Predictive Density Estimation in Nonparametric Statistical Models

博士論文 (要約)

Predictive Density Estimation in Nonparametric Statistical Models (ノンパラメトリック統計モデルにおける予測密度推定)

Keisuke Yano (矢野 恵佑)

Abstract

Prediction in nonparametric statistical models has garnered much attention regarding both theory and application. In the present thesis, we present four new contributions related to prediction in nonparametric statistical models.

First, we study asymptotically minimax predictive distributions in a function model. By attributing prediction in a function model to prediction in an infinite sequence model, we construct an asymptotically minimax predictive distribution for the case in which the parameter space is a known ellipsoid. We show that a Bayesian predictive distribution based on the Gaussian prior distribution is asymptotically minimax in the ellipsoid. We construct an asymptotically minimax predictive distribution for any Sobolev ellipsoid. We show that the Bayesian predictive distribution based on the product of Stein's priors is asymptotically minimax for any Sobolev ellipsoid. We present an efficient sampling method from the proposed Bayesian predicitve distribution.

Second, we consider refinements of estimators in nonparametric statistical models by focusing on the scale ratio of the parameter and noise. The refinements avoid a large risk for a fixed noise variance. Focusing on the scale ratio of the parameter and noise, we investigate the asymptotic risk in the case in which the ratio is large. The asymptotics as the ratio becomes large includes the asymptotics as the noise variance becomes small. We propose a prior distribution whose the Bayes estimator is minimax up to a logarithmic factor when the ratio is large and that satisfies a weak form of admissibility.

Third, we study weak admissibility and minimaxity under an estimative Kullback–Leibler risk in a Poisson sequence model. This study is motivated by estimation of an intensity function by using an estimative Kullback–Leibler risk in a Poisson point process model. We derive a minimax estimator and the corresponding exact minimax risk in a one-dimensional setting. We derive upper and lower bounds of the quantity related to weak admissibility.

Fourth, we study prediction when distributions of current and future observations might differ. We derive the asymptotic Kullback–Leibler risks of Bayesian predictive distributions for the case in which the numbers of current and future observations both grow to infinity. Based on these results, we construct model selection criteria for the case in which the true distributions of current and future observations might differ. Through numerical experiments, we show that Bayesian predictive distributions based on the proposed model selection criteria work well and that this study is extensible to prediction in a high-dimensional parametric model.

Contents

Chapter 1

Introduction

In this thesis, we discuss prediction in nonparametric statistical models. The thesis is based on Yano and Komaki (2016a,b,c).

Statistical prediction is important. In applications, the purpose of statistical analysis is often to forecast how future observations will behave based on current observations. Information regarding future observations is useful in decision-making. For example, from a weather forecast we obtain information regarding how the weather will change and then decide on a today's schedule based on the information. There is a vast literature on and application of prediction; see for example, Geisser (1993), Parzen et al. (1998), Cesa-Bianchi and Lugosi (2006), and Grünwald (2007) for details.

We discuss batch predictive density estimation. There are many kinds of prediction, including batch prediction, sequential prediction, point prediction, and predictive density estimation. Batch prediction is the prediction of the behavior of future observations one time and sequential prediction is the prediction of the behavior of future observations sequentially. Point prediction is the estimation of realized values of future observations based on current observations. Predictive density estimation is the estimation of a distribution of future observations based on current observations. Theoretical properties are dependent on the predictive setting; for sequential prediction, see Cesa-Bianchi and Lugosi (2006) and Grünwald (2007). For batch prediction, see Geisser (1993) and Parzen et al. (1998).

Predictive density estimation is more important than point prediction because a distribution of future observations contains full information. If we obtain an estimate of a distribution of future observations, we produce anything concerning future observations from the estimate. For example, predictive density estimation enables us to construct a predictive interval of future observations. Construction of a predictive interval is important in application. Consider horse racing as an example. In a horse race, we are interested in knowing not only which horse will win the race, but also knowing the winning percentages of all the horses.

Predictive density estimation in parametric models has been widely investigated. Akaike (1973) discussed model selection from the viewpoint of predictive density estimation. Dawid (1984) investigated a sequential version of predictive density estimation. Komaki (1996) and Hartigan (1998) compared Bayesian predictive densities with plugin predictive densities using the Kullback–Leibler divergence. Details of Bayesian predictive densities and plugin predictive densities are described in Chapter 2. Komaki (2001), George et al. (2006), and Brown et al. (2008) discussed the performance of Bayesian predictive densities based on shrinkage priors in finite dimensional Gaussian settings. Liang and Barron (2004), Aslan (2006), Komaki (2011), and Yano and Komaki (2014) investigated minimax predictive densities. Geisser also discussed minimax Bayesian predictive densities in a rejoinder to Bernardo (1979).

However, assumptions regarding parametric models are sometimes stringent. Considering model misspecification can relax such restrictions to some extent. Predictive density estimation using parametric models under misspecification has been discussed for example, by Takeuchi (1976), Sin and White (1996), Fushiki (2005), and Lv and Liu (2014). Recently, a further weak model misspecification assumption called local misspecification has been widely used, for example, by Shimodaira (1997), Claeskens and Hjort (2003), and Hjort and Claeskens (2003), because of its usability. Details of local misspecification are described in Chapter 5. However, considering only model misspecification is not sufficient to take the approximation capability of parametric models into account: in predictive density estimation, model misspecification takes only terms that relate to the (quasi-)distance from the distribution closest to the true distribution of future observations in the prepared model to a predictive density into consideration. There still remain terms that relate to the (quasi-)distance from the distribution closest to the true distribution of future observations among the prepared model to the true distribution of future observations.

Shifting our attention from parametric models to nonparametric models provides more flexible modeling and introduces new consideration into the approximation capability; Gaussian process regression, kernel density estimation, and estimation by using an orthogonal sequence are well known for providing flexibility to modeling. For details, see Wasserman (2007), Tsybakov (2009), Giné and Nickl (2016), and reference therein. Some reviews on nonparametric estimation are presented in Chapter 2. For new considerations regarding approximation capability, Shibata (1980, 1981) and Baraud (2000) derived the optimality of point prediction using parametric models from a nonparametric viewpoint. Such validation of parametric method from a nonparametric viewpoint is widely used; for example, see Grenander (1981) and Massart (2007).

A theory of predictive density estimation in nonparametric statistical models has been under construction, notwithstanding the importance of constructing the theory and of making its implementation practical. There are many studies on point prediction in nonparametric models; for example, see Steinwart and Christmann (2008), Hastie et al. (2009), Shibata (1980, 1981), Goldenshluger and Tsybakov (2003), Chapelle et al. (1999), and Cortes and Mohri (2007). However, little investigation has been conducted regarding predictive density estimation in nonparametric models because of its difficulty. Xu and Liang (2010), Xu and Zhou (2011), and Mukherjee and Johnstone (2015) discussed predictive density estimation in high-dimensional parametric models. High-dimensional parametric models are related to nonparametric models. In fact, Xu and Liang (2010) consider predictive density estimation in nonparametric regression models with equispaced designs by the approximation by high-dimensional parametric models.

In existing studies of predictive density estimation in nonparametric models, there are three problematic issues. First, there are no results that do not depend on the approximation of nonparametric models by high-dimensional parametric models. Approximation using high-dimensional parametric models affects the precision of predictive density. Using a nonparametric framework quantifies such an effect. Second, there are no results for the case in which a noise distribution is not Gaussian. There are many settings in which observations take discrete values instead of continuous values. Third, there are no results concerning predictive settings in which the distributions of current and future observations are different. Such settings are sometimes called extrapolation and appear frequently, as discussed in detail in Chapter 6.

In this thesis, we present four new contributions that advance the application of predictive density estimation in nonparametric models. The results involve studies concerning the above three issues. In Chapter 3, we construct an asymptotically optimal predictive density in a nonparametric model. In Chapter 4, we improve the estimators in nonpara-

metric models by focusing on the scale ratio of the parameter and the noise. The results in Chapters 3 and 4 concern the first issue. The results in Chapter 4 tell us that an estimator with approximation using high-dimensional parametric models sometimes shows poor precision, as discussed in the introduction to Chapter 4. In Chapter 5, we discuss estimators in non-homogeneous Poisson process models. The results in Chapter 5 concern the second issue. In Chapter 6, we consider predictive density estimation when the distributions of current and future observations are different. The results in Chapter 6 concern the third issue. In the remainder of the introduction, we present brief introductions to the results. A brief summary of the setting in each Chapter is provided in Table 1.1.

Table 1.1. Comparison of the setting in each chapter.

Chapter number	Prediction or estimation	Parametric or nonparametric	Noise
3	Prediction	Nonparametric	Gaussian
4	Estimation	Nonparametric	Gaussian
5	Estimation	Parametric	Poisson
6	Prediction	Parametric	General

In Chapter 3, we provide a theory of predictive density estimation in nonparametric models. We consider asymptotically optimal Bayesian predictive distributions in function models. We provide the statistical equivalence between prediction in function models and prediction in infinite sequence models. Using our results, we construct an asymptotically optimal Bayesian predictive distribution for the case in which the parameter space is a known ellipsoid. Further, using the product of Stein's priors based on the division of the parameter into blocks, we construct a more practical asymptotically optimal Bayesian predictive distribution for the case in which the parameter space is in the family of Sobolev ellipsoids. The construction is based on the parallelism between estimation and prediction. Finally, we provide some numerical experiments using efficient computational methods.

In Chapter 4, we investigate refinements of nonparametric methods focusing on the scale ratio of the parameter and the noise. Though the refinement is discussed through estimation in a Gaussian infinite sequence model because of its simplicity, the refinement is useful for predictive density estimation in nonparametric models because it is Bayesian and because prediction includes estimation, as discussed in Chapter 2. We present large scale-ratio asymptotics for nonparametric estimation in a Gaussian infinite sequence model for the case in which the parameter space is a Sobolev ellipsoid. We point out that typical Bayes estimators that are asymptotically optimal as the variance of the noise diminishes to zero suffer from excessive risk when the scale of the parameter is large. To resolve this issue, we define scale-ratio minimaxity and then construct the prior distribution of which the Bayes estimator is scale-ratio minimax up to a logarithmic factor without using knowledge of the scale of the parameter space. Further, we investigate the mass of the proposed prior on the parameter space and show that the Bayes estimator based on the proposed prior satisfies a weak form of admissibility on the parameter space. We consider an extension of the proposed prior to settings in which the smoothness of the true parameter is unknown. We also provide some numerical experiments related to our proposed prior distribution and its extension.

In Chapter 5, we investigate properties of estimation using the estimative Kullback–Leibler risk in a Poisson sequence model. Estimation in a Poisson sequence model is motivated by estimation of an intensity function in a non-homogeneous Poisson point process model. The estimative Kullback–Leibler risk is the expected Kullback–Leibler divergence from a true distribution to a plugin predictive distribution and has connections to the Kullback–Leibler risk for prediction. We derive an exact minimax estimator and

the corresponding risk and derive several bounds related to weak admissibility.

In Chapter 6, we provide predictive density estimation theory for extrapolation in parametric models. We consider prediction in the case in which the distributions of current and future observations might differ with an identical unknown finite-dimensional parameter. We show that Bayesian predictive distributions have lower risks than plugin predictive distributions, when both sample sizes of current and future observations simultaneously grow to infinity. The asymptotic form of the risk is different from that when the sample size of current distributions grows to infinity but the sample size of future observations is one. Based on the results, we propose a model selection criterion for predictive settings in which the true distributions of current and future observations might differ. Though the content focuses on prediction in parametric models, we show through numerical experiments that it is useful for prediction in a high-dimensional model, such as a sieve regression model. Through numerical experiments, we also show that our proposed model selection criterion works effectively.

The remainder of the thesis is organized as follows. In Chapter 2, we present some reviews on predictive density estimation in parametric models and on nonparametric estimation. In Chapter 7, we conclude the thesis. Each chapter has an introduction and a conclusion.

Chapter 2

Preliminaries

In this chapter, we present the backgrounds for this thesis. We present brief surveys of predictive density estimation in parametric models and of nonparametric models.

2.1 Predictive density estimation

Let $X^{(n)} = (X^1, \ldots, X^n)$ be n current observations distributed according to a probability distribution P_θ on a measurable space $\mathcal{X}$ and let $Y^{(m)} = (Y^1, \ldots, Y^m)$ be m future observations distributed according to a probability distribution Q_θ on $\mathcal{Y}$, where θ is included in Θ. When Θ is a subset of a d-dimensional Euclidean space with some $d \in \mathbb{N}$, we refer to this framework as prediction in parametric models. When Θ is not a subset of a d-dimensional Euclidean space with any $d \in \mathbb{N}$, we refer to this framework as prediction in nonparametric models. In this paper, we assume that $X^{(n)}$ and $Y^{(m)}$ are independent. We use the term "predictive model" for $\mathcal{M} := \{P_\theta \otimes Q_\theta : \theta \in \Theta\}$. Let p_θ and q_θ be the densities of P_θ and Q_θ with respect to some reference measure μ, respectively. The choice of the reference measure is provided in the individual chapters.

Remark 2.1. We assume independence only for theoretical simplicity. Relaxing this assumption is discussed by e.g., Ing and Wei (2005) and Tanaka and Komaki (2011).

Plugin and Bayesian predictive densities are introduced.

Definition 2.2. A plugin predictive density (distribution) $q_{\hat{\theta}}$ ($Q_{\hat{\theta}}$) base on an estimator $\hat{\theta}$ for θ is an estimator of q_θ (Q_θ) obtained by substituting $\hat{\theta}$ into θ of q_θ (Q_θ, respectively):

$$q_{\hat{\theta}}(y^{(m)}; x^{(n)}) = q_{\hat{\theta}(x^{(N)})}(y^{(M)}) \text{ for almost all } x^{(N)}, y^{(M)}.$$

Definition 2.3. A Bayesian predictive density (distribution) q_Π (Q_Π) based on a prior Π, a σ-finite measure on Θ, is an estimator of q_θ (Q_θ) obtained by averaging q_θ (Q_θ, respectively) by the posterior distribution $\Pi(\cdot|X^{(n)})$ based on Π:

$$q_\Pi(y^{(m)}|x^{(n)}) = \int q_\theta(y^{(m)}) \mathrm{d}\Pi(\theta|x^{(n)}) \text{ for almost all } y^{(m)}$$

and

$$Q_\Pi(A) = \int Q_\theta(A) \mathrm{d}\Pi(\theta|x^{(n)}) \text{ for all measurable } A \subset \mathcal{Y}$$

for almost all $x^{(n)} \in \mathcal{X}$.

2.1.1 Decision theory for predictive density estimation

We discuss the properties of these predictive densities from the viewpoint of statistical decision theory. Let $\mathcal{A}$ be a set of all probability distributions on $\mathcal{Y}$. Let $L(\cdot,\cdot) : \Theta \times \mathcal{A} \to \mathbb{R}_+ \cup \{\infty\}$ be the Kullback–Leibler loss: for all $Q \in \mathcal{A}$ and all $\theta \in \Theta$, if Q_θ is absolutely continuous with respect to Q, then

$$L(\theta, Q) := \int \log \frac{\mathrm{d}Q_\theta}{\mathrm{d}Q}(y^{(M)})\mathrm{d}Q_\theta(y^{(M)})$$

and otherwise $L(\theta, Q) = \infty$. Let $\mathcal{D}$ be $\{\hat{Q} : \mathcal{X} \to \mathcal{A}\}$. Let $R(\cdot,\cdot) : \Theta \times \mathcal{D} \to \mathbb{R}_+ \cup \{\infty\}$ be $\int L(\theta, \hat{Q}(\cdot; x^{(n)}))\mathrm{d}P_\theta$.

We use the Kullback–Leibler loss as a loss function because it has invariance with respect to a measurable one-to-one map: if Φ is a measurable one-to-one map from $\mathcal{X}$ to some measurable space $\mathcal{X}'$, then

$$L(\theta, Q) = \int \log \frac{\mathrm{d}Q_\theta \circ \Phi^{-1}}{\mathrm{d}Q \circ \Phi^{-1}}\mathrm{d}Q_\theta \circ \Phi^{-1},$$

where $Q_\theta \circ \Phi^{-1}$ and $Q \circ \Phi^{-1}$ are induced probability distributions of Q_θ and Q by Φ, respectively. It is also because it has connections to information theory: in fact, the loss function is the Kullback–Leibler divergence from Q_θ to Q.

Remark 2.4. We could use other (quasi-)distances satisfying the above requirements as loss functions; for such choices, see Corcuera and Giummolè (1999), Suzuki and Komaki (2010), Maruyama and Strawderman (2012), and Chang and Strawderman (2014). In particular, we could use the Hellinger distance

$$H(Q, Q_\theta) = \frac{1}{2}\int (\sqrt{\mathrm{d}Q_\theta/\mathrm{d}\lambda} - \sqrt{\mathrm{d}Q/\mathrm{d}\lambda})^2 \mathrm{d}\lambda$$

between the true distribution Q_θ and a predictive distribution Q, where λ is some reference measure. Though analysis of Hellinger distance is easy in a product type structure, it is difficult in general.

Remark 2.5. If we are interested in constructing predictive intervals, then the upper bound of the Kullback–Leibler loss is helpful. In fact, the Pinsker inequality yields

$$\|Q_\theta - \hat{Q}(\cdot; X^{(n)})\|_{\mathrm{TV}} \le \frac{1}{\sqrt{2}}\sqrt{L(\theta, \hat{Q}(\cdot; X^{(n)}))},$$

where $\|Q - \widetilde{Q}\|_{\mathrm{TV}}$ is the total variation between two probability distributions Q and $\widetilde{Q}$ on $\mathcal{Y}$:

$$\|Q - \widetilde{Q}\|_{\mathrm{TV}} := \sup_{A \subset \mathcal{Y}:\text{measurable}} |Q(A) - \widetilde{Q}(A)|.$$

Thus, for any measurable set A,

$$-\frac{1}{\sqrt{2}}\sqrt{L(\theta, \hat{Q}(\cdot; X^{(n)}))} \le Q_\theta(A) - \hat{Q}(A; X^{(n)}) \le \frac{1}{\sqrt{2}}\sqrt{L(\theta, \hat{Q}(\cdot; X^{(n)}))}.$$

If the Kullback–Leibler loss diminishes with respect to n in some sense, then a predictive interval based on the predictive distribution is consistent with a predictive interval based on the true distribution.

Optimality criteria in decision theory are introduced.

Definition 2.6 (Admissibility; e.g.,Lehmann and Casella (2003))**.** A predictive distribution $\hat{Q}$ is said to be admissible if there exists no predictive distribution $\widetilde{Q}$ such that $R(\theta, \widetilde{Q}) \leq R(\theta, \hat{Q})$ for all $\theta \in \Theta$ and $R(\theta, \widetilde{Q}) < R(\theta, \hat{Q})$ for some $\theta \in \Theta$.

Definition 2.7 (Minimaxity; e.g.,Lehmann and Casella (2003))**.** A predictive distribution $\hat{Q}$ is said to be minimax if

$$\sup_{\theta \in \Theta} R(\theta, \hat{Q}) = \inf_{Q \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, Q).$$

In nonparametric models, it is essentially difficult to attain exact admissibility and exact minimaxity. In such a setting, weak concepts of admissibility and minimaxity are useful.

Definition 2.8 (Weak admissibility; e.g., Ferguson (1967))**.** A predictive distribution $\hat{Q}$ is said to be $\gamma > 0$-admissible, if there exists no predictive distribution $\widetilde{Q}$ such that $R(\theta, \widetilde{Q}) < R(\theta, \hat{Q}) - \gamma$ for all $\theta \in \Theta$.

Definition 2.9 (Asymptotic minimaxity; e.g., Tsybakov (2009))**.** A predictive distribution $\hat{Q}$ is said to be asymptotically minimax, if the asymptotic equality

$$\lim_{n \to \infty} \left[\sup_{\theta \in \Theta} R(\theta, \hat{Q}) \Big/ \inf_{Q \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, Q) \right] = 1$$

holds.

Remark 2.10. There are many asymptotics in prediction concerning the relationship between the numbers n and m. For example, one-step ahead prediction is a setting in which n grows to infinity and m is fixed to one. Multi-step ahead prediction is a setting in which both n and m simultaneously grow to infinity.

There is a useful criterion for ensuring admissibility.

Theorem 2.11 (e.g.,Lehmann and Casella (2003))**.** *A unique Bayes solution $\hat{Q}(\cdot; X^{(n)})$, a unique minimizer with respect to $\hat{Q}$ of the Bayes risk $\int R(\theta, \hat{Q}) \mathrm{d}\Pi(\theta)$ based on a proper prior distribution Π up to a null set of the true distribution P_θ, is admissible*

In our setting, a Bayesian predictive distribution is a Bayes solution, and then the Bayesian predictive distribution based on a proper prior distribution is admissible when it is unique.

Theorem 2.12 (Aitchison (1975))**.** *If Π is a proper distribution on Θ, then Q_Π is a Bayes solution with respect to the Kullback–Leibler risk:*

$$\int R(\theta, Q_\Pi) \mathrm{d}\Pi(\theta) = \inf_{\hat{Q} \in \mathcal{D}} \int R(\theta, \hat{Q}) \mathrm{d}\Pi(\theta).$$

2.1.2 Prediction includes estimation

To restrict a predictive distribution to a narrow class, prediction contains estimation. Note that $\{Q_\theta : \theta \in \Theta\} \subset \mathcal{A}$. Then, letting $L_{\mathrm{e}} : \Theta \times \Theta \to \mathbb{R}_+ \cup \{\infty\}$ be $L_{\mathrm{e}}(\theta, \tilde{\theta}) = L(\theta, Q_{\tilde{\theta}})$ implies that prediction contains estimation. We call the risk R_{e} the estimative Kullback–Leibler risk. In Gaussian settings, the loss function L_{e} is a constant multiple of the mean squared loss and the corresponding risk R_{e} is a constant multiple of the mean squared risk.

2.1.3 Existing results in predictive density estimation

The following fundamental asymptotic results in parametric models were obtained by Komaki (1996) and Hartigan (1998).

Theorem 2.13 (Komaki (1996) and Hartigan (1998))**.** *Assume that* $X^1, \ldots, X^n, Y^1, \ldots, Y^m$ *are independent and identically distributed. Then, for any* $\theta \in \Theta$*, the Kullback–Leibler risk of any Bayesian predictive distribution based on a prior* Π *satisfying some regularity condition is smaller than that of any plugin predictive distribution based on* $\hat{\theta}$ *in the second-order term with respect to* n*:*

$$\lim_{n\to\infty} n^2 \left[R(\theta, Q_\Pi) - \frac{d}{n} \right] \leq \lim_{n\to\infty} n^2 \left[R(\theta, Q_{\hat{\theta}}) - \frac{d}{n} \right].$$

For details of the regularity condition, see Hartigan (1998).

2.2 Estimation in nonparametric models

Estimation in nonparametric models has been widely investigated. Some basic existing results are provided in this section.

2.2.1 Basic nonparametric models

Four basic nonparametric models are introduced.

Example 2.14 (White noise model)**.** Let $X(\cdot) = (X(t))_{t\in[0,1]}$ be an observation from

$$\mathrm{d}X(t) = f(t)\mathrm{d}t + \varepsilon \mathrm{d}W(t) \text{ for } t \in [0,1],$$

where $f : [0,1] \to \mathbb{R}$ is an unknown L_2-function and $W(\cdot) = (W(t))_{t\in[0,1]}$ is a standard Brownian motion.

Example 2.15 (Infinite sequence model)**.** Let $X = (X_i)_{i\in\mathbb{N}}$ be an observation from

$$X_i = \theta_i + \varepsilon W_i \text{ for } i \in \mathbb{N},$$

where $\theta = (\theta_i)_{i\in\mathbb{N}}$ is an unknown l_2-sequence and $W = (W_i)_{i\in\mathbb{N}}$ is a random sequence from $\otimes_{i=1}^{\infty}\mathcal{N}(0,1)$.

Example 2.16 (Nonparametric regression model using an equispaced design)**.** Let $X^{(n)}$ be an observation from

$$X^{(n)} = \begin{pmatrix} f(t_1) \\ \cdots \\ f(t_n) \end{pmatrix} + W^{(n)},$$

where $t_i := i/n$ for $i \in \{1, \ldots, n\}$, $f : [0,1] \to \mathbb{R}$ is an unknown L_2-function and $W^{(n)}$ is a random variable from $\otimes_{i=1}^{n}\mathcal{N}(0,1)$.

Example 2.17 (Nonparametric regression model using a random design)**.** Let $(X^{(n)}, T^{(n)})$ be an observation from

$$T^{(n)} = (T_1, \ldots, T_n) \sim \otimes_{i=1}^{n} \mathrm{Unif.}[0,1] \text{ for } i = 1, \ldots, n$$

and from

$$X^{(n)} = \begin{pmatrix} f(T_1) \\ \cdots \\ f(T_n) \end{pmatrix} + W_{(n)},$$

where $f : [0,1] \to \mathbb{R}$ is an unknown L_2-function and $W_{(n)}$ is a random variable from $\otimes_{i=1}^{n} \mathcal{N}(0,1)$.

Remark 2.18. In a nonparametric regression model using a random design, the assumption that the distribution of $T^{(n)}$ is the product of the uniform distributions is relaxed to the assumption that the distribution of $T^{(n)}$ has a known density with respect to a Lebesgue measure that is bounded below and above regardless of n.

Remark 2.19 (Sieve regression model)**.** In place of a nonparametric regression model, the following model, called a sieve regression model, is often used. Let $X^{(n)}$ be an observation from

$$X^{(n)} = \begin{pmatrix} \phi_1(t_1) & \cdots & \phi_{d_n}(t_1) \\ \phi_1(t_2) & \cdots & \phi_{d_n}(t_2) \\ \cdots & \cdots & \cdots \\ \phi_1(t_n) & \cdots & \phi_{d_n}(t_n) \end{pmatrix} \begin{pmatrix} \theta_1 \\ \cdots \\ \theta_{d_n} \end{pmatrix} + W^{(n)},$$

where $\{\phi_i\}_{i \in \mathbb{N}}$ is a system and d_n is a cut-off dimension. See e.g., Bontemps (2011).

The fundamental results of equivalence between these nonparametric models were obtained by, e.g., Brown and Low (1996) and Nussbaum (1996). In the following lemma, ε in the white noise and infinite-sequence models is chosen to be $1/\sqrt{n}$.

Theorem 2.20 (e.g., Brown and Low (1996) and Nussbaum (1996))**.** *Under some regularity condition on the parameter space, the above four statistical models are asymptotically statistically equivalent in the sense of Le Cam as $n \to \infty$.*

The statement of statistical equivalence in the sense of Le Cam and its proof are very complicated and are omitted. The important point is that these statistical models are strongly related in the sense that statistical solutions in some model can have corresponding solutions in any of the above models.

If we use the mean integrated squared loss as the loss function, then the equivalence result is simply derived from the L_2-isometry. We provide a sketch for the equivalence of the white noise and infinite sequence models.

Set ε in the white noise model and infinite sequence models to $1/\sqrt{n}$. Let $\{\phi_i(\cdot)\}_{i \in \mathbb{N}}$ be any orthonormal basis in L_2. By the Parseval identity, the correspondence between the true function $f(\cdot)$ and the true sequence θ is

$$f(\cdot) = \sum_{i=1}^{\infty} \theta_i \phi_i(\cdot)$$

and the correspondence between the estimator $\hat{f}(\cdot)$ of $f(\cdot)$ and the estimator $\hat{\theta}$ of the sequence θ is

$$\hat{f}(\cdot) = \sum_{i=1}^{\infty} \hat{\theta}_i \phi_i(\cdot).$$

Since $W(f) := \int f \mathrm{d}W(t)$ is an isonormal process on L_2, the correspondence between $W(\cdot)$ and W is given by

$$W_i = \int \phi_i(t)\mathrm{d}W(t) \text{ for } i \in \mathbb{N}$$

and

$$W(t) = \sum_{i=1}^{\infty} W_i \int_0^t \phi_i(t')\mathrm{d}t'.$$

2.3 Function space

In this section, we discuss function space. An orthogonal basis in $L_2[0,1]$ provides a clear description of a function space. A typical example is a periodic Sobolev space.

A periodic Sobolev space is defined as follows.

Definition 2.21. A periodic Sobolev space of order $\alpha \in \mathbb{N}$ is defined by

$$\mathcal{F}_{\mathrm{Sobolev}}(B,\alpha) := \left\{ f \in L_2[0,1] : f^{(\alpha)}(0) = f^{(\alpha)}(1) \text{ and } \int \{f^{(\alpha)}(t)\}^2 \mathrm{d}t \le \pi^{2\alpha} B \right\},$$

where $f^{(\alpha)}$ is the weak derivative of order α of f recursively defined by

$$\int_{[0,1]} f(t)\phi^{(1)}(t)\mathrm{d}t = -\int_{[0,1]} f^{(1)}(t)\phi(t)\mathrm{d}t \text{ for all } \phi \in C^{\infty}[0,1] \text{ with } \phi(0) = \phi(1),$$

and

$$\int_{[0,1]} f^{(\alpha-1)}(t)\phi^{(1)}(t)\mathrm{d}t = -\int_{[0,1]} f^{(\alpha)}(t)\phi(t)\mathrm{d}t \text{ for all } \phi \in C^{\infty}[0,1] \text{ with } \phi(0) = \phi(1).$$

The trigonometric series yields a simple description of a periodic Sobolev space. The trigonometric series $\{\psi_{\mathrm{tri},k}\}_{k\in\mathbb{N}}$ is defined by

$$\psi_{\mathrm{tri},k}(t) = \begin{cases} 1, & (k=1), \\ \sqrt{2}\cos\left(2\pi\frac{k}{2}t\right), & (k\text{:even}), \\ \sqrt{2}\sin\left(2\pi\frac{k-1}{2}t\right), & (k\text{:odd}). \end{cases}$$

For $f \in L_2[0,1]$, let $\theta_f = (\theta_{f,1}, \theta_{,f2}, \cdots)$ be the coefficients of f with respect to $\{\psi_{\mathrm{tri},k}\}_{k\in\mathbb{N}}$: $\theta_{f,i} = \int_{[0,1]} f(t)\psi_{\mathrm{tri},i}(t)\mathrm{d}t$.

Theorem 2.22 (e.g.,pp. 196–198 in Tsybakov (2009))**.** *A function $f \in L_2[0,1]$ is included in $\mathcal{F}_{\mathrm{Sobolev}}(B,\alpha)$ if and only if the corresponding coefficients θ_f with respect to $\{\psi_{\mathrm{tri},k}\}_{k\in\mathbb{N}}$ are included in an ellipsoid*

$$\left\{ \theta \in l_2 : \sum_{i:even} i^{2\alpha}\theta_i^2 + \sum_{i:odd} (i-1)^{2\alpha}\theta_i^2 \le B \right\}.$$

2.4 List of notations

In this section, we provide a list of notations and symbols.

- X and Y are current and future observations, respectively.
- n and m are numbers of current and future observations, respectively.
- ε and $\tilde{\varepsilon}$ are deviations of current and future observations, respectively.
- $R(\cdot, \cdot)$ is the Kullback–Leibler risk for predictive density estimation.
- $R_{\mathrm{e}}(\cdot, \cdot)$ is the estimative Kullback–Leibler risk for parameter estimation.
- P and Q are distributions of current and future observations, respectively.
- θ and λ are unknown parameters.
- $\hat{\theta}$ and $\hat{\lambda}$ are estimators.
- $\hat{Q}$ is a predictive distribution.
- δ is a predictive distribution (an estimator) other than $\hat{Q}$ ($\hat{\theta}$, respectively).
- γ and c are constants.
- d is the dimension of a parameter.
- t is a time index.

Chapter 3

Asymptotically minimax predictive density estimation in nonparametric statistical models

This part is scheduled to be published as part of a journal.

Chapter 4

Nonparametric estimation using scale ratio asymptotics

This part is scheduled to be published as part of a journal.

Chapter 5

On minimaxity and weak admissibility in Poisson sequence models

This part is scheduled to be published as part of a journal.

Chapter 6

Prediction when distributions of current and future observations differ

This part is scheduled to be published as part of a journal.

Chapter 7

Conclusion

In this thesis, we provided four results related to prediction in nonparametric models.

In Chapter 3, we provided a construction method for an asymptotically minimax predictive distribution in a nonparametric model. First, we showed the statistical equivalence between predictive density estimation in the function model and that in the infinite sequence model. Second, focusing on the above equivalence, we constructed an asymptotically minimax Bayesian predictive distribution for the case in which the parameter space is a known ellipsoid. Third, using the product of Stein's priors based on the division of the parameter into blocks, we constructed an asymptotically minimax Bayesian predictive distribution for any Sobolev ellipsoid. Finally, we provided efficient computational methods for the construction of the above Bayesian predictive distribution.

In Chapter 4, we investigated refinements of nonparametric estimation focusing on the scale ratio of the parameter and the noise. We presented large scale-ratio asymptotics for nonparametric estimation in a Gaussian sequence model for the case in which the parameter space is a Sobolev ellipsoid. First, we pointed out that typical Bayes estimators that are asymptotically minimax as the variance of the noise diminishes to zero do not work well when the scale of the parameter is large. Second, to resolving this issue, we defined scale-ratio minimaxity and then constructed the prior distribution of which the Bayes estimator is scale-ratio minimax up to a logarithmic factor without using knowledge of the scale of the parameter space. Third, we investigated the mass of the proposed prior on the parameter space and investigated weak admissibility of the Bayes estimator based on the proposed prior. Finally, we considered an extension of the proposed prior to settings in which the smoothness of the true parameter is unknown and provided some numerical experiments related to our proposed prior distribution and its extension.

In Chapter 5, we investigated the estimative Kullback–Leibler risk for estimation in a Poisson sequence model motivated by estimation of an intensity function in a nonhomogeneous Poisson point process. We derived an exact minimax estimator and the corresponding risk and derived several bounds related to weak admissibility.

In Chapter 6, we investigated the performance of predictive densities in parametric models when there is an extrapolation. For extrapolation, we considered prediction when the distributions of current and future observations might differ with an identical unknown finite-dimensional parameter. First, we derived asymptotic forms of Kullback–Leibler risks and showed that Bayesian predictive distributions have smaller risks than plugin predictive distributions when the sample sizes of both current and future observations simultaneously grow to infinity. The asymptotic form of the risk is different from that when the sample size of the current distribution grows to infinity but the sample size of future observations is one. We presented the interpretation of the asymptotic form. Second, we proposed a model selection criterion for predictive settings in which the true distributions of current and future observations might differ. Through numerical experiments we showed that our

proposed model selection criterion works effectively. The numerical experiments indicated that our work on prediction in parametric models is extensible to prediction in high-dimensional models.

Acknowledgements

I am deeply grateful to my supervisor Professor Fumiyasu Komaki for all his support and all his encouragement. He gave me many fruitful comments on my works and spared time to discuss with me during my master and doctor courses. I have been deeply impressed by his thoughtful and powerful work, which is the motivation of my works.

I am thankful to the members of my defence committee for reading the thesis and for giving their valuable comments: the members are Professor Yuzo Maruyama, Professor Kenji Yamanishi, Professor Hiromichi Nagao, and Professor Tomonari Sei. I would like to thank Professor Tomonari Sei and Professor Hiromichi Nagao also for giving their valuable comments in Mathematical Informatics 4th Laboratory's meetings. I thank Professor Hiromichi Nagao for inviting me to his lab seminar of data assimilation.

When I was in the final year, the laboratory that I belonged to changed from Mathematical Informatics 5th Laboratory to Mathematical Informatics 4th Laboratory. I am grateful to Professor Yoshihiro Kanno and Professor Taiji Suzuki for giving their fruitful comments in Mathematical Informatics 5th Laboratory's meetings. When I was a master student, I was stimulated to Professor Taiji Suzuki's work on nonparametric Bayesian statistics.

I am grateful to Professor Kengo Kato for giving his valuable comments on my works and for permitting me to attend to his lab seminar. He taught me way of thinking in nonparametric statistics and probability theory.

I am grateful to Dr. Yoshihiro Hirose for giving his valuable advises.

I thank my fellows Mr. Takeru Matsuda, Mr. Masaaki Imaizumi, and Mr. Shiro Ozaki. Mr. Takeru Matsuda is my fellow student in Mathematical Informatics 4th Laboratory. He has always been there, which has been the encouragement to me during five years. I was stimulated by his interesting way of thinking and his deep insight. Mr. Masaaki Imaizumi is also my fellow in the University of Tokyo. He spared time to discuss about recent studies on nonparametric statistics and machine learning. Mr. Shiro Ozaki was my fellow student when I was in the master course.

I also thank Mr. Kentaro Minami. Mr. Kentaro Minami is a Ph. D. student in Mathematical Informatics 4th Laboratory. He gave me much information about recent studies on machine learning and discussed on many interesting topics with me.

I am grateful to the secretary of Mathematical Informatics 4th Laboratory Ms. Yumiko Fujii and to the secretary of Mathematical Informatics 5th Laboratory Ms. Yuri Inoue.

I give my gratitude to all the laboratory members of Mathematical Informatics 4th Laboratory at Graduate School of Information Science and Technology, the University of Tokyo. I also give my gratitude to all the members of Mathematical Informatics 5th Laboratory at Graduate School of Information Science and Technology, the University of Tokyo.

I have been financially supported by Japan Society for the Promotion of Science from April 2015.

Lastly, I thank my family for all their support to my daily life.

Bibliography

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62**, pp. 547–554.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Caski, eds., *Proc. of the 2nd International Symposium of Information Theory*. Akadimiai Kiado, pp. 267–281.

Aslan, M. (2006). Asymptotically minimax Bayes predictive densities. *Ann. Statist.* **34**, pp. 2921–2938.

Baraud, Y. (2000). Model selection for regression on a fixed design. *Probab. Theory Relat. Fields* **117**, pp. 467–493.

Bernardo, J. (1979). Reference posterior distributions for bayesian inference. *J. Royal. Statist. Soc. B* **41**, pp. 113–147.

Bontemps, D. (2011). Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *Ann. Statist.* **39**, pp. 2557–2584.

Brown, L., George, E., and Xu, X. (2008). Admissible predictive density estimation. *Ann. Statist.* **36**, pp. 1156–1170.

Brown, L. and Low, M. (1996). Asymptotic equaivalence of nonparametric regression and white noise. *Ann. Statist.* **24**, pp. 2384–2398.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press.

Chang, Y.-T. and Strawderman, W. (2014). Stochastic domination in predictive density estimation for ordered normal means under α-divergence loss. *J. Multi. Analysis* **128**, pp. 1–9.

Chapelle, O., Vapnik, V., and Weston, J. (1999). Transfuctive inference for estimating values of functions. In *Advances in Neural Information Processing Systems 12*.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.* **98**, pp. 900–916.

Corcuera, J. and Giummolè, F. (1999). A generalized Bayes rule for prediction. *Scand. J. Statist.* **26**, pp. 265–279.

Cortes, C. and Mohri, M. (2007). On transductive regression. In *Advances in Neural Information Processing Systems 19*.

Dawid, P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *J. Royal Statist. Society A* **147**, pp. 278–292.

Ferguson, T. (1967). *Mathematical statistics: A decision theoretic approach*. Academic Press.

Fushiki, T. (2005). Bootstrap prediction and bayesian prediction under misspecified models. *Bernoulli* **11**, pp. 747–758.

Geisser, S. (1993). *Predictive inference: An introduction*. CRC Press.

George, E., Liang, F., and Xu, X. (2006). Improved minimax predictive densities under kullback–leibler loss. *Ann. Statist.* **34**, pp. 78–91.

Giné, E. and Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press.

Goldenshluger, A. and Tsybakov, A. (2003). Optimal prediction for linear regression with

infinitely many parameters. *J. Multivariate Anal.* **84**, pp. 40–60.
Grenander, U. (1981). *Abstract inference*. John Wiley and Sons Inc.
Grünwald, P. (2007). *The minimum description length principle*. MIT Press.
Hartigan, J. (1998). The maximum likelihood prior. *Ann. Statist.* **26**, pp. 2083–2103.
Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: Data Mining, Inference, and Prediction*. 2nd edn.
Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98**, pp. 879–899.
Ing, C.-K. and Wei, C.-Z. (2005). Order selection for same-realization predictions in autoregressive processes. *Ann. Statist.* **33**, pp. 2423–2474.
Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika* **83**, pp. 299–313.
Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observations. *Biometrika* **88**, pp. 859–864.
Komaki, F. (2011). Bayesian predictive densities based on latent information priors. *J. Statist. Plann. Infer.* **141**, pp. 3705–3715.
Lehmann, E. and Casella, G. (2003). *Theory of Point Estimation*. Springer, 2nd edn.
Liang, F. and Barron, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE tran. on INFORMATION THEORY* **50**, pp. 2708–2726.
Lv, L. and Liu, J. (2014). Model selection principles in misspecified models. *J. R. Statist. Soc. B* **76**, pp. 141–167.
Maruyama, Y. and Strawderman, W. (2012). Bayesian predictive densities for linear regression models under α-divergence loss: Some results and open problems. *IMS Collections* **8**, pp. 42–56.
Massart, P. (2007). Concentration inequalities and model selection .
Mukherjee, G. and Johnstone, I. (2015). Exact minimax estimation of the predictive density in sparse gaussian models. *Ann. Statist.* **43**, pp. 937–961.
Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24**, pp. 2399–2430.
Parzen, E., Tanabe, K., and Kitagawa, G., eds. (1998). *Selected papers of hirotugu akaike*. Springer Science+Business Media.
Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist* **8**, pp. 147–164.
Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, pp. 45–54.
Shimodaira, H. (1997). Assessing the error probability of the model selection test. *Ann. Inst. Statist. Math.* **49**, pp. 395–410.
Sin, C. and White, H. (1996). Information criteria for selecting possibly misspecified parameter models. *J. Econom.* **71**, pp. 207–225.
Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer.
Suzuki, T. and Komaki, F. (2010). On prior selection and covariate shift of beta-Bayesian prediction under alpha-divergence risk. *Comm. in Statist. Theory* **39**, pp. 1655–1673.
Takeuchi, K. (1976). Distribution of information statistics and a criterion of model fitting. *Suri-Kagaku* **153**, pp. 12–18. In Japanese.
Tanaka, F. and Komaki, F. (2011). Asymptotic expansion of the risk difference of the Bayesian spectral density in the autoregressive moving average model. *Snkhyā* **73-A**, pp. 162–184.
Tsybakov, A. (2009). *Introduction to nonparametric estimation*. Springer Science+Business Media.
Wasserman, L. (2007). *All of nonparametric statistics*. Springer, 3rd edn.

Xu, X. and Liang, F. (2010). Asymptotic minimax risk of predictive density estimation for non-parametric regression. *Bernoulli* **16**, pp. 543–560.

Xu, X. and Zhou, D. (2011). Empirical Bayes predictive densities for high-dimensional normal models. *J. Multi. Anal.* **102**, pp. 1417–1428.

Yano, K. and Komaki, F. (2014). Asymptotically constant-risk predictive densities when the distributions of data and target variables are different. *Entropy* **16**, pp. 3026–3048.

Yano, K. and Komaki, F. (2016a). Asymptotically minimax prediction in infinite sequence models. Submitted.

Yano, K. and Komaki, F. (2016b). Information criteria for prediction when the distributions of current and future observations differ. Accepted for publication in *Statist. Sinica*.

Yano, K. and Komaki, F. (2016c). Large scale-ratio asymptotics for nonparametric estimation. Submitted.