. Then we have for sufficiently large $|x|$

$$\begin{aligned}
\frac{f_{\mathrm{EH}}(x)}{|x|^{-1}(\log|x|)^{-1-\gamma}} &= \int_0^\infty \frac{\mathrm{N}(x;0,u)H(u;\gamma)}{|x|^{-1}(\log|x|)^{-1-\gamma}} du \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{u}} e^{-x^2/(2u)} \frac{\gamma|x|}{1+u} \Big\{ \frac{\log|x|}{1+\log(1+u)} \Big\}^{1+\gamma} du \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{v}} e^{-1/(2v)} \frac{\gamma x^2}{1+x^2 v} \Big\{ \frac{\log|x|}{1+\log(1+x^2 v)} \Big\}^{1+\gamma} dv,
\end{aligned}$$

where the last equality follows by making the change of variables $u = x^2 v$. Now, by part (a) of Lemma 8.7.2, the integrand is bounded by

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{v}} e^{-1/(2v)} \frac{\gamma}{v} \Big\{ \frac{\log|x|}{1+\log(1+x^2)} \frac{1+\log(1+x^2)}{1+\log(1+x^2 v)} \Big\}^{1+\gamma} \\
&\le \frac{\gamma}{\sqrt{2\pi}} \frac{e^{-1/(2v)}}{v^{3/2}} \Big(\frac{1}{2} \max\{1, v^{-1}\} \Big)^{1+\gamma} = \frac{\gamma/2^{1+\gamma}}{\sqrt{2\pi}} \frac{e^{-1/(2v)}}{v^{3/2}} \max\{1, v^{-(1+\gamma)}\} \\
&\le \frac{\gamma/2^{1+\gamma}}{\sqrt{2\pi}} \{ v^{-3/2} e^{-1/(2v)} + v^{-5/2-\gamma} e^{-1/(2v)} \},
\end{aligned}$$

where the right-hand side is an integrable function of $v \in (0,\infty)$ which does not depend on x. By part (b) of Lemma 8.7.2, the integrand converges to

$$\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{v}} e^{-1/(2v)} \frac{\gamma}{v} \Big\{ \lim_{|x|\to\infty} \frac{\log|x|}{1+\log(1+x^2)} \frac{1+\log(1+x^2)}{1+\log(1+x^2 v)} \Big\}^{1+\gamma} = \frac{\gamma/2^{1+\gamma}}{\sqrt{2\pi}} v^{-3/2} e^{-1/(2v)}$$

as $|x| \to \infty$ for each $v \in (0,\infty)$. Thus, by the dominated convergence theorem, we obtain

$$\lim_{|x|\to\infty} \frac{f_{\mathrm{EH}}(x)}{|x|^{-1}(\log|x|)^{-1-\gamma}} = \int_0^\infty \frac{\gamma/2^{1+\gamma}}{\sqrt{2\pi}} v^{-3/2} e^{-1/(2v)} dv = \frac{\gamma}{2^{1+\gamma}}.$$

This complete the proof.

8.7.3 Proof of Theorem 8.2.1

Let $\boldsymbol{y}_n = \mathcal{D}$ and $\boldsymbol{y}_k = \mathcal{D}^*$. Let $\mathcal{K} = \{i \in \{1, \ldots, n\} | y_i \in \boldsymbol{y}_k\}$ and $\mathcal{L} = \{1, \ldots, n\} \setminus \mathcal{K}$. Let

$$f_1(z) = \int_0^\infty \mathrm{N}(z|0, u) H(u; \gamma, \delta) du,$$
$$f_0(z) = \int_0^\infty \mathrm{N}(z|0, u) \mathrm{Ga}(u; a, a) du,$$

and $f(z) = s f_1(z) + (1 - s) f_0(z)$ for $z \in \mathbb{R}$, so that the ratio of $p(\boldsymbol{\beta}, \sigma | \boldsymbol{y}_n)$ to $p(\boldsymbol{\beta}, \sigma | \boldsymbol{y}_k)$ is

$$\begin{aligned}
\frac{p(\boldsymbol{\beta}, \sigma | \boldsymbol{y}_n)}{p(\boldsymbol{\beta}, \sigma | \boldsymbol{y}_k)} &= \frac{p(\boldsymbol{y}_k)}{p(\boldsymbol{y}_n)} \frac{\pi(\boldsymbol{\beta}, \sigma) \prod_{i=1}^n f((y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta})/\sigma)}{\pi(\boldsymbol{\beta}, \sigma) \prod_{i \in \mathcal{K}} f((y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta})/\sigma)} \\
&= \frac{p(\boldsymbol{y}_k) \prod_{i \in \mathcal{L}} f(y_i)}{p(\boldsymbol{y}_n)} \prod_{i \in \mathcal{L}} \frac{f((y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta})/\sigma)/\sigma}{f(y_i)} \\
&\propto \prod_{i \in \mathcal{L}} \frac{f((y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta})/\sigma)/\sigma}{f(y_i)}.
\end{aligned}$$

We prove that the right-hand side converges to $\sigma^{2|\mathcal{L}|\delta}$ uniformly in $(\boldsymbol{\beta}, \sigma) \in K$ as $\omega \to \infty$ for any nonempty compact set $K \subset \mathbb{R}^p \times (0, \infty)$. For this purpose, it is sufficient to show that

$$\frac{f((y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta})/\sigma)/\sigma}{f(y_i)} \to \sigma^{2\delta}$$

uniformly in $(\boldsymbol{\beta}, \sigma) \in K$ as $\omega \to \infty$ for every $i \in \mathcal{L}$. Fix $i \in \mathcal{L}$. Let $M = \sup_{(\beta, \sigma) \in K} |{\boldsymbol{x}_i}^\top \boldsymbol{\beta}| \in [0, \infty)$. Let $\underline{\sigma} = \inf_{(\boldsymbol{\beta}, \sigma) \in K} \sigma \in (0, \infty)$ and $\overline{\sigma} = \sup_{(\boldsymbol{\beta}, \sigma) \in K} \sigma \in (0, \infty)$. Assume without loss of generality that ω is sufficiently large so that $|y_i| \geq 2M + 1$.

We first consider the case of $s = 1$. Then

$$\begin{aligned}
\frac{f((y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta})/\sigma)/\sigma}{f(y_i)} &= \frac{f_1((y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta})/\sigma)/\sigma}{f_1(y_i)} \\
&= \frac{1}{\sigma} \frac{\int_0^\infty \mathrm{N}((y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta})/\sigma | 0, u) H(u; \gamma, \delta) du}{\int_0^\infty \mathrm{N}(y_i | 0, u) H(u; \gamma, \delta) du} \\
&= \frac{|y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta}|}{\sigma^2 |y_i|} \frac{\int_0^\infty v^{-1/2} e^{-1/(2v)} H((|y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta}|^2/\sigma^2) v | \gamma, \delta) dv}{\int_0^\infty v^{-1/2} e^{-1/(2v)} H(|y_i|^2 v | \gamma, \delta) dv},
\end{aligned}$$

where the last equality follows by making the change of variables $u = (|y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta}|/\sigma)^2 v$ in the numerator and by making the change of variables $u = |y_i|^2 v$ in the denominator. Therefore,

$$\left| \frac{f((y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta})/\sigma)/\sigma}{f(y_i)} - \sigma^{2\delta} \right| \leq \overline{\sigma}^{2\delta} \frac{\int_0^\infty v^{-1/2} e^{-1/(2v)} H(|y_i|^2 v | \gamma, \delta) G(v) dv}{\int_0^\infty v^{-1/2} e^{-1/(2v)} H(|y_i|^2 v | \gamma, \delta) dv},$$

where

$$\begin{aligned}
G(v) = G(v; \beta, \sigma, \gamma, \delta, y_i, x_i) &= \left| \frac{|y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta}|}{\sigma^{2(1+\delta)} |y_i|} \frac{H((|y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta}|^2/\sigma^2) v | \gamma, \delta)}{H(|y_i|^2 v | \gamma, \delta)} - 1 \right| \\
&= \left| \frac{|y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta}|}{|y_i|} \Big(\frac{1 + |y_i|^2 v}{\sigma^2 + |y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta}|^2 v} \Big)^{1+\delta} \Big[\frac{1 + \log(1 + |y_i|^2 v)}{1 + \log\{1 + (|y_i - {\boldsymbol{x}_i}^\top \boldsymbol{\beta}|^2/\sigma^2) v\}} \Big]^{1+\gamma} - 1 \right|
\end{aligned}$$

for $v > 0$. Note that

$$F_1(v) \le \frac{|y_i - \boldsymbol{x}_i{}^\top \boldsymbol{\beta}|}{|y_i|}\Big(\frac{1+|y_i|^2 v}{\sigma^2 + |y_i - \boldsymbol{x}_i{}^\top \boldsymbol{\beta}|^2 v}\Big)^{1+\delta}\Big[\frac{1+\log(1+|y_i|^2 v)}{1+\log\{1+(|y_i - \boldsymbol{x}_i{}^\top \boldsymbol{\beta}|^2/\sigma^2)v\}}\Big]^{1+\gamma} \le F_2(v),$$

where

$$F_1(v) = \frac{|y_i| - M}{|y_i|}\Big\{\frac{1+|y_i|^2 v}{\overline{\sigma}^2 + (|y_i|+M)^2 v}\Big\}^{1+\delta}\Big(\frac{1+\log(1+|y_i|^2 v)}{1+\log[1+\{(|y_i|+M)^2/\underline{\sigma}^2\}v]}\Big)^{1+\gamma},$$
$$F_2(v) = \frac{|y_i| + M}{|y_i|}\Big\{\frac{1+|y_i|^2 v}{\underline{\sigma}^2 + (|y_i|-M)^2 v}\Big\}^{1+\delta}\Big(\frac{1+\log(1+|y_i|^2 v)}{1+\log[1+\{(|y_i|-M)^2/\overline{\sigma}^2\}v]}\Big)^{1+\gamma}.$$

Then

$$G(v) \le |F_1(v) - 1| + |F_2(v) - 1|.$$

Therefore,

$$\begin{aligned}&\Big|\frac{f((y_i - \boldsymbol{x}_i{}^\top \boldsymbol{\beta})/\sigma)/\sigma}{f(y_i)} - \sigma^{2\delta}\Big| \\ &\le \overline{\sigma}^{2\delta}\frac{\int_0^\infty v^{-1/2}e^{-1/(2v)}\tilde{H}(v)\{|F_1(v)-1| + |F_2(v)-1|\}dv}{\int_0^\infty v^{-1/2}e^{-1/(2v)}\tilde{H}(v)dv},\end{aligned} \tag{8.7.3}$$

where

$$\tilde{H}(v) = \frac{H(|y_i|^2 v|\gamma,\delta)}{H(|y_i|^2|\gamma,\delta)}.$$

The right-hand side of (8.7.3) is independent of $(\boldsymbol{\beta}, \sigma)$. We have that $\lim_{\omega\to\infty}(|F_1(v)-1| + |F_2(v)-1|) = 0$ for each $v > 0$ and that for $|y_i| \ge 1$,

$$v^{-1/2}e^{-1/(2v)}\tilde{H}(v) = v^{-1/2}\Big(\frac{1+|y_i|^2}{1+|y_i|^2 v}\Big)^{1+\delta}\Big\{\frac{1+\log(1+|y_i|^2)}{1+\log(1+|y_i|^2 v)}\Big\}^{1+\gamma}e^{-1/(2v)}$$
$$\begin{cases}\le 2^{1+\delta}v^{-1/2-1-\delta}\max\{1, v^{-(1+\gamma)}\}e^{-1/(2v)} \\ \to v^{-1/2-1-\delta}e^{-1/(2v)} \quad \text{as } \omega\to\infty\end{cases}$$

for all $v > 0$ by Lemma 8.7.2. Furthermore,

$$|F_1(v)-1| + |F_2(v)-1| \le 2 + |F_1(v)| + |F_2(v)| \le 2\{1 + F_2(v)\}$$

and, since $|y_i| \ge 2M + 1 > M$, we have

$$\begin{aligned}F_2(v) &= \frac{|y_i|+M}{|y_i|}\Big\{\frac{1+|y_i|^2 v}{\underline{\sigma}^2 + (|y_i|-M)^2 v}\Big\}^{1+\delta}\Big(\frac{1+\log(1+|y_i|^2 v)}{1+\log[1+\{(|y_i|-M)^2/\overline{\sigma}^2\}v]}\Big)^{1+\gamma} \\ &\le 2\Big\{\frac{1}{\underline{\sigma}^2} + \frac{|y_i|^2}{(|y_i|-M)^2}\Big\}^{1+\delta}\Big(1 + \frac{\log\frac{1+|y_i|^2 v}{1+\{(|y_i|-M)^2/\overline{\sigma}^2\}v]}}{1+\log[1+\{(|y_i|-M)^2/\overline{\sigma}^2\}v]}\Big)^{1+\gamma} \\ &\le 2\Big(\frac{1}{\underline{\sigma}^2} + 4\Big)^{1+\delta}\Big[1 + \Big|\log\frac{1+|y_i|^2 v}{1+\{(|y_i|-M)^2/\overline{\sigma}^2\}v}\Big|\Big]^{1+\gamma},\end{aligned}$$

where

$$\begin{aligned}
&\Big|\log\frac{1+|y_i|^2v}{1+\{(|y_i|-M)^2/\overline{\sigma}^2\}v}\Big| \\
&= \Big|\int_{(|y_i|-M)^2/(|y_i|\overline{\sigma})^2}^{1}\frac{|y_i|^2v}{1+|y_i|^2vt}dt\Big| \le \int_{\min\{1,(|y_i|-M)^2/(|y_i|\overline{\sigma})^2\}}^{\max\{1,(|y_i|-M)^2/(|y_i|\overline{\sigma})^2\}}\frac{1}{t}dt \\
&\le \frac{\max\{1,(|y_i|-M)^2/(|y_i|\overline{\sigma})^2\}-\min\{1,(|y_i|-M)^2/(|y_i|\overline{\sigma})^2\}}{\min\{1,(|y_i|-M)^2/(|y_i|\overline{\sigma})^2\}} \\
&= \frac{|(|y_i|\overline{\sigma})^2-(|y_i|-M)^2|}{\min\{(|y_i|\overline{\sigma})^2,(|y_i|-M)^2\}} \le \frac{(|y_i|\overline{\sigma})^2}{(|y_i|-M)^2}+\frac{(|y_i|-M)^2}{(|y_i|\overline{\sigma})^2} \le (2\overline{\sigma})^2+(1/\overline{\sigma})^2.
\end{aligned}$$

Thus, by the dominated convergence theorem, the right-hand side of (8.7.3) converges to zero as $\omega\to\infty$.

Next we consider the case of $s\in(0,1)$. Then we have

$$\frac{f((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)/\sigma}{f(y_i)} = \frac{f_1((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)/\sigma}{f_1(y_i)}\frac{s+(1-s)\dfrac{f_0((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)}{f_1((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)}}{s+(1-s)\dfrac{f_0(y_i)}{f_1(y_i)}}.$$

Therefore,

$$\begin{aligned}
\Big|\frac{f((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)/\sigma}{f(y_i)}-\sigma^{2\delta}\Big| &\le \overline{\sigma}^{2\delta}\Big|\frac{f((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)/\sigma}{f(y_i)\sigma^{2\delta}}-1\Big| \\
&\le \overline{\sigma}^{2\delta}\Big[\Big\{\Big|\frac{f_1((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)/\sigma}{f_1(y_i)\sigma^{2\delta}}-1\Big|+1\Big\} \\
&\quad\times\Big\{\Big|\frac{s+(1-s)\dfrac{f_0((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)}{f_1((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)}}{s+(1-s)\dfrac{f_0(y_i)}{f_1(y_i)}}-1\Big|+1\Big\}-1\Big].
\end{aligned}$$

By the result for $s=1$,

$$\sup_{(\boldsymbol{\beta},\sigma)\in K}\Big|\frac{f_1((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)/\sigma}{f_1(y_i)\sigma^{2\delta}}-1\Big| \le \frac{1}{\underline{\sigma}^{2\delta}}\sup_{(\boldsymbol{\beta},\sigma)\in K}\Big|\frac{f_1((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)/\sigma}{f_1(y_i)}-\sigma^{2\delta}\Big|\to 0$$

as $\omega\to\infty$. On the other hand,

$$\Big|\frac{s+(1-s)\dfrac{f_0((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)}{f_1((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)}}{s+(1-s)\dfrac{f_0(y_i)}{f_1(y_i)}}-1\Big| \le \Big|\frac{s}{s+(1-s)\dfrac{f_0(y_i)}{f_1(y_i)}}-1\Big|+\frac{1-s}{s}\frac{f_0((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)}{f_1((y_i-\boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)}. \tag{8.7.4}$$

Since $\lim_{u\to\infty}\mathrm{Ga}(u|a,a)/H(u|\gamma,\delta)=0$,

$$\lim_{z\to\infty}\frac{f_0(z)}{f_1(z)} = \lim_{z\to\infty}\frac{\int_0^\infty \mathrm{N}(z|0,u)\mathrm{Ga}(u|a,a)du}{\int_0^\infty \mathrm{N}(z|0,u)H(u|\gamma,\delta)du}=0$$

by Lemma 8.7.1 and the first term on the right side of (8.7.4) converges to zero as $\omega \to \infty$. Since $f_0(z) = f_0(|z|)$ and $f_1(z) = f_1(|z|)$ are nonincreasing functions of $|z|$ and since $M \le |y_i|/2 \le |y_i|$, it follows that

$$\begin{aligned}\frac{f_0((y_i - \boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)}{f_1((y_i - \boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)} &\le \frac{f_0((|y_i| - M)/\overline{\sigma})}{f_1((|y_i| + M)/\underline{\sigma})} = \frac{f_0((|y_i| - M)/\overline{\sigma})}{f_1((|y_i| - M)/\overline{\sigma})}\frac{f_1((|y_i| - M)/\overline{\sigma})}{f_1((|y_i| + M)/\underline{\sigma})} \\ &\le \frac{f_0((|y_i| - M)/\overline{\sigma})}{f_1((|y_i| - M)/\overline{\sigma})}\frac{f_1(|y_i|/(2\overline{\sigma}))}{f_1(|y_i|/(\underline{\sigma}/2))},\end{aligned}$$

where

$$\lim_{\omega\to\infty} \frac{f_0((|y_i| - M)/\overline{\sigma})}{f_1((|y_i| - M)/\overline{\sigma})} = 0.$$

Furthermore,

$$\begin{aligned}\frac{f_1(|y_i|/(2\overline{\sigma}))}{f_1(|y_i|/(\underline{\sigma}/2))} &= \frac{\int_0^\infty \mathrm{N}(|y_i|/(2\overline{\sigma})|0, u)H(u;\gamma,\delta)du}{\int_0^\infty \mathrm{N}(|y_i|/(\underline{\sigma}/2)|0, u)H(u;\gamma,\delta)du} \\ &= \frac{\underline{\sigma}}{4\overline{\sigma}}\frac{\int_0^\infty \mathrm{N}(|y_i||0, v)H(v/(2\overline{\sigma})^2;\gamma,\delta)dv}{\int_0^\infty \mathrm{N}(|y_i||0, v)H(v/(\underline{\sigma}/2)^2;\gamma,\delta)dv} \\ &\to \Big(\frac{4\overline{\sigma}}{\underline{\sigma}}\Big)^{1+2\delta}\end{aligned}$$

as $\omega \to \infty$ by Lemma 8.7.1 since

$$\frac{H(v/(2\overline{\sigma})^2;\gamma,\delta)}{H(v/(\underline{\sigma}/2)^2;\gamma,\delta)} = \Big\{\frac{1 + v/(\underline{\sigma}/2)^2}{1 + v/(2\overline{\sigma})^2}\Big\}^{1+\delta}\Big[\frac{1 + \log\{1 + v/(\underline{\sigma}/2)^2\}}{1 + \log\{1 + v/(2\overline{\sigma})^2\}}\Big]^{1+\gamma} \to \Big(\frac{4\overline{\sigma}}{\underline{\sigma}}\Big)^{2(1+\delta)}$$

as $v \to \infty$ by Lemma 8.7.2. Thus, we conclude that

$$\sup_{(\boldsymbol{\beta},\sigma)\in K}\Big|\frac{f((y_i - \boldsymbol{x}_i{}^\top\boldsymbol{\beta})/\sigma)/\sigma}{f(y_i)} - \sigma^{2\delta}\Big| \to 0$$

as $\omega \to \infty$.

8.7.4 Additional experiment in simulation study

The LPTN models are estimated by the random-walk Metropolis-Hastings algorithm, which requires many iterations in posterior sampling for convergence. While keeping the fairness in the number of iterations, we conduct another experiment that favors the LPTN models by partly eliminating the convergence issue in the LPTN models. The additional simulation study is based on the same settings in Section 8.4, except that the number of predictors is now $p = 10$.

The results are summarized in Table 8.2. The IFs of the LPTN models are improved, but still significantly higher than the others. The LPTN model with $\rho = 0.9$ improves the accuracy of point and interval estimations and is now competitive with the proposed models, while the other LPTN model with $\rho = 0.7$ still provides interval estimates with lower coverage probabilities. This result illustrates the difficulty in tuning the hyperparameters in the class of LPTN distributions, which contrasts the proposed model with no hyperparameter that is sensitive to the posterior result.

Table 8.2: Average values of RMSE, CP and AL of the proposed extremely-heavy tailed distribution with fixed γ (EH) and estimated gamma (aEH), log-Pareto normal distribution with $\rho = 0.9$ (LP1) and $\rho = 0.7$ (LP2), Cauchy distribution (C), t-distribution with 3 degrees of freedom (T3) and estimated degrees of freedom (T), based on 500 replications in 9 combinations of $(100\omega, \mu)$ with $p = 10$. All values are multiplied by 100.

	$(100\omega, \mu)$	EH	aEH	LP1	LP2	C	T3	aT	N
RMSE	(0, –)	6.12	6.14	6.40	7.68	7.69	6.58	6.36	6.13
	(5, 5)	6.75	7.40	6.76	7.95	7.78	6.95	7.31	11.61
	(10, 5)	8.63	8.68	8.63	9.36	8.10	8.31	10.22	18.84
	(5, 10)	6.34	6.57	6.45	7.66	7.63	6.66	6.98	20.63
	(10, 10)	6.97	7.49	6.84	8.00	7.90	7.31	10.39	35.83
	(5, 15)	6.44	6.60	6.49	7.76	7.78	6.73	6.99	30.97
	(10, 15)	6.77	7.09	6.62	7.92	7.76	6.90	10.54	53.29
	(5, 20)	6.46	6.60	6.56	7.68	7.80	6.74	6.84	39.81
	(10, 20)	6.85	7.06	6.70	8.04	7.81	6.83	10.37	70.04
CP	(0, –)	94.8	94.8	93.0	84.6	87.6	92.7	94.4	94.8
	(5, 5)	94.9	92.1	94.3	86.0	89.2	94.2	95.4	88.1
	(10, 5)	93.4	90.8	92.1	86.6	90.2	92.9	92.9	85.9
	(5, 10)	95.3	94.3	94.6	86.7	89.9	95.8	97.6	86.4
	(10, 10)	94.1	92.6	94.6	87.4	91.1	96.8	97.6	86.1
	(5, 15)	94.8	93.7	94.1	86.6	88.9	95.0	98.0	86.3
	(10, 15)	94.6	93.7	94.7	87.1	91.6	97.6	98.7	86.6
	(5, 20)	94.4	94.1	93.4	86.4	89.6	95.0	98.4	86.4
	(10, 20)	93.8	93.1	94.2	85.7	90.5	97.4	99.4	86.3
AL	(0, –)	23.8	23.8	23.5	22.5	23.9	23.9	24.2	23.8
	(5, 5)	26.3	26.3	25.8	24.1	25.2	26.6	29.4	35.0
	(10, 5)	29.7	29.3	30.0	26.6	27.1	30.6	36.4	43.0
	(5, 10)	24.9	25.0	25.1	23.9	25.2	26.7	32.5	56.4
	(10, 10)	26.3	26.7	27.2	25.0	27.0	31.3	48.4	75.1
	(5, 15)	25.0	25.1	25.1	23.8	25.2	26.9	34.8	80.8
	(10, 15)	26.1	26.2	26.6	24.9	26.9	31.3	58.6	109.5
	(5, 20)	24.9	24.9	24.9	23.6	25.2	26.8	35.5	105.0
	(10, 20)	25.9	26.0	26.4	24.6	26.6	31.2	66.7	144.1
IF	(0, –)	1.02	1.56	28.10	40.69	4.35	2.10	1.85	0.98
	(5, 5)	2.36	5.33	27.47	39.81	4.06	1.97	1.83	0.98
	(10, 5)	4.21	5.91	27.77	38.83	3.79	1.87	1.88	0.98
	(5, 10)	2.17	3.77	27.66	40.14	4.02	1.89	1.82	0.98
	(10, 10)	3.53	5.70	27.28	38.91	3.71	1.71	2.02	0.98
	(5, 15)	2.20	3.47	27.68	40.19	3.98	1.89	1.80	0.97
	(10, 15)	3.54	5.02	27.37	39.38	3.68	1.68	2.13	0.97
	(5, 20)	2.16	3.20	27.75	40.42	4.01	1.89	1.80	0.98
	(10, 20)	3.51	4.68	27.39	39.65	3.63	1.67	2.22	0.98

Part IV

Conclusion

Chapter 9

Conclusion

In Part II of the thesis, we derived Bayesian shrinkage estimators and predictive density estimators and proved that they dominate usual procedures under suitable conditions. We first considered in Chapter 2 the problem of simultaneously estimating parameters of independent Poisson distributions in the presence of possibly unbalanced sample sizes. We broadened the class of shrinkage priors of Komaki (2015) to include the proper priors of Clevenson and Zidek (1975). By using Lemma 5 of Komaki (2015) to evaluate integrals, we obtained sufficient conditions under which the corresponding Bayes estimators dominate the ML estimator for the standardized squared error loss and which are applicable to priors not considered in Theorem 1 of Komaki (2015). We compared symmetric priors with asymmetric priors depending on sample sizes both analytically and through application to real data and saw that the former lead to heterogeneous estimators which shrink the ML estimator more toward the origin when sample sizes are smaller. Next, in Chapter 3, we considered the case of negative multinomial observations. We showed that empirical Bayes and hierarchical Bayes estimators of negative multinomial probability vectors dominate the UMVU estimator under suitable conditions. Since the denominator of each component of the UMVU estimator is not a constant in contrast to the Poisson case, additional complication arose in examining the risk function of the empirical Bayes estimator. We found that this complication can be overcome when the row dimension of the observation matrix is large enough. In order to obtain a generalized Bayes estimator which dominates the UMVU estimator, we introduced a class of hierarchical shrinkage priors for negative multinomial parameters constructed by imitating those for Poisson parameters as used by Clevenson and Zidek (1975) and Komaki (2015). Although we were mainly concerned with balanced cases, we utilized the method developed by Komaki (2015), who considered an unbalanced problem. Part (2) of Lemma 5 of Komaki (2015) has a direct counterpart in our negative multinomial case. On the other hand, it was an inequality that we used as a counterpart to the equality in part (1) of that lemma. In chapter 4, we went on to consider the problems of estimating negative multinomial parameter vectors and the joint predictive density of multinomial tables on the basis of observations of negative multinomial variables in unbalanced settings. We first obtained new conditions for empirical Bayes estimators to dominate the UMVU estimator and then showed that our hierarchical shrinkage priors are useful in deriving improved Bayesian predictive densities for multinomial observations. Predictive density estimation for the negative multinomial distribution was also discussed. Finally, in Chapter 5, we considered the predictive density estimation problem under the Kullback-Leibler divergence which corresponds to the classical and suggestive result of Stein's phenomenon for the estimation of a normal variance with

unknown mean. We provided a class of Bayesian shrinkage predictive densities and showed that they dominate the minimum risk equivariant predictive density under appropriate conditions.

In Part III of the thesis, we considered fully Bayesian posterior inference based on heavy-tailed distributions. First, in Chapter 6, we discussed global-local shrinkage priors for analyzing sequence of counts. We showed that the asymptotic bias of a Bayes estimator of a Poisson rate can be characterized by the tail behavior of the corresponding local prior. We obtained a general sufficient condition for tail-robustness. Then we proposed priors which satisfy the sufficient condition approximately or exactly and, in particular, introduced extremely heavy-tailed priors. Moreover, we introduced a novel augmentation approach using latent variables to develop an efficient posterior computation algorithm for Bayesian inference. We demonstrated the proposed methods through simulation and an application to a real dataset and observed the theoretically guaranteed tail-robustness property. The theoretical results are related to those in Part II. For example, when we consider the Poisson part in the hierarchical representation of the negative multinomial distribution as a likelihood, the extremely heavy-tailed priors can be viewed as special cases of the shrinkage priors for negative multinomial parameters we used in Part II. Augmentation approaches using latent gamma variables are also useful in the context of decision theory. Conversely, Properties (iii) and (iv) of Proposition 3.3.1 are analogous to the tail-robustness property considered in Chapter 6. Next, in Chapter 7, we broadened the class of the extremely heavy-tailed priors of Chapter 6 in order to achieve desirable shrinkage and robustness properties for the case of normal observations. The novel feature of the prior of our interest is its potential of further generalization, by which one may modify the proper prior "as robust as possible". We confirmed that the marginal density of our proposed prior has a spike around the origin so that our prior has a large shrinkage effect on noises. We proved the superiority of the proposed prior to existing ones explicitly via improvement of the mean squared error for a large signal. This theoretical property was supported by extensive simulation studies. Although our prior has an intractable normalizing constant, we showed that we can sample from the posterior distribution of a hyperparameter by using the accept-reject algorithm. Finally, in Chapter 8, we proposed a new approach to robust Bayesian linear regression by introducing the extremely heavy-tailed error distribution for the noise terms. More specifically, we considered the finite mixture of two components with thin and heavy tails as the error distribution and, for the heavily-tailed component, we used the novel class of distributions as considered in Chapters 6 and 7. Since both components are expressed as scale mixtures of normals, we can easily construct a simple Gibbs sampling for posterior inference. We proved the robustness to outliers of the posterior distributions under the proposed models. The improved performance of our model was shown in simulation and empirical studies.

Bibliography

[1] Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, **62**, 547–554.

[2] Andrade, J.A.A. and O'Hagan, A. (2006). Bayesian robustness modeling using regularly varying distributions. *Bayesian Analysis*, **1**, 169–188.

[3] Andrade, J.A.A. and O'Hagan, A. (2011). Bayesian robustness modelling of location and scale parameters. *Scandinavian Journal of Statistics*, **38**, 691–711.

[4] Armagan, A., Clyde, M. and Dunson, D.B. (2011). Generalized beta mixtures of Gaussians. In *Advances in neural information processing systems*, 523–531.

[5] Bai, R. and Ghosh, M. (2019). Large-scale multiple hypothesis testing with the normal-beta prime prior. *Statistics*, **53**, 1210–1233.

[6] Bai, R. and Ghosh, M. (2020). On the Beta Prime Prior for Scale Parameters in High-Dimensional Bayesian Regression Models. *Statistica Sinica, to appear.*

[7] Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, **8**, 716–76.

[8] Berry, L.R., Helman, P. and West, M. (2019). Probabilistic forecasting of heterogeneous consumer transaction-sales time series. *arXiv preprint arXiv:1808.04698.*

[9] Berry, L.R. and West, M. (2019). Bayesian forecasting of many count-valued time series. *Journal of Business and Economic Statistics*, to appear.

[10] Bhadra, A., Datta, J., Polson, N.G. and Willard, B.T. (2016). Default Bayesian analysis with global-local shrinkage priors. *Biometrika*, **103**, 955–969.

[11] Bhadra, A., Datta, J., Polson, N.G. and Willard, B.T. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, **12**, 1105–1131.

[12] Bhadra, A., Datta, J., Polson, N.G. and Willard, B.T. (2019). Lasso Meets Horseshoe: A Survey. *Statistical Science, to appear.*

[13] Bhattacharya, A., Pati, D., Pillai, N.S., and Dunson, D.B. (2015). Dirichlet-Laplace Priors for Optimal Shrinkage. *Journal of the American Statistical Association*, **110**, 1479–1490.

[14] Boisbunon, A. and Maruyama, Y. (2014). Inadmissibility of the best equivariant predictive density in the unknown variance case. *Biometrika*, **101**, 733–740.

[15] Box, G. and Tiao, G.C. (1968). A Bayesian approach to some outlier problems. *Biometrika*, **55**, 119–129.

[16] Brewster, J.F. and Zidek, J.V. (1974). Improving on equivariant estimators. *The Annals of Statistics*, **2**, 21–38.

[17] Brown, L.D., George, E.I., and Xu, X. (2008). Admissible predictive density estimation. *The Annals of Statistics*, **36**, 1156–1170.

[18] Brown, L.D., Greenshtein, E. and Ritov, Y. (2013). The Poisson compound decision problem revisited. *Journal of the American Statistical Association*, **108**, 741–749.

[19] Carter, C.K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, **81**, 541–553.

[20] Carvalho, C.M., Polson, N.G. and Scott, J.G. (2009). Handling Sparsity via the Horseshoe. In *AISTATS*, Volume 5, pp. 73–80.

[21] Carvalho, C.M., Polson, N.G., and Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.

[22] Chang, Y.-T. and Shinozaki, N. (2019). New types of shrinkage estimators of Poisson means under the normalized squared error loss. *Communications in Statistics - Theory and Methods*, **48**, 1108–1122.

[23] Chou, J.-P. (1991). Simultaneous estimation in discrete multivariate exponential families. *The Annals of Statistics*, **19**, 314–328.

[24] Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

[25] Clevenson, M.L. and Zidek, J.V. (1975). Simultaneous estimation of the means of independent Poisson laws. *Journal of the American Statistical Association*, **70**, 698–705.

[26] Datta, J. and Dunson, D.V. (2016). Bayesian inference on quasi-sparse count data. *Biometrika*, **103**, 971–983.

[27] Desgagné, A. (2015). Robustness to outliers in location–scale parameter model using log-regularly varying distributions. *The Annals of Statistics*, **43**, 1568–1595.

[28] Desgagné, A. and Gagnon, P. (2019). Bayesian robustness to outliers in linear regression and ratio estimation. *Brazilian Journal of Probability and Statistics*, **33**, 205–221.

[29] Devroye, L. (1981). The series method for random variate generation and its application to the Kolmogorov-Smirnov distribution. *American Journal of Mathematical and Management Sciences*, **1**, 359–379.

[30] Devroye, L. (2009). On exact simulation algorithms for some distributions related to Jacobi theta functions. *Statistics & Probability Letters*, **79**, 2251–2259.

[31] Dey, D. and Chung, Y. (1992). Compound Poisson distributions: Properties and estimation. *Communications in Statistics - Theory and Methods*, **21**, 3097–3121.

[32] Dey, D.K., Ghosh, M. and Srinivasan, C. (1987). Simultaneous estimation of parameters under entropy loss. *Journal of Statistical Planning and Inference*, **15**, 347–363.

[33] Efron, B. (2010). The future of indirect evidence. *Statistical Science*, **25**, 145–157.

[34] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least Angle regression. *The Annals of Statistics*, **32**, 407–499.

[35] Frühwirth-Schnatter, S. (2006). Finite mixture and Markov switching models. *Springer Science & Business Media*, **5**, 81–102.

[36] Gagnon, P., Desgagné, A., and Bédard, M. (2020). A new Bayesian approach to robustness against outliers in linear regression. *Bayesian Analysis*, **15**, 389–414.

[37] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, **1**, 515–534.

[38] George, E.I., Liang, F. and Xu, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *The Annals of Statistics*, **34**, 78–91.

[39] George, E.I., Liang, F. and Xu, X. (2012). From minimax shrinkage estimation to minimax shrinkage prediction. *Statistical Science*, **27**, 82–94.

[40] George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

[41] Ghosh, M., Hwang, J.T. and Tsui, K.-W. (1983). Construction of improved estimators in multiparameter estimation for discrete exponential families. *The Annals of Statistics*, **11**, 351–367.

[42] Ghosh, M. and Parsian, A. (1981). Bayes minimax estimation of multiple Poisson parameters. *Journal of Multivariate Analysis*, **11**, 280–288.

[43] Ghosh, M. and Yang, M.-C. (1988). Simultaneous estimation of Poisson means under entropy loss. *The Annals of Statistics*, **16**, 278–291.

[44] Griffin, J.E. and Brown, P.J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**, 171–188.

[45] Hamura, Y. (2020). Bayesian shrinkage approaches to unbalanced problems of estimation and prediction on the basis of negative multinomial samples. *arXiv preprint arXiv:2010.03141*.

[46] Hamura, Y., Irie, K. and Sugasawa, S. (2020a). On global-local shrinkage priors for count data. R&R for *Bayesian Analysis*. *arXiv preprint arXiv:1907.01333v2*.

[47] Hamura, Y., Irie, K. and Sugasawa, S. (2020b). Shrinkage with robustness: log-adjusted priors for sparse signals. *arXiv preprint arXiv:2001.08465v2*.

[48] Hamura, Y., Irie, K. and Sugasawa, S. (2020c). Log-regularly varying scale mixture of normals for robust regression. *arXiv preprint arXiv:2005.02800*.

[49] Hamura, Y. and Kubokawa, T. (2019a). Bayesian predictive distribution for a negative binomial model. *Mathematical Methods of Statistics*, **28**, 1–17.

[50] Hamura, Y. and Kubokawa, T. (2019b). Simultaneous estimation of parameters of Poisson distributions with unbalanced sample sizes. *Japanese Journal of Statistics and Data Science*, **2**, 405–435.

[51] Hamura, Y. and Kubokawa, T. (2020a). Bayesian predictive distribution for a Poisson model with a parametric restriction. *Communications in Statistics - Theory and Methods*, **49**, 3257–3266.

[52] Hamura, Y. and Kubokawa, T. (2020b). Bayesian shrinkage estimation of negative multinomial parameter vectors. *Journal of Multivariate Analysis*, **179**, 104653.

[53] Hamura, Y. and Kubokawa, T. (2020c). Proper Bayes minimax estimation of parameters of Poisson distributions in the presence of unbalanced sample sizes. *Brazilian Journal of Probability and Statistics*, **34**, 728–751.

[54] Hamura, Y. and Kubokawa, T. (2020d). Bayesian predictive density estimation for a chi-squared model using information from a normal observation with unknown mean and variance. *arXiv preprint arXiv:2006.07052v2*.

[55] Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics & Management*, **5**, 81–102.

[56] Hudson, H.M. (1978). A natural identity for exponential families with applications in multiparameter estimation. *The Annals of Statistics*, **6**, 473–484.

[57] Hwang, J.T. (1982). Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases. *The Annals of Statistics*, **10**, 857-867.

[58] Ishwaran, H. and Rao, J.S. (2015). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, **33**, 730–773.

[59] Kato, K. (2009). Improved prediction for a multivariate normal distribution with unknown mean and variance. *Annals of the Institute of Statistical Mathematics*, **61**, 531–542.

[60] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, **27**, 887–906.

[61] Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rule. *Journal of the American Statistical Association*, *109*, 674–685.

[62] Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika*, **88**, 859–864.

[63] Komaki, F. (2004). Simultaneous prediction of independent Poisson observables. *The Annals of Statistics*, **32**, 1744–1769.

[64] Komaki, F. (2006). A class of proper priors for Bayesian simultaneous prediction of independent Poisson observables. *Journal of Multivariate Analysis*, **97**, 1815–1828.

[65] Komaki, F. (2009). Bayesian predictive densities based on superharmonic priors for the 2-dimensional Wishart model. *Journal of Multivariate Analysis*, **100**, 2137–2154.

[66] Komaki, F. (2012). Asymptotically minimax Bayesian predictive densities for multinomial models. *Electronic Journal of Statistics*, **6**, 934–957.

[67] Komaki, F. (2015). Simultaneous prediction for independent Poisson processes with different durations. *Journal of Multivariate Analysis*, **141**, 35–48.

[68] Kubokawa, T. (1994). A unified approach to improving equivariant estimators. *The Annals of Statistics*, **22**, 290–299.

[69] Lawson, A.B. (2013). Bayesian disease mapping: hierarchical modeling in spatial epidemiology. Chapman and Hall/CRC.

[70] Lehmann, E.L. and Casella, G. (1998). Theory of Point Estimation, 2nd ed. (Springer, New York, 1998).

[71] Li, H., Graubardn, B.I. and Gail, M.H. (2010). Covariate Adjustment and Ranking Methods to Identify Regions with High and Low Mortality Rates. *Biometrics*, **66**, 613–620.

[72] Liang, F. and Barron, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*, **50**, 2708–2726.

[73] L'Moudden, A., Marchand, É., Kortbi, O. and Strawderman, W.E. (2017). On Predictive density estimation for Gamma models with parametric constraints. *Journal of Statistical Planning and Inference*, **185**, 56–68.

[74] Maruyama, Y. (1998). Minimax estimators of a normal variance. *Metrika*, **48**, 209–214.

[75] Maruyama, Y. and Strawderman W.E. (2005). A new class of generalized Bayes minimax ridge regression estimators. *The Annals of Statistics*, **33**, 1753–1770.

[76] Maruyama, Y. and Strawderman W.E. (2012). Bayesian predictive densities for linear regression models under α-divergence loss: Some results and open problems. In *IMS Collections, Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman*, D. Fourdrinier, É. Marchand & A. Rukhin, eds., vol. 8. Beachwood, USA: Institute of Mathematical Statistics, 42–56.

[77] Maruyama, Y. and Strawderman W.E. (2020a). Admissible Bayes equivariant estimation of location vectors for spherically symmetric distributions with unknown scale. *The Annals of Statistics*, **48**, 1052–1071.

[78] Maruyama, Y. and Strawderman W.E. (2020b). Admissible estimators of a multivariate normal mean vector when the scale is unknown. *arXiv preprint arXiv:2003.08571.*

[79] Morris, C.N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, **78**, 47-55.

[80] Muldoon, M.E. (1978). Some monotonicity properties and characterizations of the gamma function. *Aequationes Mathematicae*, **18**, 54–63.

[81] O'Hagan, A. (1979). On outlier rejection phenomena in Bayes inference. *Journal of the Royal Statistical Society: Series*, **41**, 358–367.

[82] O'Hagan, A. and Pericchi, L. (2012). Bayesian heavy-tailed models and conflict resolution: A review. *Brazilian Journal of Probability and Statistics*, **26**, 372–401.

[83] Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–686.

[84] Pérez, M., Pericchi, L.R., and Ramirez, I.C. (2017). The Scaled Beta2 distribution as a robust prior for scales. *Bayesian Analysis*, **12**, 615–637.

[85] Polson, N.G. (1991). A representation of the posterior mean for a location model. *Biometrika*, **78**, 426–430.

[86] Polson, N.G. and Scott, J.G. (2012a). Local shrinkage rules, Lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**, 287–311.

[87] Polson, N.G. and Scott, J.G. (2012b). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, **7**, 887–902.

[88] Polson, N.G., Scott, J.G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, **108**, 1339–1349.

[89] Robert, C.P. (1996). Intrinsic losses. *Theory and Decision*, **40**, 191–214.

[90] Seneta, E. (1976). Regularly varying functions. Springer-Verlag Berlin Heidelberg.

[91] Sibuya, M., Yoshimura, I. and Shimizu, R. (1964). Negative multinomial distribution. *Annals of the Institute of Statistical Mathematics*, **16**, 409–426.

[92] Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., and Sellers, W.R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

[93] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, **1**, University of California Press, 197–206.

[94] Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Annals of the Institute of Statistical Mathematics*, **16**, 155–160.

[95] Stoltenberg, E.A. and Hjort, N.L. (2019). Multivariate estimation of Poisson parameters. *Journal of Multivariate Analysis*, **175**, 1–19.

[96] Strawderman, W.E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Statistics*, **42**, 385–388.

[97] Tak, H., Ellis, J.A and Ghosh, S.K. (2019). Robust and Accurate Inference via a Mixture of Gaussian and Student's t Errors. *Journal of Computational and Graphical Statistics*, **28**, 415–426.

[98] Tang, X., Ghosh, M., Ha, N. and Sedransk, J. (2018). Modeling Random Effects Using Global–Local Shrinkage Priors in Small Area Estimation. *Journal of the American Statistical Association*, **113**, 1476–1489.

[99] Tsui, K.-W. (1979a). Multiparameter estimation of discrete exponential distributions. *The Canadian Journal of Statistics*, **7**, 193–200.

[100] Tsui, K.-W. (1979b). Estimation of Poisson means under weighted squared error loss. *The Canadian Journal of Statistics*, **7**, 201–204.

[101] Tsui, K.-W. (1984). Robustness of Clevenson-Zidek-type estimators. *Journal of the American Statistical Association*, **79**, 152–157.

[102] Tsui, K.-W. (1986a). Further developments on the robustness of Clevenson-Zidek-type means estimators. *Journal of the American Statistical Association*, **81**, 176–180.

[103] Tsui, K.-W. (1986b). Multiparameter estimation for some multivariate discrete distributions with possibly dependent components. *Annals of the Institute of Statistical Mathematics*, **38** 45–56.

[104] Tsui, K.-W. and Press, S.J. (1982). Simultaneous estimation of several Poisson parameters under K-normalized squared error loss. *TheAnnals of Statistics*, **10**, 93–100.

[105] van Dyk, D.A. and Park, T. (2019). Partially collapsed Gibbs samplers: Theory and methods *Journal of the American Statistical Association*, **103**, 790–796.

[106] Wakefield, J. (2006). Disease mapping and spatial regression with count data. Oxford University Press.

[107] West, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, **46**, 431–439.

[108] West, M. (1997). Modelling and robustness issues in Bayesian time series analysis (with discussion). *Institute for Mathematical Statistics*, **5**, 231–252.

[109] Womack, A. and Yang, Z. (2019). Heavy Tailed Horseshoe Priors. *arXiv preprint arXiv:1903.00928*.

[110] Yano, K., Kaneko, R. and Komaki, F. (2019). Exact Minimax Predictive Density for Sparse Count Data. *arXiv preprint arXiv:1812.06037*.

[111] Zhou, M. and Carin, L. (2013). Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 307–320.

[112] Zhu, A., Ibrahim, J.G. and Love, M.I. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*, **35**, 2084–2092.