博士論文

論文題目

Penalized Least Squares Approximation Methods and Their Applications to Stochastic Processes

（罰則付き最小二乗近似法とその確率過程への応用）

氏　名　　鈴木　拓海

Penalized Least Squares Approximation Methods and Their Applications to Stochastic Processes

SUZUKI Takumi

January 8, 2020

Acknowledgements

I would like to express my sincere thanks to Professor Nakahiro Yoshida for giving me many valuable instructions and advices. It was a great pleasure in my life that I had received his solicitous education and generous support.

Contents

Chapter 1

Introduction

In this thesis, we are interested in two topics: one is the sparse estimation and the other is statistical inference for stochastic processes. As a prologue, we shall begin to introduce the general history of these two topics briefly.

1.1 A brief review of sparse estimation

For several decades, the sparse modeling has received attention from various fields. The most commonly used sparse modeling in statistics is L^1 regularization, and typical one is the LASSO (least absolute shrinkage and selection operator), which is proposed by Tibshirani [35]. LASSO is a useful and widely studied approach to the problem of variable selection. Compared with other estimation methods, LASSO's major advantage is simultaneous execution of both parameter estimation and variable selection ([16], [35]). Originally, LASSO was introduced for linear regression problems. Suppose that $\mathbf{y} = [y_1, \ldots, y_T]'$ is a response vector and $\mathbf{x}_j = [x_{1j}, \ldots, x_{Tj}]', j = 1, \ldots, d$, are the linearly independent predictors.[1] Then the LASSO estimator is defined by

$$\hat{\theta}_{\mathrm{LASSO}} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \left\| \mathbf{y} - \sum_{j=1}^{d} \mathbf{x}_j \theta_j \right\|^2 + \lambda \sum_{j=1}^{d} |\theta_j| \right\}, \tag{1.1}$$

where λ is a nonnegative regularization parameter. The second term on the right-hand side of (1.1) is the so-called L^1 penalty. Thanks to the singularity of the L^1 penalty at the origin, LASSO can perform automatic variable selection.

[1] The prime denotes the matrix transpose.

LASSO has evolved in various directions over the last two decades, and the directions of its development are roughly categorized into the following three.

(1) Extension of models and penalties.

(2) Investigations to the properties of estimators.

(3) Developments of algorithms to calculate estimators.

First, regarding (1), LASSO is applied to not only the linear regression model, but also the generalized linear regression model ([20], [33]), the graphical model ([29], [48]), the multivariate analysis ([44], [53]), the (quasi-) likelihood analysis ([14]), etc. Regarding the penalty terms, there are various extensions according to the analysis. One extension is to replace the L^1-penalty with another penalty like the adaptive LASSO ([52]), Bridge ([18]), SCAD (smoothly clipped absolute deviation [16]), MCP (minimax concave penalty [49]), etc. Another extension is using the relation among parameters like the Group LASSO ([47]), the fused LASSO ([36]), etc.

In relation to (2), the concept of the LASSO is that "*we sacrifice a little bias to reduce the variance of the predicted values and hence may improve the overall prediction accuracy*" ([36]), but it is preferable that the estimator of interest is consistent or asymptotically normal as the sample size is large enough. Thus Knight and Fu ([26]) studied the asymptotic properties of the LASSO estimator and showed that the LASSO type estimator is consistent and asymptotically normal. Recently, the convergence rate and selection consistency of the estimators have been well studied in the case where, not only the sample size, the number of variables is also large ([5], [42], [50]). However, when the number of variables is quite large, strong conditions are required to derive the good properties of the LASSO estimator. Thus, the LASSO cannot be used in the practical situation with high dimensional data.

With regards to (3), the LARS algorithm (Least Angle Regression [15]) was the first noticed algorithm for the LASSO. The estimator obtained by the LARS algorithm is quite similar to the LASSO estimator and, moreover, those two estimators coincide by using the slightly modified version of the LARS algorithm. In recent years, fast and versatile algorithms such as the coordinate decent ([19]) or the ADMM algorithm (Alternating Direction Method and Multipliers [8]) have been proposed.

Based on the above, in this thesis, we will discuss the following.

Regarding (1), by replacing the first term on the right-hand side of (1.1) with a general loss function $\mathcal{L}_T$, we can easily apply the LASSO to various models. However, the asymptotic and numerical theories are established in a

case-by-case manner. One of the solutions to this problem is the least squares approximation (LSA) method proposed by Wang and Leng ([43]). The LSA is defined by a simple approximation to the original loss function:

$$\frac{1}{T}\mathcal{L}_T(\theta) \approx (\theta - \tilde{\theta})'\hat{G}(\theta - \tilde{\theta})$$

where $\hat{G}$ is a non-singular matrix depending on the data and $\tilde{\theta}$ is the minimizer of the loss function $\mathcal{L}_T$. Using the LSA method, we can deal with many different models in a unified frame. Choice of the penalty term is also a crucial issue in regularization techniques. In this thesis, we adopt the Bridge and the adaptive lasso type penalty, i.e., a weighted L^q penalty because the Bridge estimator with $0 < q < 1$ and the adaptive LASSO estimator have the "oracle properties". Oracle properties were recognized by Fan and Li ([16]) and a good estimator with variable selection should have these properties. Let $\theta^* = [\theta_j^*]_j$ is the true value of θ and $\mathcal{A} = \{j; \theta_j^* \neq 0\}$. An estimator $\hat{\theta}$ has oracle properties if $\hat{\theta}$ satisfies

- selection consistency: $P[\hat{\theta}_{\mathcal{A}^c} = 0] \to 1$, and
- asymptotic normality: $\sqrt{T}(\hat{\theta}_{\mathcal{A}} - \theta^*_{\mathcal{A}}) \to^d N(0, G^{-1})$, for some $|\mathcal{A}| \times |\mathcal{A}|$ positive definite symmetric matrix G.

Then our objective function $Q_T^{(q)}(\theta)$ consists of an LSA term and an weighted L^q penalty term:

$$Q_T^{(q)}(\theta) = (\theta - \tilde{\theta})'\hat{G}(\theta - \tilde{\theta}) + \lambda_T \sum_{j=1}^{d} \hat{w}_j |\theta_j|^q. \tag{1.2}$$

Moreover, for $0 < q \leq 1$, we define a penalized least squares approximation (penalized LSA, pLSA) estimator by $\hat{\theta}^{(q)} = \mathrm{argmin}_\theta Q_T^{(q)}(\theta)$.

Next, respecting (2), we first show that the pLSA estimator defined above has the oracle properties. Moreover, we investigate the properties of the pLSA estimator in more detail. Specifically, we consider the L^p-boundedness of the initial estimator. This concept is often used and plays an important role in the statistical inference for stochastic processes. By assuming the L^p-boundedness of the initial estimator, we have the L^p-boundedness of the pLSA estimator and evaluate the convergence rate of selection consistency.

Finally, with respect to (3), since the pLSA objective function is non-convex when $q < 1$, the pLSA methods have a disadvantage in optimization in comparison with the L^1 regularization methods. However, by simplifying the objective function (1.2) by replacing the coefficient matrix $\hat{G}$ with the

identity matrix, optimization of the pLSA objective function comes down to the one-dimensional optimization. Due to that simplification, the pLSA estimator loses the efficiency, but we can obtain the efficient estimator by estimating the parameter again under the model which is selected by the pLSA estimator.

1.2 A brief review of statistical inference for stochastic processes

A stochastic process is the mathematical concept which describes random phenomenon depending on time or space. Studies on the statistics of stochastic processes began in the 1970s, and it grew rapidly, especially with the development of the martingale limit theorem in the 1980s. In recent years, research has bean conducted using not only martingale theory, but also the Malliavin calculus and the theory of empirical processes, and the basic theory has become wider and deeper. On the other hand, its applications to various fields such as actually finance, actuarial science, biostatistics, and survival analysis have been actively studied, and its importance has been widely recognized.

This thesis treats some specific examples of continuous-time stochastic processes, thus, we will mention this a little. The asymptotic inference for continuous-time stochastic processes has been studied based on the likelihood theory by many authors. There are many examples of this: Markov processes with the general state space (Billingsley [6]), Markov branching processes (Athreya and Keiding [2], Feigin [17]), point processes (Brown [11], Brillinger [10]) and general Levy processes (Akritas and Johnson [1]), for instance. In the 1990s, the asymptotic inference for stochastic processes began to be discussed in the general framework of semimartingales that was a wide class of stochastic processes.The LAMN property discussed by Jeganathan [24] and Basawa and Scott [4] in general framework was applied to the class of semimartingales by Luschgy [27], and he introduced the new concept of the local asymptotic quadraticity (LAQ). Yoshida [46] gave a polynomial type large deviation inequality in this LAQ setting to carry out the Ibragimov-Has'minskii-Kutoyants scheme for stochastic processes. The polynomial type large deviation inequality works in various settings, in particular, it is applicable to our examples. As mentioned in [46], L^p-boundedness is derived from the PLDI.

1.3 Organization of this thesis

This thesis is mostly based on [34] and consists of three chapters.

In Chapter 2, we first define the penalized least squares approximation estimator and derive the oracle properties of the pLSA estimator and the convergence rate of selection consistency. This is the main part of this thesis. Next, we define P-O estimator, which does not have the oracle properties but is consist. "P-O" is the abbreviation of "penalized method to ordinary method", i.e., P-O estimator is obtained by the following two steps : (i) obtain an estimator which satisfies the selection consistency (not necessarily having the oracle properties), (ii) obtain an estimator by using the ordinary method under the model which is selected by the estimator in Step (i). Specifically, instead of using the pLSA objective function $Q_T^{(q)}(\theta)$, we use the objective function

$$Q_{T,I}^{(q)}(\theta) = |\theta - \tilde{\theta}|^2 + \lambda_T \sum_{j=1}^{d} \hat{w}_j |\theta_j|^q.$$

Thus, the optimization of the objective function is reduced the one-dimensional one. This gives a computational advantage. We sill prove the theorems at the end of Chapter 2.

In Chapter 3, we apply the pLSA methods to stochastic processes. We focus on two types of processes: point process and diffusion type process. For the point process, we first introduce the general theory of the point process. Next, a Cox model and a Hawkes model will be considered as the specific example. For the Cox model, we consider the intensity with following:

$$\lambda(t, \theta) = \exp\Big(\sum_{j \in \mathbf{J}} \theta_j X_t^j\Big),$$

where $X^j = (X_t^j)_t$ are ergodic covariate processes. For the Hawkes model, in particular, we treat the exponential Hawkes process. In this model, since the parameter space is a subset of $\mathbb{R}_+^{\mathtt{d}} \times \mathbb{R}_+^{\mathtt{d}\times\mathtt{d}} \times \mathbb{R}_+^{\mathtt{d}\times\mathtt{d}}$, the true value of the parameter is on the boundary of the parameter space under the sparse situation. We will derive the selection consistency of the pLSA estimator even in this case. There are applications to various research field on the Hawkes process and recently we often consider the high dimensional situation. Therefore, it is worth considering the variable selection of the Hawkes model. For the diffusion type process, we treat the both ergodic and non-ergodic cases and also use QLA. In the ergodic case, drift parameter and volatility parameter are simultaneously estimated by QLA and we show that the variable selection is also executed simultaneously by using the pLSA methods. In regard

to the non-ergodic case, we consider only the volatility parameter. Since the QMLE (or QBE) has the asymptotic mixed normality, this is an example where the limit G of coefficient matrix $\hat{G}$ is random matrix.

In Chapter 4, we report three simulations: (i) Cox model, (ii) Hawkes model and (iii) non-ergodic diffusion type process. For the Cox model, we take 20 Ornstein-Uhlenbeck processes as the covariates. We use the P-O estimator to avoid optimization involving the 20-dimensional parameter. For the Hawkes process, we first show the results that the initial estimator (i.e. QMLE) performs well even if the case where the true value of parameter is on the boundary of the parameter space. It will be shown that the pLSA estimator performs well. In the cases (i) and (iii), we also use the unified LASSO type estimator and the Bridge type estimator for comparison.

Chapter 2

Penalized LSA estimator

In this chapter, we will discuss the theory of penalized LSA estimation. As mentioned in the previous chapter, we can deal with various kinds of loss functions in a unified way by using the penalized least squares approximation (pLSA) methods defined in this chapter. Penalized methods in the general case has often been discussed. For example, [28] gives some results for penalized methods in M-estimation. Since optimization is usually not easy for a high dimensional parameter, it is worth considering LSA type estimation.

For the pLSA estimator, we will show the oracle property, and derive the L^p-boundedness of pLSA estimator. Moreover we will obtain the convergence rate of variable selection consistency from the L^p-boundedness of the initial estimator. We also construct an objective function in a simpler form than the conventional LSA type estimation by replacing the coefficient matrix by the identity matrix. Optimization for such objective function is easy even if the case $q < 1$.

2.1 Definition of penalized LSA estimator

Suppose that $\theta = [\theta_1, \ldots, \theta_{\mathsf{p}}]' \in \Theta \subset \mathbb{R}^{\mathsf{p}}$ is a parameter of interest and $\tilde{\theta} = [\tilde{\theta}_1, \ldots, \tilde{\theta}_{\mathsf{p}}]' \in \mathbb{R}^{\mathsf{p}}$ is an estimator of θ, where the parameter space Θ is an open bounded subset of $\mathbb{R}^{\mathsf{p}}$. In many cases, $\tilde{\theta}$ minimizes some loss function $\mathcal{L}_T(\theta)$, but we will not assume the existence of the loss function. Here, T is a time index and we often consider the case where $T \in \mathbb{N}$ with discrete time observation or $T \in \mathbb{R}_+$ with continuous time observation. $\tilde{\theta}$ depends on T, however, we omit T for the sake of notational simplicity : $\tilde{\theta} = \tilde{\theta}_T$.

Example. Consider a linear regression model $y_t = \mathbf{x}_t'\theta + \epsilon_t$, $(t = 1, \ldots, T, T \in \mathbb{N})$, where $\{\epsilon_t\}_t$ are independent and identically distributed random variables

with mean 0 and covariance σ^2 and $\{\mathbf{x}_t\}_t$ is independent of $\{\epsilon_t\}_t$. Then we can take $\tilde{\theta}$ as the least square estimator for $\mathcal{L}_T(\theta) = \sum_t |y_t - \mathbf{x}_t'\theta|^2$.

Example. If we consider a negative log-likelihood function as the loss function, then $\tilde{\theta}$ is the maximum likelihood estimator (MLE) of θ.

Hereafter, we assume that there exists a true value $\theta^* = [\theta_1^*, \ldots, \theta_{\mathsf{p}}^*]' \in \mathbb{R}^{\mathsf{p}}$ of θ and that p^0 components of θ^* do not equal to 0, $\mathsf{p}^0 = \#\{j; \theta_j^* \neq 0\}$. Here, for convenience of explanation, we consider a loss function $\mathcal{L}_T(\theta)$. In order to carry out parameter estimation and variable selection simultaneously, we add a penalty term to the loss function $\mathcal{L}_T(\theta)$. For example, we can take a penalized loss function as the adaptive lasso objective function by Zou (2006 [52]):

$$\frac{1}{T}\mathcal{L}_T(\theta) + \sum_{j=1}^{\mathsf{p}} \kappa_T^j |\theta_j|, \tag{2.1}$$

where $\kappa_T^j = \alpha_T |\tilde{\theta}_j|^{-\gamma}$ for a deterministic sequence $(\alpha_T)_T$ and a $\sqrt{T}$-consistent estimator $\tilde{\theta}$.

We consider quadratic approximation of the loss function instead of the first term of (2.1). Thanks to this approximation, we can discuss the various cases into a unified methodology, and because the behavior as $T \to \infty$ is simply described, we can have a more in-depth discussion like large deviation. Moreover, we replace L^1 penalty with L^q penalty ($0 < q \leq 1$) instead. Under this setting, we will show that the parameter estimation and the variable selection can be executed simultaneously in this case. More precisely, for a $\mathsf{p} \times \mathsf{p}$ almost surely positive definite symmetric random matrix $\hat{G}$ depending on T, we use the objective function

$$Q_T^{(q)}(\theta) = \hat{G}[(\theta - \tilde{\theta})^{\otimes 2}] + \sum_{j=1}^{\mathsf{p}} \kappa_T^j |\theta_j|^q,$$

where κ_T^j are nonnegative random variables, $A^{\otimes 2} = AA'$ for a matrix or a vector A, and $A[B] = \mathrm{Tr}(AB')$ for matrices A and B of the same size.

For twice differentiable $\mathcal{L}_T(\theta)$, $T^{-1}\mathcal{L}_T(\theta)$ is approximated as

$$\frac{1}{T}\mathcal{L}_T(\theta) \approx \frac{1}{T}\mathcal{L}_T(\tilde{\theta}) + \frac{1}{T}(\theta - \tilde{\theta})'\partial_\theta \mathcal{L}_T(\tilde{\theta}) + \frac{1}{2}\Big\{\frac{1}{T}\partial_\theta^2 \mathcal{L}_T(\tilde{\theta})\Big\}[(\theta - \tilde{\theta})^{\otimes 2}].$$

Here, the first term on the right hand side is constant with respect to θ and the second term vanishes by the definition of $\tilde{\theta}$. Thus, instead of minimizing $T^{-1}\mathcal{L}_T(\theta)$, we may minimize $\{T^{-1}\partial_\theta^2 \mathcal{L}_T(\tilde{\theta})\}[(\theta - \tilde{\theta})^{\otimes 2}]$ and in this case we can take $\hat{G} = T^{-1}\partial_\theta^2 \mathcal{L}_T(\tilde{\theta})$ for example.

Let $\hat{\theta}^{(q)} = [\hat{\theta}_1^{(q)}, \ldots, \hat{\theta}_{\mathsf{p}}^{(q)}]'$ be a minimizer of this objective function $Q_T^{(q)}(\theta)$:

$$\hat{\theta}^{(q)} \in \operatorname*{argmin}_{\theta \in \bar{\Theta}} Q_T^{(q)}(\theta)$$

We call $\hat{\theta}^{(q)}$ the penalized least squares approximation (penalized LSA) estimator.

2.2 Main theorem

In this section, we will show asymptotic properties of the penalized LSA estimator $\hat{\theta}^{(q)}$ based on $Q_T^{(q)}(\theta)$. Suppose that the statistics are realized on a probability space $(\Omega, \mathcal{F}, P)$. To describe the results, we may suppose that $\theta_1^* \neq 0, \ldots, \theta_{\mathsf{p}^0}^* \neq 0$ and $\theta_{\mathsf{p}^0+1}^* = \cdots = \theta_{\mathsf{p}}^* = 0$ without loss of generality. Let

$$a_T = \max\{\kappa_T^j; j \leq \mathsf{p}^0\} \quad \text{and} \quad b_T = \min\{\kappa_T^j; j > \mathsf{p}^0\}.$$

For a vector $v = [v_1, \ldots, v_{\mathsf{p}}]' \in \mathbb{R}^{\mathsf{p}}$, we denote subvectors $[v_1, \ldots, v_{\mathsf{p}^0}]'$ and $[v_{\mathsf{p}^0+1}, \ldots, v_{\mathsf{p}}]'$ by $v_{\mathcal{J}^1}$ and $v_{\mathcal{J}^0}$ respectively.

We consider the following conditions with respect to $\tilde{\theta}$ and $\hat{G}$. Let r_T be a sequence of positive numbers tending to 0 as $T \to \infty$. We often consider the case that $r_T = T^{-1/2}$.

Assumption 1. There exists a positive definite symmetric random matrix G such that $\hat{G} \to^p G$.

Assumption 2. $\tilde{\theta}$ is r_T^{-1}-consistent, i.e., $r_T^{-1}(\tilde{\theta} - \theta^*) = O_p(1)$.

Assumption 3. $r_T^{-1}(\tilde{\theta} - \theta^*) \to^{d_s} \Gamma^{-\frac{1}{2}}\zeta$ holds, where Γ is a $\mathsf{p} \times \mathsf{p}$ positive definite random symmetric matrix, ζ is a p-dimensional standard Gaussian random vector defined on an extended probability space of $(\Omega, \mathcal{F}, P)$ and independent of $\mathcal{G}$, and d_s denotes the $\mathcal{G}$-stable convergence for some σ-field $\mathcal{G}$ such that $\sigma(\Gamma) \subset \mathcal{G} \subset \mathcal{F}$.

Of course, Assumption 3 is stronger than Assumption 2, but r_T^{-1}-consistency and selection consistency of the penalized LSA estimator $\hat{\theta}^{(q)}$ are derived from Assumptions 1 and 2. We need Assumption 3 to show asymptotic normality of penalized LSA estimator $\hat{\theta}^{(q)}$.

For a $\mathsf{p} \times \mathsf{p}$ matrix $M = [m_{ij}]_{1 \leq i \leq \mathsf{p}, 1 \leq j \leq \mathsf{p}}$, we denote the $\mathsf{p}^0 \times \mathsf{p}^0$ matrix $[m_{ij}]_{1 \leq i \leq \mathsf{p}^0, 1 \leq j \leq \mathsf{p}^0}$, $\mathsf{p}^0 \times (\mathsf{p} - \mathsf{p}^0)$ matrix $[m_{ij}]_{1 \leq i \leq \mathsf{p}^0, \mathsf{p}^0 < j \leq \mathsf{p}}$, $(\mathsf{p} - \mathsf{p}^0) \times \mathsf{p}^0$ matrix $[m_{ij}]_{\mathsf{p}^0 < i \leq \mathsf{p}, 1 \leq j \leq \mathsf{p}^0}$ and $(\mathsf{p} - \mathsf{p}^0) \times (\mathsf{p} - \mathsf{p}^0)$ matrix $[m_{ij}]_{\mathsf{p}^0 < i \leq \mathsf{p}, \mathsf{p}^0 < j \leq \mathsf{p}}$ by $M_{\mathcal{J}^{11}}, M_{\mathcal{J}^{10}}, M_{\mathcal{J}^{01}}$ and $M_{\mathcal{J}^{00}}$ respectively:

$$M = \begin{bmatrix} M_{\mathcal{J}^{11}} & M_{\mathcal{J}^{10}} \\ M_{\mathcal{J}^{01}} & M_{\mathcal{J}^{00}} \end{bmatrix}.$$

Theorem 2.1 (r_T^{-1}-consistency)**.** Under Assumptions 1 and 2, if $r_T^{-1}a_T = O_p(1)$, then

$$r_T^{-1}(\hat{\theta}^{(q)} - \theta^*) = O_p(1).$$

Let $\hat{\mathcal{J}}^1 = \{j = 1, \ldots, \mathsf{p}; \hat{\theta}_j^{(q)} \neq 0\}$.

Theorem 2.2 (Selection consistency)**.** Under Assumptions 1 and 2, if $r_T^{-1}a_T = O_p(1)$ and $r_T^{-(2-q)}b_T \to^p \infty$, then

$$P[\hat{\mathcal{J}}^1 = \{1, \ldots, \mathsf{p}^0\}] \to 1. \tag{2.2}$$

Theorem 2.3 (Asymptotic normality)**.** Let $\mathfrak{G} = \begin{bmatrix} I_{\mathsf{p}^0} & (G_{\mathcal{J}^{11}})^{-1}G_{\mathcal{J}^{10}} \end{bmatrix}$ for $\mathsf{p}^0 \times \mathsf{p}^0$ identity matrix I_{p^0}. Under Assumptions 1 and 2, if $r_T^{-1}a_T = o_p(1)$ and $r_T^{-(2-q)}b_T \to^p \infty$, then

$$r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)_{\mathcal{J}^1} - \mathfrak{G}\{r_T^{-1}(\tilde{\theta} - \theta^*)\} \to^p 0.$$

In particular, under Assumption 3 and $G = \Gamma$, we have

$$r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)_{\mathcal{J}^1} \to^{d_s} \mathfrak{G}\Gamma^{-\frac{1}{2}}\zeta \sim \mathrm{MN}_{\mathsf{p}^0}(0, (\Gamma_{\mathcal{J}^{11}})^{-1}).$$

Hereafter, we consider $\kappa_T^j = \alpha_T|\tilde{\theta}_j|^{-\gamma}$, where γ is a constant satisfying $\gamma > -(1-q)$ and $(\alpha_T)_T$ is a deterministic sequence. If $(\alpha_T)_T$ satisfies the conditions

$$r_T^{-(2-q+\gamma)}\alpha_T \to \infty \qquad \text{and} \qquad r_T^{-1}\alpha_T = o(1), \tag{2.3}$$

then the conditions in Theorems 2.1-2.3 are fulfilled. Moreover we will show that the probability $P[\hat{\theta}_{\mathcal{J}^0}^{(q)} = 0]$ can be evaluated by any power of r_T.

Let $\tilde{u} = r_T^{-1}(\tilde{\theta} - \theta^*)$ and $\hat{u} = r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)$.

Definition 2.4. For a stochastic process $X = \{X_T\}_T$ is $L^{\infty-}$-bounded if and only if $\sup_T E[|X_T|^p] < \infty$ holds for all $p \geq 1$.

Additionally, we consider the following conditions:

Assumption 4. $\{\hat{G}\}_T$, $\{\hat{G}^{-1}\}_T$ and $\{\tilde{u}\}_T$ are $L^{\infty-}$-bounded.

Remark. The L^p-boundedness of a sequence of estimators can be obtained by the quasi-likelihood analysis with a polynomial type large deviation inequality for an associated statistical random field. See [46] for details.

Theorem 2.5. Let $\epsilon \in (-1+q, \gamma)$. Assume that $r_T^{1+\gamma-\epsilon}\alpha_T^{-1} = O(1)$ and $r_T^{-1}\alpha_T = O(1)$. Then under Assumptions 1 and 4, $\{\hat{u}\}_T$ is $L^{\infty-}$-bounded. Moreover, for all $L > 0$, there exists a constant C_L such that

$$P[\hat{\mathcal{J}}^1 = \{1, \ldots, \mathsf{p}^0\}] \geq 1 - C_L r_T^{2L}. \tag{2.4}$$

for all $T > 0$.

2.3 P-O estimator

We now discuss the coefficient matrix $\hat{G}$. In the above theorems, we assume convergence of $\hat{G}$ to G or $L^{\infty-}$-boundedness of $\{\hat{G}\}$ and $\{\hat{G}^{-1}\}$ but we should not necessarily find such coefficient matrix $\hat{G}$. In fact, if we take $\hat{G} = I_{\mathsf{p}}$, then we can apply Theorems 2.1-2.5 except that the conditional asymptotic variance in Theorem 3 becomes $(\Gamma^{-1})_{\mathcal{J}^{11}}$. Since $(\Gamma_{\mathcal{J}^{11}})^{-1} = (\Gamma^{-1})_{\mathcal{J}^{11}} - (\Gamma^{-1})_{\mathcal{J}^{10}}((\Gamma^{-1})_{\mathcal{J}^{00}})^{-1}(\Gamma^{-1})_{\mathcal{J}^{01}}$, this estimator is not efficient. However, the objective function has the following simple form

$$Q_T^{(q)}(\theta) = \sum_{j=1}^{\mathsf{p}} \Big((\theta_j - \tilde{\theta}_j)^2 + \kappa_T^j |\theta_j|^q \Big).$$

From a computational point of view, this fact is useful because it is difficult to optimize the non-convex function in the high-dimensional case. Then we obtain a new estimator under the model selected by the penalized LSA estimator with coefficient matrix I_{p}. We call this estimator the P-O (penalized method to ordinary method) estimator and denote it by $\check{\theta}$. More precisely, we define the P-O estimator as follows.

Let Θ be a bounded open subset of $\mathbb{R}^{\mathsf{p}}$. First, we assume the r_T^{-1}-consistency of the initial estimator $\tilde{\theta}$. Second, we get the penalized LSA estimator $\hat{\theta}_{I_{\mathsf{p}}}^{(q)}$ with coefficient matrix I_{p} defined by

$$\hat{\theta}_{I_{\mathsf{p}}}^{(q)} \in \operatorname*{argmin}_{\theta \in \bar{\Theta}} Q_T^{(q)}(\theta),$$

where $\kappa_T^j = \alpha_T |\tilde{\theta}_j|^{-\gamma}$, $\gamma > -(1-q)$ and α_T satisfies (2.3). Let $\hat{\mathcal{J}}_{I_{\mathsf{p}}}^0 = \{j = 1, \ldots, \mathsf{p}; \hat{\theta}_{I_{\mathsf{p}},j}^{(q)} = 0\}$ and $\hat{\Theta} = \{\theta \in \Theta; \theta_j = 0, j \in \hat{\mathcal{J}}_{I_{\mathsf{p}}}^0\}$. Here, we consider another loss function $\mathbb{L}_T(\theta)$. Then, we define the P-O estimator $\check{\theta}$ by

$$\check{\theta} \in \operatorname*{argmin}_{\theta \in \bar{\hat{\Theta}}} \mathbb{L}_T(\theta).$$

Before we turn to the statement of the results for the P-O estimator $\check{\theta}$, we consider some conditions. We denote a parameter $\theta = \begin{bmatrix} \phi \\ \psi \end{bmatrix} \in \mathbb{R}^{\mathsf{p}^0 + (\mathsf{p} - \mathsf{p}^0)}$ and its true value $\theta^* = \begin{bmatrix} \phi^* \\ \psi^* \end{bmatrix} = \begin{bmatrix} \phi^* \\ 0 \end{bmatrix}$. Let $\bar{\mathbb{L}}_T(\phi) = \mathbb{L}_T\Big(\begin{bmatrix} \phi \\ 0 \end{bmatrix} \Big)$ and

$$\bar{\phi} \in \operatorname{argmin}_{\phi} \bar{\mathbb{L}}_T(\phi). \tag{2.5}$$

Assumption 5. For any sequence of estimations $\{\bar{\phi}\}$ satisfying (2.5),

(i) $\{\tilde{u}\}_T = \{r_T^{-1}(\tilde{\theta} - \theta^*)\}_T$ is $L^{\infty-}$-bounded.

(ii) $r_T^{-1}(\bar{\phi} - \phi^*) \to^{d_s} \Lambda^{-\frac{1}{2}}\eta$, where Λ is a $\mathsf{p}^0 \times \mathsf{p}^0$ positive definite symmetric random matrix, η is a p^0-dimensional standard Gaussian random vector independent of Λ.

(iii) $\{r_T^{-1}(\bar{\phi} - \phi^*)\}_T$ is $L^{\infty-}$-bounded.

Remark. In many cases, we take $\mathbb{L}_T(\theta) = \mathcal{L}_T(\theta)$ and $\Lambda = \Gamma_{\mathcal{J}^{11}}$. Then, the sufficient condition for Assumption 5 is as follows. We define the random field $\mathbb{Z}_T : \mathbb{U}_T \to \mathbb{R}_+$ by $\mathbb{Z}_T(u) = \exp\{-\mathcal{L}_T(\theta^* + r_T u) + \mathcal{L}_T(\theta^*)\}$, where $\mathbb{U}_T = \{u \in \mathbb{R}^{\mathsf{p}}; \theta^* + r_T u \in \Theta\}$. We denote $B(R) = \{u \in \mathbb{R}^{\mathsf{p}}; |u| \leq R\}$. If $\mathbb{Z}_T(u) \to^{d_s} \mathbb{Z}(u)$ in $C(B(R))$ for every $R > 0$ as $T \to \infty$ and $\{\bar{\phi}\}$ is tight, then Assumption 5 (ii) holds. Here, $\mathbb{Z}$ is a random field defined by $\mathbb{Z}(u) = \exp\Big(u'\Gamma^{\frac{1}{2}}\zeta - \frac{1}{2}u'\Gamma u\Big)$. Moreover, if the random field $\mathbb{Z}_T$ satisfies the polynomial type large deviation inequality (Theorem 1 in [46]), then by using Proposition 1 in [46], Assumptions 5(i) and (iii) hold.

Theorem 2.6. (a) Under Assumption 2,

$$P\Big[\check{\theta}_{\mathcal{J}^1} \in \mathrm{argmin}_\phi \bar{\mathbb{L}}_T(\phi)\Big] \to 1. \tag{2.6}$$

Additionally, under Assumption 5(ii),

$$r_T^{-1}(\check{\theta} - \theta^*)_{\mathcal{J}^1} \to^{d_s} \Lambda^{-\frac{1}{2}}\eta \sim \mathrm{MN}_{\mathsf{p}^0}(0, \Lambda^{-1}).$$

(b) Let $\epsilon \in \big(-1 + q, \gamma\big)$. Assume Assumptions 5(i) and (iii), $r_T^{1+\gamma-\epsilon}\alpha_T^{-1} = O(1)$ and $r_T^{-1}\alpha_T = O(1)$. Then $\{r_T^{-1}(\check{\theta} - \theta^*)\}_T$ is $L^{\infty-}$-bounded. Moreover, for all $L > 0$ there exists a constant C_L such that

$$P[\check{\mathcal{J}}^1 = \{1, \ldots, \mathsf{p}^0\}] \geq 1 - C_L r_T^{2L}. \tag{2.7}$$

for all $T > 0$ where $\check{\mathcal{J}}^1 = \{j = 1, \ldots, \mathsf{p}; \check{\theta}_j^{(q)} \neq 0\}$.

2.4 Proofs of main theorems

2.4.1 Proof of Theorem 2.1

Since $\hat{\theta}^{(q)}$ minimizes $Q_T^{(q)}(\theta)$, we obtain

$$\begin{aligned}
0 &\geq Q_T^{(q)}(\hat{\theta}^{(q)}) - Q_T^{(q)}(\theta^*) \\
&= \hat{G}[(\hat{\theta}^{(q)} - \tilde{\theta})^{\otimes 2}] + \sum_{j=1}^{\mathsf{p}} \kappa_T^j |\hat{\theta}_j^{(q)}|^q - \hat{G}[(\theta^* - \tilde{\theta})^{\otimes 2}] - \sum_{j=1}^{\mathsf{p}} \kappa_T^j |\theta_j^*|^q \\
&= \hat{G}[(\hat{\theta}^{(q)} - \theta^*)^{\otimes 2}] + 2(\hat{\theta}^{(q)} - \theta^*)'\hat{G}(\theta^* - \tilde{\theta}) + \sum_{j=1}^{\mathsf{p}} \kappa_T^j |\hat{\theta}_j^{(q)}|^q - \sum_{j=1}^{\mathsf{p}} \kappa_T^j |\theta_j^*|^q.
\end{aligned} \tag{2.8}$$

Since $0 \leq |\hat{\theta}_j^{(q)}| < |\theta_j^*|$ implies $(|\theta_j^*|^q - |\hat{\theta}_j^{(q)}|^q)/(|\theta_j^*| - |\hat{\theta}_j^{(q)}|) \leq |\theta_j^*|^q/|\theta_j^*| = |\theta_j^*|^{q-1}$, we obtain $|\hat{\theta}_j^{(q)}|^q - |\theta_j^*|^q \geq -K^*|\hat{\theta}_j^{(q)} - \theta_j^*|$ where $K^* = \max_{1 \leq j \leq \mathsf{p}^0} |\theta_j^*|^{q-1}$. Thus

$$\begin{aligned}
\sum_{j=1}^{\mathsf{p}} \kappa_T^j |\hat{\theta}_j^{(q)}|^q - \sum_{j=1}^{\mathsf{p}} \kappa_T^j |\theta_j^*|^q &\geq \sum_{j=1}^{\mathsf{p}^0} \kappa_T^j (|\hat{\theta}_j^{(q)}|^q - |\theta_j^*|^q) \\
&\geq -\sum_{j=1}^{\mathsf{p}^0} K^* \kappa_T^j |\hat{\theta}_j^{(q)} - \theta_j^*| \\
&\geq -\mathsf{p}^0 K^* a_T |\hat{\theta}^{(q)} - \theta^*|.
\end{aligned}$$

Therefore, by multiplying both sides of (2.8) by r_T^{-2}, we obtain

$$\begin{aligned}
0 &\geq \hat{G}\big[\{r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)\}^{\otimes 2}\big] + 2\{r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)\}'\hat{G}\{r_T^{-1}(\theta^* - \tilde{\theta})\} - \\
&\qquad \mathsf{p}^0 K^* r_T^{-1} a_T |r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)| \\
&\geq \tau_{\min}(\hat{G}) |r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)|^2 - 2\tau_{\max}(\hat{G}) |r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)| |r_T^{-1}(\tilde{\theta} - \theta^*)| - \\
&\qquad \mathsf{p}^0 K^* r_T^{-1} a_T |r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)|,
\end{aligned}$$

where $\tau_{\min}(\hat{G})$ and $\tau_{\max}(\hat{G})$ are the minimum and maximum eigenvalues of matrix $\hat{G}$, respectively. After all,

$$|r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)| \leq \left\{ \frac{1}{\tau_{\min}(\hat{G})} \Big(2\tau_{\max}(\hat{G}) |r_T^{-1}(\tilde{\theta} - \theta^*)| + \mathsf{p}^0 K^* r_T^{-1} a_T \Big) \right\}. \tag{2.9}$$

Since the right hand side is $O_p(1)$ by the assumption, we obtain $r_T^{-1}(\hat{\theta}^{(q)} - \theta^*) = O_p(1)$. □

2.4.2 Proof of Theorem 2.2

First, we assume that $\hat{\theta}^{(q)} \in \partial\Theta$. Then, since $|\hat{\theta}^{(q)} - \theta^*| \geq \epsilon_0$ where $\epsilon_0 = \inf\{|\theta - \theta^*|; \theta \in \partial\Theta\} > 0$, we obtain

$$P[\hat{\theta}^{(q)} \in \partial\Theta] \leq P[r_T^{-1}|\hat{\theta}^{(q)} - \theta^*| \geq r_T^{-1}\epsilon_0] \to 0. \tag{2.10}$$

Next, we assume $\hat{\theta}^{(q)} \notin \partial\Theta$ and $\hat{\theta}_j^{(q)} \neq 0$ for some $j(\mathsf{p}^0 < j \leq \mathsf{p})$. Since $Q_T^{(q)}(\theta)$ is differentiable at $\theta = \hat{\theta}^{(q)}$ with respect to the j-th component and $\hat{\theta}^{(q)}$ minimizes $Q_T^{(q)}(\theta)$,

$$\begin{aligned} 0 &= r_T^{-1}\frac{\partial Q_T^{(q)}(\theta)}{\partial\theta_j}|_{\theta=\hat{\theta}^{(q)}} \\ &= 2\hat{G}^{(j)}\{r_T^{-1}(\hat{\theta}^{(q)} - \tilde{\theta})\} + r_T^{-1}\kappa_T^j q|\hat{\theta}_j^{(q)}|^{q-1}\mathrm{sgn}(\hat{\theta}_j^{(q)}), \end{aligned}$$

where $\hat{G}^{(j)}$ means the j-th row vector of $\hat{G}$. Therefore, we have

$$\begin{aligned} 2|\hat{G}^{(j)}\{r_T^{-1}(\hat{\theta}^{(q)} - \tilde{\theta})\}||r_T^{-1}\hat{\theta}_j^{(q)}|^{1-q} &= qr_T^{-(2-q)}\kappa_T^j \\ &\geq qr_T^{-(2-q)}b_T. \end{aligned} \tag{2.11}$$

Since, by Theorem 2.1 and the assumption, the left hand side of above equation is $O_p(1)$ and $r_T^{-(2-q)}b_T \to^p \infty$, we obtain

$$\begin{aligned} &P\Big[\hat{\theta}_j^{(q)} \neq 0, \hat{\theta}^{(q)} \notin \partial\Theta\Big] \\ &\leq P\Big[|2\hat{G}^{(j)}\{r_T^{-1}(\hat{\theta}^{(q)} - \tilde{\theta})\}||r_T^{-1}\hat{\theta}_j^{(q)}|^{1-q} \geq qr_T^{-(2-q)}b_T\Big] \to 0 \end{aligned} \tag{2.12}$$

for $j = \mathsf{p}^0 + 1, \ldots, \mathsf{p}$.
Thus we have

$$P[\hat{\theta}_{\mathcal{J}^0}^{(q)} \neq 0] \leq P[\hat{\theta}^{(q)} \in \partial\Theta] + \sum_{j=\mathsf{p}^0+1,\ldots,\mathsf{p}} P\Big[\hat{\theta}_j^{(q)} \neq 0, \hat{\theta}^{(q)} \notin \partial\Theta\Big] \to 0.$$

In particular, since $\hat{\mathcal{J}}^1 \neq \{1, \ldots, \mathsf{p}^0\}$ implies

$$|(\hat{\theta}^{(q)} - \theta^*)_{\mathcal{J}^1}| \geq \min_{j=1,\ldots,\mathsf{p}^0}|\theta_j^*| > 0$$

or $\hat{\theta}_{\mathcal{J}^0}^{(q)} \neq 0$, we have (2.2) by Theorem 2.1. □

2.4.3 Proof of Theorem 2.3

For $\theta = \begin{bmatrix} \theta_{\mathcal{J}^1} \\ \theta_{\mathcal{J}^0} \end{bmatrix} \in \mathbb{R}^{\mathsf{p}}$,

$$\begin{aligned} Q_T^{(q)}(\theta) &= \hat{G}[(\theta - \tilde{\theta})^{\otimes 2}] + \sum_{j=1}^{\mathsf{p}} \kappa_T^j |\theta_j|^q \\ &= \hat{G}_{\mathcal{J}^{11}}[(\theta - \tilde{\theta})_{\mathcal{J}^1}^{\otimes 2}] + 2(\theta - \tilde{\theta})'_{\mathcal{J}^1} \hat{G}_{\mathcal{J}^{10}} (\theta - \tilde{\theta})_{\mathcal{J}^0} + \hat{G}_{\mathcal{J}^{00}}[(\theta - \tilde{\theta})_{\mathcal{J}^0}^{\otimes 2}] \\ &\qquad + \sum_{j=1}^{\mathsf{p}^0} \kappa_T^j |\theta_j|^q + \sum_{j=\mathsf{p}^0+1}^{\mathsf{p}} \kappa_T^j |\theta_j|^q. \end{aligned}$$

In particular, for $\theta^\ddagger = \begin{bmatrix} \theta_{\mathcal{J}^1} \\ 0 \end{bmatrix} \in \mathbb{R}^{\mathsf{p}}$,

$$Q_T^{(q)}(\theta^\ddagger) = \hat{G}_{\mathcal{J}^{11}}[(\theta - \tilde{\theta})_{\mathcal{J}^1}^{\otimes 2}] - 2(\theta - \tilde{\theta})'_{\mathcal{J}^1} \hat{G}_{\mathcal{J}^{10}} \tilde{\theta}_{\mathcal{J}^0} + \hat{G}_{\mathcal{J}^{00}}[\tilde{\theta}_{\mathcal{J}^0}^{\otimes 2}] + \sum_{j=1}^{\mathsf{p}^0} \kappa_T^j |\theta_j|^q.$$

Let

$$A_T = \Big\{ \min_{1 \le j \le \mathsf{p}^0} |\hat{\theta}_j^{(q)}| > 0, \hat{\theta}_{\mathcal{J}^0}^{(q)} = 0, \det(\hat{G}_{\mathcal{J}^{11}}) \neq 0 \Big\}.$$

Then Theorems 2.1 and 2.2 imply $P[A_T] \to 1$. Let $\mathbb{R}_0^{\mathsf{p}} = \{\theta \in \mathbb{R}^{\mathsf{p}}; \theta_{\mathcal{J}^0} = 0\}$. Since $Q_T^{(q)}(\hat{\theta}^{(q)}) = \min_{\theta^\ddagger \in \mathbb{R}_0^{\mathsf{p}}} Q_T^{(q)}(\theta^\ddagger)$ on A_T,

$$\begin{aligned} 0 &= \frac{1}{2} \frac{\partial Q_T^{(q)}(\theta)}{\partial \theta_{\mathcal{J}^1}} \Big|_{\theta = \hat{\theta}^{(q)}} \\ &= \hat{G}_{\mathcal{J}^{11}}(\hat{\theta}^{(q)} - \tilde{\theta})_{\mathcal{J}^1} - \hat{G}_{\mathcal{J}^{10}} \tilde{\theta}_{\mathcal{J}^0} + V(\hat{\theta}_{\mathcal{J}^1}^{(q)}) \end{aligned}$$

holds on A_T, where $V(\hat{\theta}_{\mathcal{J}^1}^{(q)}) = \big[2^{-1} q \kappa_T^j |\hat{\theta}_j^{(q)}|^{q-1} \mathrm{sgn}(\hat{\theta}_j^{(q)})\big]_{j=1,\ldots,\mathsf{p}^0} \in \mathbb{R}^{\mathsf{p}^0}$. Let $\hat{\mathfrak{G}} = \big[I_{\mathsf{p}^0} \quad (\hat{G}_{\mathcal{J}^{11}})^{-1} \hat{G}_{\mathcal{J}^{10}}\big]$. Since $\hat{\mathfrak{G}} \to^p \mathfrak{G}$ and $1_{A_T}\{r_T^{-1}(\hat{G}_{\mathcal{J}^{11}})^{-1} V(\hat{\theta}_{\mathcal{J}^1}^{(q)})\} \to^p$

0, we have

$$
\begin{aligned}
& r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)_{\mathcal{J}^1} - \mathfrak{G}\{r_T^{-1}(\tilde{\theta} - \theta^*)\} \\
&= 1_{A_T}\Big\{ r_T^{-1}(\tilde{\theta} - \theta^*)_{\mathcal{J}^1} + r_T^{-1}(\hat{G}_{\mathcal{J}^{11}})^{-1}\hat{G}_{\mathcal{J}^{10}}\tilde{\theta}_{\mathcal{J}^0} \\
&\qquad - r_T^{-1}(\hat{G}_{\mathcal{J}^{11}})^{-1}V(\hat{\theta}^{(q)}_{\mathcal{J}^1}) - \mathfrak{G}\{r_T^{-1}(\tilde{\theta} - \theta^*)\}\Big\} \\
&\qquad\qquad + 1_{A_T^c}\Big\{ r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)_{\mathcal{J}^1} - \mathfrak{G}\{r_T^{-1}(\tilde{\theta} - \theta^*)\}\Big\} \\
&= 1_{A_T}\Big\{ (\hat{\mathfrak{G}} - \mathfrak{G})\{r_T^{-1}(\tilde{\theta} - \theta^*)\} - r_T^{-1}(\hat{G}_{\mathcal{J}^{11}})^{-1}V(\hat{\theta}^{(q)}_{\mathcal{J}^1})\Big\} \\
&\qquad\qquad + 1_{A_T^c}\Big\{ r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)_{\mathcal{J}^1} - \mathfrak{G}\{r_T^{-1}(\tilde{\theta} - \theta^*)\}\Big\} \\
&\to^p 0.
\end{aligned}
$$

□

2.4.4 Proof of Theorem 2.5

By (2.9) and Assumption 4, $\{\hat{u}\}$ is $L^{\infty-}$-bounded. By (2.10) in the Proof of theorem 2.2, we obtain

$$
\begin{aligned}
P[\hat{\theta}^{(q)} \in \partial\Theta] &\le P[r_T^{-1}|\hat{\theta}^{(q)} - \theta^*| \ge r_T^{-1}\epsilon_0] \\
&\le \frac{1}{(r_T^{-1}\epsilon_0)^L} E[|\hat{u}|^L],
\end{aligned}
$$

for all $L \ge 1$. For $j > \mathsf{p}^0$, by the equation (2.11) and the Markov's inequality, we have

$$
\begin{aligned}
& P\Big[\hat{\theta}_j^{(q)} \ne 0, \hat{\theta}^{(q)} \notin \partial\Theta\Big] \\
& \le P\Big[2|\hat{G}^{(j)}| \cdot |\hat{u} - \tilde{u}| \ge r_T^{-1}\kappa_T^j q \big| r_T\hat{u}\big|^{-(1-q)}\Big] \\
& \le \frac{1}{r_T^{-(1-q+\epsilon)M}} 2^M q^{-M} E\left[|\hat{G}^{(j)}|^M |\hat{u} - \tilde{u}|^M |\hat{u}|^{M(1-q)} \Big(\frac{1}{r_T^{\epsilon-1}\kappa_T^j} \Big)^M \right],
\end{aligned}
$$

where $M = M(L) = 2L(1 - q + \epsilon)^{-1}$ and $L > 0$ is an arbitrary constant. Here, by Hölder inequality, we have

$$
\begin{aligned}
& E\left[|\hat{G}^{(j)}|^M |\hat{u} - \tilde{u}|^M |\hat{u}|^{M(1-q)} \Big(\frac{1}{r_T^{\epsilon-1}\kappa_T^j} \Big)^M \right] \\
& \le E\Big[|\hat{G}^{(j)}|^{4M} \Big]^{\frac{1}{4}} E\Big[|\hat{u} - \tilde{u}|^{4M} \Big]^{\frac{1}{4}} E\Big[|\hat{u}|^{4M(1-q)} \Big]^{\frac{1}{4}} E\left[\Big(\frac{1}{r_T^{\epsilon-1}\kappa_T^j} \Big)^{4M} \right]^{\frac{1}{4}}. \quad (2.13)
\end{aligned}
$$

Since

$$E\left[\Big(\frac{1}{r_T^{\epsilon-1}\kappa_T^j}\Big)^{4M}\right] = E\left[\Big(\frac{|\tilde{\theta}_j|^\gamma}{r_T^{\epsilon-1}\alpha_T}\Big)^{4M}\right] \leq \left(\frac{1}{r_T^{-(1+\gamma-\epsilon)}\alpha_T}\right)^{4M} E\Big[|\tilde{u}|^{4\gamma M}\Big]$$

and $\{\hat{u}\}_T$,$\{\tilde{u}\}_T$ and $\{\hat{G}\}$ are $L^{\infty-}$-bounded, the right-hand side of (2.13) is bounded uniformly in T. Finally, the inequality (2.4) is obtained in a similar way as the Proof of Theorem 2.2. □

2.4.5 Proof of Theorem 2.6

(a) Since $\hat{\mathcal{J}}^0_{I_\mathsf{p}} = \{\mathsf{p}^0+1,\ldots,\mathsf{p}\}$ implies $\check{\theta}_{\mathcal{J}^1} \in \mathrm{argmin}_\phi \bar{\mathbb{L}}_T(\phi)$, we obtain (2.6) by Theorem 2.2.
(b) Next, let $B_T = \{\hat{\mathcal{J}}^0_{I_\mathsf{p}} = \{\mathsf{p}^0+1,\ldots,\mathsf{p}\}\}$ and $\mathrm{diam}(\Theta) = \sup\{|\theta_1-\theta_2|;\theta_1,\theta_2\in\Theta\}$. By Theorem 2.5, $P[B_T^c]$ is evaluated by any power of r_T. Since

$$\begin{aligned}&\sup_T E[|r_T^{-1}(\check{\theta}-\theta^*)|^p] \\ &\leq \sup_T E[|r_T^{-1}(\bar{\phi}-\phi^*)|^p] + \sup_T\Big\{P[B_T^c]\cdot(r_T^{-1}\mathrm{diam}(\Theta))^p\Big\} < \infty\end{aligned}$$

for $\bar{\phi} = \check{\theta}_{\mathcal{J}^1}$ and all $p>0$, we have $L^{\infty-}$-boundedness of $\{r_T^{-1}(\check{\theta}-\theta^*)\}_T$. By the definition of $\check{\theta}$, $\hat{\mathcal{J}}^0_{I_\mathsf{p}} = \{\mathsf{p}^0+1,\ldots,\mathsf{p}\}$ and $|\check{\theta}-\theta^*| < \min_{1\leq j\leq \mathsf{p}^0}|\theta_j^*|$ imply $\check{\mathcal{J}}^1 = \{1,\ldots,\mathsf{p}^0\}$. Thus

$$P[\check{\mathcal{J}}^1 \neq \{1,\ldots,\mathsf{p}^0\}] \leq P[B_T^c] + P\Big[|r_T^{-1}(\check{\theta}-\theta^*)| \geq r_T^{-1}\min_{1\leq j\leq\mathsf{p}^0}|\theta_j^*|\Big]$$

Therefore, by $L^{\infty-}$-boundedness of $\{r_T^{-1}(\check{\theta}-\theta^*)\}_T$, we obtain the inequality (2.7). □

Chapter 3

Applications

In the previous chapter, we discussed the theory of pLSA estimator. In this chapter, we consider the its application to analysis of stochastic processes. In particular, we are interested in the point processes and the diffusion processes. For the point process, first, we consider the general theory of ergodic intensity model. Then, as an example, we treat the Cox process and Hawkes process using the quasi-likelihood analysis (QLA) method. For the diffusion process, we consider the ergodic and non-ergodic diffusion processes. We also use QLA method in this case.

3.1 Point process

In this section, we will apply the results in Section 3 to a point process with parameters containing zero components. We consider a multivariate point process $N = (N_t^\alpha)_{\alpha\in\mathbf{I},t\in\mathbb{R}_+}$ with intensity process $\lambda(t,\theta) = (\lambda^\alpha(t,\theta))_{\alpha\in\mathbf{I}}$, $t \in \mathbb{R}_+$, where $\mathbf{I} = \{1, 2, \ldots, \mathsf{d}\}$ is an index set. More precisely, given a stochastic basis $\mathcal{B} = (\Omega, \mathcal{F}, \mathbf{F}, P)$ with a filtration $\mathbf{F} = (\mathcal{F}_t)_{t\in\mathbb{R}_+}$, we suppose that N and $\lambda(\cdot,\theta)$ are defined on $\mathcal{B}$, the simple counting process N is $\mathbf{F}$-adapted right-continuous, $\lambda(\cdot,\theta)$ is predictable locally integrable for every $\theta \in \Theta$, and that $N - \int_0^\cdot \lambda(s,\theta^*)ds$ is a d-dimensional local martingale with respect to $\mathbf{F}$. Assume that the components of N have no common jumps. The parameter space Θ is a bounded open set in $\mathbb{R}^{\mathsf{p}}$ that admits Sobolev's inequality

$$\|f\|_\infty \le C_\Theta \sum_{i=0,1} \|\partial_\theta^i f\|_{L^r(\Theta)}$$

for elements f of the Sobolev space $f \in W^{1,r}(\Theta)$, with a constant C_Θ independent of f, for $r > \mathsf{p}$. We suppose that $0 \in \mathbb{R}^{\mathsf{p}}$ is in Θ and that the mapping $\theta \mapsto \lambda(t,\theta)$ is continuously extended to $\bar{\Theta}$.

We will use the quasi-likelihood method ([12]) with the quasi-log likelihood function

$$\ell_T(\theta) = \sum_{\alpha\in\mathbf{I}} \int_0^T \log\left(\lambda^\alpha(t,\theta)\right) dN_t^\alpha - \sum_{\alpha\in\mathbf{I}} \int_0^T \lambda^\alpha(t,\theta)dt. \tag{3.1}$$

Then $\mathcal{L}_T(\theta) = -\ell_T(\theta)$ becomes a loss function. The conditions stated later ensure the existence of the function (3.1). For the initial estimator $\tilde{\theta}$, we can use, for example, the quasi-maximum likelihood estimator $\tilde{\theta}^M$ and the quasi-Bayesian estimator $\tilde{\theta}^B$ given by

$$\tilde{\theta}^M \in \operatorname*{argmax}_{\theta\in\bar{\Theta}} \ell_T(\theta)$$

and

$$\tilde{\theta}^B = \left[\int_\Theta \exp(\ell_T(\theta))\pi(\theta)d\theta\right]^{-1} \int_\Theta \theta \exp(\ell_T(\theta))\pi(\theta)d\theta,$$

respectively, where π is a prior density satisfying $0 < \inf_\theta \pi(\theta) \leq \sup_\theta \pi(\theta) < \infty$.

For ergodic point processes, asymptotic normality and convergence of moments of $\tilde{\theta}^M$ and $\tilde{\theta}^B$ were proved in [12]. We recall their results briefly. Hereafter $\theta^* \in \Theta$ denotes the true value of θ and the distribution of the data is expressed by a multivariate point process N with intensity process $\lambda(t,\theta^*)$. For a random variable X, we denote $\|X\|_p = E[|X|^p]^{\frac{1}{p}}$. We write $C_\uparrow(\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^{\mathsf{p}})$ the set of functions f satisfying the following conditions : (i) f is continuous on $(\mathbb{R}_+ - \{0\}) \times (\mathbb{R}_+ - \{0\}) \times \mathbb{R}^{\mathsf{p}}$, (ii) for any $(u,v,w) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^{\mathsf{p}}$, $f(0,v,w) = f(u,0,w) = 0$ and (iii) f is of polynomial growth in $(u,v,w,\frac{1_{\{u>0\}}}{u},\frac{1_{\{v>0\}}}{v})$, i.e. there exists a constant $C_0 > 0$ and $m_1,m_2,m_3,m_4,m_5 \in \mathbb{N}$ such that $|f(u,v,w)| \leq C_0\big(1 + u^{m_1} + v^{m_2} + |w|^{m_3} + (\frac{1_{\{u>0\}}}{u})^{m_4} + (\frac{1_{\{v>0\}}}{v})^{m_5}\big)$ holds for all $(u,v,w) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^{\mathsf{p}}$.

Assumption 6. The mapping $\lambda : \Omega \times \mathbb{R}_+ \times \Theta \to \mathbb{R}_+^{\mathsf{d}}$ is $\mathcal{F} \times \mathbb{B}(\mathbb{R}_+) \times \mathbb{B}(\Theta)$-measurable and almost surely satisfies

(i) for every $\theta \in \Theta$, the mapping $s \mapsto \lambda(s,\theta)$ is left continuous,

(ii) for every $s \in \mathbb{R}_+$, the mapping $\theta \mapsto \lambda(s,\theta)$ is in $C^4(\Theta)$ and admits a continuous extension to $\bar{\Theta}$.

Assumption 7. **(i)** $\sup_{t\in\mathbb{R}_+} \sum_{i=0}^4 \left\| \sup_{\theta\in\Theta} \partial_\theta^i \lambda(t,\theta) \right\|_p < \infty$ for every $p > 1$.

(ii) $\sup_{t\in\mathbb{R}_+} \big\| \sup_{\theta\in\Theta} |\lambda^\alpha(t,\theta)^{-1} 1_{\{\lambda^\alpha(t,\theta)\neq 0\}}| \big\|_p < \infty$ for $p > 1$ and $\alpha \in \mathbf{I}$.

(iii) For any $\theta \in \Theta$ and $\alpha \in \mathbf{I}$, $\lambda^\alpha(t,\theta) = 0$ if and only $\lambda^\alpha(t,\theta^*) = 0$.

Assumption 8. For every $(\alpha,\theta) \in \mathbf{I}\times\Theta$, there exists a probability measure $\nu^\alpha(\cdot,\theta)$ on $\mathbb{R}_+\times\mathbb{R}_+\times\mathbb{R}^{\mathsf{p}}$ and $0 < \delta < \frac{1}{2}$ such that

$$\sup_{\theta\in\Theta} T^\delta \left\| \frac{1}{T}\int_0^T f\big(\lambda^\alpha(t,\theta^*), \lambda^\alpha(t,\theta), \partial_\theta\lambda^\alpha(t,\theta)\big)dt \right. \\ \left. - \int f(x,y,z)\nu^\alpha(dx,dy,dz,\theta) \right\|_p \to 0$$

as $T\to\infty$ for $p > 1$ and $f \in C_\uparrow(\mathbb{R}_+\times\mathbb{R}_+\times\mathbb{R}^{\mathsf{p}})$.

Let $\nu^\alpha(dx,dy,\theta) = \int_{\mathbb{R}^{\mathsf{p}}} \nu^\alpha(dx,dy,dz,\theta)$. Define $\mathbb{Y}_T(\theta)$ by

$$\mathbb{Y}_T(\theta) = \frac{1}{T}(\ell_T(\theta) - \ell_T(\theta^*)),$$

and $\mathbb{Y}(\theta)$ by the limit in probability of $\mathbb{Y}_T(\theta)$, where

$$\mathbb{Y}(\theta) = \sum_{\alpha\in\mathbf{I}} \int_{\mathbb{R}_+\times\mathbb{R}_+} 1_{\{x,y>0\}} \big\{ x\log(y/x) - (y-x) \big\} \nu^\alpha(dx,dy,\theta).$$

Remark. From the above expression of $\mathbb{Y}(\theta)$, we easily obtain $\mathbb{Y}(\theta^*) = 0$ and for all $\theta\in\Theta$,

$$\mathbb{Y}(\theta) \le 0. \tag{3.2}$$

Then Lemma 3.10 of [12] gives

$$\sup_{\theta\in\Theta} \big| \mathbb{Y}_T(\theta) - \mathbb{Y}(\theta) \big| \to^p 0$$

as $T\to\infty$.

The index χ_0 is defined by

$$\chi_0 = \inf_{\theta\in\Theta\setminus\{\theta^*\}} \frac{-\mathbb{Y}(\theta)}{|\theta-\theta^*|^2}.$$

Then identifiability is ensured by the condition

Assumption 9. $\chi_0 > 0$.

The Fisher information matrix is well defined by

$$\Gamma = \sum_{\alpha \in \mathbf{I}} \int_{\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^{\mathsf{p}}} z^{\otimes 2} x^{-1} 1_{\{x>0\}} \nu^{\alpha}(dx, dy, dz, \theta^*).$$

The matrix Γ is non-degenerate by Assumption 9.

By Theorem 3.14 of [12], we have

Theorem 3.1. Suppose that Assumptions 6-9 are satisfied. Then for $\tilde{\theta} = \tilde{\theta}^M$ and $\tilde{\theta}^B$, the convergence

$$\lim_{T \to \infty} E\big[\mathsf{f}\big(\sqrt{T}(\tilde{\theta} - \theta^*)\big)\big] = E\big[\mathsf{f}\big(\Gamma^{-1/2}\zeta\big)\big]$$

holds for all $f \in C(\mathbb{R}^{\mathsf{p}})$ of polynomial growth, where ζ is a p-dimensional standard normal random variable.

Now we are on the point of applying it to the penalized methods. Take $\tilde{\theta} = \tilde{\theta}^M$ or $\tilde{\theta}^B$. The penalized estimator will be denoted by $\hat{\theta}$. Let $r_T = T^{-\frac{1}{2}}$ and let

$$\hat{G} = -T^{-1}\partial_\theta^2 \ell_T(\tilde{\theta}) 1_{\{-\partial_\theta^2 \ell_T(\tilde{\theta}) \in \mathcal{S}_+\}} + T^{-1} I_{\mathsf{p}}$$

where $\mathcal{S}_+$ is the set of $\mathsf{p} \times \mathsf{p}$ positive definite symmetric matrices.

It is easy to show

$$\lim_{T \to \infty} \big\| T^{\delta}\big(\hat{G} - \Gamma\big) \big\|_p = 0$$

for every $p > 1$ and $0 < \delta < \frac{1}{2}$. Therefore the conditions in Theorems 2.1-2.5 are fulfilled in this situation.

3.1.1 Cox type of process with ergodic covariates

Regularization methods for Cox proportional hazards model are proposed by, for example, [9] and [23]. Here, we consider the multivariate point process N in Section 3.1 with intensity processes

$$\lambda^{\alpha}(t, \theta) = \exp\Big(\sum_{j \in \mathbf{J}} \theta_j^{\alpha} X_t^j\Big), \quad (\alpha \in \mathbf{I}) \tag{3.3}$$

where $\mathbf{J} = \{1, \ldots, J\}$ is an index set and $X^j = (X_t^j)_{t \in \mathbb{R}_+}$ $(j \in \mathbf{J})$ are left-continuous adapted stochastic covariate processes satisfying the following conditions.

Assumption 10. The J-dimensional process $(X^j)_{j\in\mathbf{J}}$ is stationary and $E[\exp(uX_0^j)] < \infty$ for all $u \in \mathbb{R}$ and $j \in \mathbf{J}$.

Denote by $\mathcal{B}_I$ the σ-field generated by $\{X_t^j;\ t \in I,\ j \in \mathbf{J}\}$ for $I \subset \mathbb{R}_+$. Let

$$\alpha(h) = \sup_{A\in\mathcal{B}_{[0,t]}, B\in\mathcal{B}_{[t+h,\infty)}} \big|P[A\cap B] - P[A]P[B]\big|$$

for $h > 0$.

Assumption 11. There exists $a > 0$ such that $\alpha(h) \le a^{-1}e^{-ah}$ for all $h > 0$.

Let $X_t = (X_t^j)_{j\in\mathbf{J}}$. For the model (3.3), $\theta = (\theta_j^\alpha)_{\alpha\in\mathbf{I}, j\in\mathbf{J}}$, $\mathsf{p} = \mathsf{d}J$ and

$$\hat{G} = \text{diag}\,[\hat{G}_1, \dots, \hat{G}_{\mathsf{d}}]$$

where

$$\hat{G}_\alpha = \frac{1}{T}\int_0^T X_t^{\otimes 2} \exp\Big(\sum_{j\in\mathbf{J}} \tilde{\theta}_j^\alpha X_t^j\Big) dt + \frac{1}{T} I_J. \tag{3.4}$$

It should be remarked that the first term on the right hand side of (3.4) may degenerate in general. Under Assumptions 10 and 11, we obtain $\hat{G} \to^p \Gamma$ for $\Gamma = \text{diag}\,[\Gamma_1(\theta^*), \dots, \Gamma_{\mathsf{d}}(\theta^*)]$, where

$$\Gamma_\alpha(\theta) = E\bigg[X_0^{\otimes 2} \exp\Big(\sum_{j\in\mathbf{J}} \theta_j^\alpha X_0^j\Big)\bigg].$$

Write $\Gamma(\theta) = \text{diag}\,[\Gamma_1(\theta), \dots, \Gamma_{\mathsf{d}}(\theta)]$.

Assumption 12. $\inf_{\theta\in\bar{\Theta}} \det \Gamma(\theta) > 0$.

We assume that Θ is an open bounded convex subset in $\mathbb{R}^{\mathsf{p}}$, that admits the Sobolev inequality in Section 3.1.

Lemma 3.2. Assumption 8 holds under Assumptions 10 and 11.

Proof. We remark that $\exp(|x|) < \exp(x) + \exp(-x)$ for all $x \in \mathbb{R}$. Thus, for all $(\theta, \alpha, j) \in \Theta \times \mathbf{I} \times \mathbf{J}$ and $p, q > 1$ and $t > 0$,

$$\begin{aligned} E\bigg[|X_t^j|^p \Big\{\exp(\theta_j^\alpha X_t^j)\Big\}^q\bigg] &\le E\Big[\exp(p|X_t^j|)\exp(q|\theta_j^\alpha||X_t^j|)\Big] \\ &= E\Big[\exp\Big\{\big(p + q|\theta_j^\alpha|\big)|X_0^j|\Big\}\Big] < C_{p,q}, \end{aligned} \tag{3.5}$$

where $C_{p,q}$ is a constant depend on p, q but not depending of θ, i, j. By the definition of $\lambda^\alpha(t,\theta)$, for all $\alpha \in \mathbf{I}$,

$$\lambda^\alpha(t,\theta) = \exp\Big(\sum_{j\in\mathbf{J}} \theta_j^\alpha X_t^j\Big),$$

$$\partial_{\theta^{\alpha'}}\lambda^\alpha(t,\theta) = \begin{cases} X_t \exp\Big(\sum_{j\in\mathbf{J}} \theta_j^\alpha X_t^j\Big) & \text{if } \alpha' = \alpha \\ 0 & \text{if } \alpha' \neq \alpha \end{cases},$$

where $\theta^\alpha = [\theta_j^\alpha]_j$. For $f \in C_\uparrow(\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^{\mathsf{p}})$, $\alpha \in \mathbf{I}$ and $\theta \in \Theta$, define $\tilde{f}_\theta^\alpha$ by

$$\tilde{f}_\theta^\alpha(x) = f(\mathrm{e}^{\sum_j \theta_j^{*\alpha} x_j}, \mathrm{e}^{\sum_j \theta_j^\alpha x_j}, \big[1_{\{\alpha'=\alpha\}} x \mathrm{e}^{\sum_j \theta_j^{\alpha'} x_j}\big]_{\alpha'\in\mathbf{I}})$$

for $x \in \mathbb{R}^J$. Then we can write for all $\alpha \in \mathbf{I}$,

$$\tilde{f}_\theta^\alpha(X_t) = f\big(\lambda^\alpha(t,\theta^*), \lambda^\alpha(t,\theta), \partial_\theta\lambda^\alpha(t,\theta)\big).$$

By (3.5), we obtain for all $\alpha \in \mathbf{I}$ and $p > 1$,

$$\sup_{\theta\in\Theta} E[|\tilde{f}_\theta^\alpha(X_t)|^p] = \sup_{\theta\in\Theta} E[|\tilde{f}_\theta^\alpha(X_0)|^p] < \infty.$$

Here, for $(\alpha,\theta) \in \mathbf{I}\times\Theta$ we define a probability measure $\nu^\alpha(\cdot,\theta)$ by

$$\nu^\alpha(A,\theta) = P\Big[\Big(\mathrm{e}^{\sum_j \theta_j^{*\alpha} X_0^j}, \mathrm{e}^{\sum_j \theta_j^\alpha X_0^j}, \big[1_{\{\alpha'=\alpha\}} X_0 \mathrm{e}^{\sum_j \theta_j^{\alpha'} X_0^j}\big]_{\alpha'\in\mathbf{I}}\Big) \in A\Big]$$

for $A \in \mathcal{B}(\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^{\mathsf{p}})$. Since

$$\int f(x,y,z)\nu^\alpha(dx,dy,dz,\theta) = E[\tilde{f}_\theta^\alpha(X_0)]$$

for all $\alpha \in \mathbf{I}$, we may show that for all $p > 1$, $\alpha \in \mathbf{I}$ and $\tilde{f}_\theta^\alpha$,

$$\sup_{\theta\in\Theta} T^{\frac{1}{2}} \left\| \frac{1}{T}\int_0^T \Big\{\tilde{f}_\theta^\alpha(X_t) - E[\tilde{f}_\theta^\alpha(X_0)]\Big\} dt \right\|_p = O(1).$$

By Assumption 10, there exists a constant C_0 such that

$$\left\| \int_{s_1}^{s_2} \Big(\tilde{f}_\theta^\alpha(X_t) - E[\tilde{f}_\theta^\alpha(X_0)]\Big) dt \right\|_p \leq C_0 (s_2 - s_1)^p$$

for $s_1 < s_2$. Then, Lemma 4 in [46] implies under Assumption 11 that

$$\begin{aligned}
&E\left[\left|\int_0^T \Big(\tilde{f}_\theta^\alpha(X_t) - E[\tilde{f}_\theta^\alpha(X_0)]\Big)dt\right|^p\right] \\
&= E\left[\left|\sum_{l=1}^{\lfloor T\rfloor} \int_{\frac{(l-1)T}{\lfloor T\rfloor}}^{\frac{lT}{\lfloor T\rfloor}} \Big(\tilde{f}_\theta^\alpha(X_t) - E[\tilde{f}_\theta^\alpha(X_0)]\Big)dt\right|^p\right] \\
&\leq C_1\lfloor T\rfloor^{\frac{p}{2}} + C_2\lfloor T\rfloor \\
&= O(T^{\frac{p}{2}}),
\end{aligned}$$

for $T \geq 1$ and $p \geq 2$, where C_1 and C_2 are constants depending on a and p. Therefore, we have

$$\sup_{\theta\in\Theta} T^{\frac{1}{2}} \left\| \frac{1}{T}\int_0^T \Big(\tilde{f}_\theta^\alpha(X_t) - E[\tilde{f}_\theta^\alpha(X_0)]\Big)dt \right\|_p = O(1).$$

□

Next, we will give a sufficient condition for Assumption 9.

Lemma 3.3. We assume that Θ is convex. Then Assumption 9 follows from Assumptions 10 and 12.

Proof. By the definition of $\mathbb{Y}(\theta)$,

$$\mathbb{Y}(\theta) = \sum_{\alpha\in\mathbf{I}} \mathbb{Y}^\alpha(\theta),$$

where $\mathbb{Y}^\alpha(\theta)$ is given by

$$\begin{aligned}
\mathbb{Y}^\alpha(\theta) = E\Bigg[&\exp\Big(\sum_{j\in\mathbf{J}} \theta_j^{*\alpha} X_0^j\Big)\Big(\sum_{j\in\mathbf{J}} (\theta_j^\alpha - \theta_j^{*\alpha}) X_0^j\Big) \\
&-\Big\{\exp\Big(\sum_{j\in\mathbf{J}} \theta_j^\alpha X_0^j\Big) - \exp\Big(\sum_{j\in\mathbf{J}} \theta_j^{*\alpha} X_0^j\Big)\Big\}\Bigg]
\end{aligned}$$

for $\alpha \in \mathbf{I}$. Thus we have

$$\partial_\theta \mathbb{Y}(\theta) = \begin{bmatrix} \partial_{\theta^1}\mathbb{Y}^1(\theta) \\ \partial_{\theta^2}\mathbb{Y}^2(\theta) \\ \vdots \\ \partial_{\theta^{\mathsf{d}}}\mathbb{Y}^{\mathsf{d}}(\theta) \end{bmatrix},$$

where

$$\partial_{\theta^\alpha}\mathbb{Y}^\alpha(\theta) = E\left[\left\{\exp\Big(\sum_{j\in\mathbf{J}}\theta^{*\alpha}_j X^j_0\Big) - \exp\Big(\sum_{j\in\mathbf{J}}\theta^{\alpha}_j X^j_0\Big)\right\}X_0\right].$$

Similarly,

$$\partial^2_\theta\mathbb{Y}(\theta) = \mathrm{diag}\Big[\partial^2_{\theta^1}\mathbb{Y}^1(\theta), \partial^2_{\theta^2}\mathbb{Y}^2(\theta), \ldots, \partial^2_{\theta^{\mathsf{d}}}\mathbb{Y}^{\mathsf{d}}(\theta)\Big],$$

where

$$\partial^2_{\theta^\alpha}\mathbb{Y}^\alpha(\theta) = -E\Big[X_0^{\otimes 2}\exp\Big(\sum_{j\in\mathbf{J}}\theta^{\alpha}_j X^j_0\Big)\Big] = -\Gamma_\alpha(\theta).$$

Therefore, we have

$$\partial^2_\theta\mathbb{Y}(\theta) = -\Gamma(\theta).$$

By Assumption 12, for all $\theta \in \bar{\Theta}$, $-\partial^2_\theta\mathbb{Y}(\theta)$ is positive definite. Therefore, for all $\theta \in \Theta$, there exists a vector $\theta^\dagger$ satisfying

$$-\mathbb{Y}(\theta) = -\frac{1}{2}\partial^2_\theta\mathbb{Y}(\theta^\dagger)[(\theta-\theta^*)^{\otimes 2}].$$

By positive definiteness of $-\partial^2_\theta\mathbb{Y}(\theta)$ and Assumption 10,

$$\begin{aligned}\inf_{\theta\in V(\theta^*)\setminus\{\theta^*\}}\frac{-\mathbb{Y}(\theta)}{|\theta-\theta^*|^2} &= \inf_{\theta\in V(\theta^*)\setminus\{\theta^*\}}\frac{-\partial^2_\theta\mathbb{Y}(\theta^\dagger)[(\theta-\theta^*)^{\otimes 2}]}{2|\theta-\theta^*|^2}\\ &\geq \inf_{\theta\in V(\theta^*)}\frac{1}{2}\tau_{\min}(\Gamma(\theta)) > 0. \qquad (3.6)\end{aligned}$$

Therefore, by (3.6), we obtain $\chi_0 > 0$. □

Assumption 6 follows from (3.3), and Assumption 7 follows from Assumption 10. Thus the conditions in Theorems 2.1-2.5 are fulfilled under Assumptions 9-11.

3.1.2 Hawkes process

Hawkes (1971a[21], 1971b[22]) introduced a multivariate model for point processes with mutually exciting components now referred to as the Hawkes model. Ordinary, it was motivated by modeling after shocks and seismological phenomena (cf. Vere-Jones 1970 [40], Vere-Jones and Ozaki 1982 [41], Ogata 1999 [32]). However, the usage of the Hawkes model has been more and more spread out to various research areas : high-frequency finance (c.f. [3]), crime activity (c.f. [30]) and analysis of social networks (c.f. [13], [7], [51], [45]).

Example. For example, consider a finite network with d nodes (each node corresponding to a user in a social network). For each node α in $\{1, \dots, \mathsf{d}\}$, we observe the timestamps $\{t_{\alpha,1}, t_{\alpha,2}, \dots\}$ of actions of node α on the network (a message, a click etc.). To each node α is associated a counting process $N^\alpha(t) = \sum_{i \geq 1} 1_{\{t_{\alpha,i} \leq t}$. If $N = (N^\alpha)_\alpha$ is a Hawkes process with intensity (3.7), then we can quantify the influence of β on α by the function $h_{\alpha\beta}(t)$.

Now we turn to the definition of a Hawkes process. Let $N_t = (N_t^\alpha)_{\alpha \in \mathbf{I}}, \mathbf{I} = \{1, \dots, \mathsf{d}\}$, $N_0 = 0$, be a multidimensional point process and write $\mathbf{F} = (\mathcal{F}_t^N)_{t \in \mathbb{R}_+}$, where we call that $\mathcal{F}_t^N = \sigma\{N_s | 0 \leq s \leq t\}$ is the canonical filtration of N. We say that N is a linear Hawkes process or Hawkes' self-exciting process starting from 0 if there exist functions $h_{\alpha\beta} : \mathbb{R}_+ \to \mathbb{R}_+^{\mathsf{d} \times \mathsf{d}}$ and $\nu \in \mathbb{R}^{\mathsf{d}}$ such that the $\mathcal{F}^N$-intensity $\lambda(t)$ of N writes

$$\lambda^\alpha(t) = \nu_\alpha + \sum_{\beta \in \mathbf{I}} \int_0^{t-} h_{\alpha\beta}(t-s) dN_s^\beta, \quad \alpha \in \mathbf{I}. \tag{3.7}$$

The baseline intensities ν_α's represent the rate of spontaneous occurrences of events, while the kernels $h_{\alpha\beta}$'s model self-interaction in the system. Indeed, if a shock occurs at time t_0 on the covariate N^β, an aftershock will happen on the covariate N^α around time t_1 with high probability if $h_{\alpha\beta}(t_1 - t_0)$ is large. When $h_{\alpha\beta} = 0$, covariate N^β has no influence on the chain of events related to N^α.

Let us define the matrix $\Phi = [\phi_{\alpha\beta}]_{\alpha\beta}$ where

$$\phi = \int_0^\infty h_{\alpha\beta}(s) ds,$$

and write $\rho(\Phi)$ its spectral radius. Finally, given $A = [a_{\alpha\beta}]_{\alpha\beta}$ and $C = [c_{\alpha\beta}]_{\alpha\beta} \in \mathbb{R}_+^{\mathsf{d} \times \mathsf{d}}$, we say that N is an exponential Hawkes process if the kernel functions $h_{\alpha\beta}$ are of the form

$$h_{\alpha\beta}(s) = c_{\alpha\beta} e^{-a_{\alpha\beta} s}.$$

Note that the matrix Φ has the representation $\Phi = [\frac{c_{\alpha\beta}}{a_{\alpha\beta}}]_{\alpha\beta}$ in this case. Hereafter, we assume $\rho(\Phi) < 1$, where $\rho(\Phi)$ is a spectral radius of Φ.

Now, we consider the exponential Hawkes process. Then the intensity has the following representation.

$$\lambda^\alpha(t) = \nu_\alpha + \sum_{\beta \in \mathbf{I}} \int_0^{t-} c_{\alpha\beta} e^{-a_{\alpha\beta}(t-s)} dN_s^\beta, \quad \alpha \in \mathbf{I}. \tag{3.8}$$

Here let $\theta = (\nu, C, A) \in \mathbb{R}_+^{\mathsf{d}} \times \mathbb{R}_+^{\mathsf{d} \times \mathsf{d}} \times \mathbb{R}_+^{\mathsf{d} \times \mathsf{d}}$. First, we consider the usual case, i.e., the true value of the parameter $\theta^* = (\nu^*, C^*, A^*)$ is interior point of Θ.

Assumption 13. For every $(\alpha, \theta) \in \mathbf{I} \times \Theta$, there exists a probability measure $\nu^\alpha(\cdot, \theta)$ on $\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^{\mathsf{p}}$ and $0 < \delta < \frac{1}{2}$ such that

$$\frac{1}{T}\int_0^T f\big(\lambda^\alpha(t,\theta^*), \lambda^\alpha(t,\theta), \partial_\theta \lambda^\alpha(t,\theta)\big)dt \to^p \int f(x,y,z)\nu^\alpha(dx,dy,dz,\theta)$$

as $T \to \infty$ for $f \in C_b(\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^{\mathsf{p}})$.

Assumption 14. For any $\theta \in \bar{\Theta} - \{\theta^*\}, \mathbb{Y}(\theta) \neq 0$.

Proposition 3.4 (Theorem 3.9. of Clinet and Yoshida 2017 ([12]))**.** Under the Assumptions 6, 7, 13 and 14, the QMLE $\tilde{\theta}$ is consistent.

$$\tilde{\theta} \to^p \theta^*.$$

Next, we are interested in the case that some component of Φ^*, the true value of Φ, is precisely 0. Then the graph structure has a form like the left side of Figure 3.1 in contrast with the right side which represents the full model.

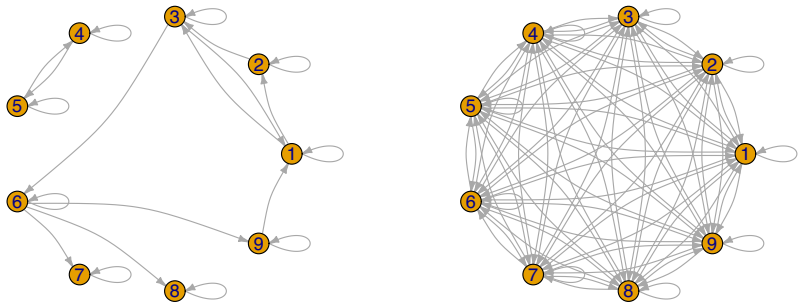

Figure 3.1: Graph structure; sparse model (left) and full model (right)

Now we consider the case where $\phi^*_{\alpha\beta} = 0$ for some (α, β). Let

$$\mathcal{J}^* = \{(\alpha,\beta) \in \mathbf{I} \times \mathbf{I}; c^*_{\alpha\beta} = 0\}.$$

For a constant $\xi \in \mathbb{R}_+$ we define a function $p^\xi : \Theta \to \Theta$ by $p^\xi(\theta) = p^\xi(\nu, C, A) = (\nu, C, p_0^\xi(A))$, where $p_0^\xi(A) = [a_{\alpha\beta}^{(\xi)}]_{\alpha\beta}$ and

$$a_{\alpha\beta}^{(\xi)} = \begin{cases} a_{\alpha\beta} & (\alpha, \beta) \in \mathcal{J}^* \\ \xi & (\alpha, \beta) \notin \mathcal{J}^*. \end{cases}$$

If $c_{\alpha\beta}^* = 0$ for some (α, β), then the value of $a_{\alpha\beta}^*$ does not affect the model and Proposition 3.4 does not hold. Thus we modify the Assumption 14.

Assumption 15. For any $\theta \in \bar{\Theta} - \{p^\xi(\theta^*) \in \Theta; \xi \in \mathbb{R}_+\}, \mathbb{Y}(\theta) \neq 0$.

Proposition 3.5. Under the assumptions 6, 7, 13 and 15, (the QMLE) $p^\xi(\tilde{\theta})$ is consistent.

$$p^\xi(\tilde{\theta}) \to^p p^\xi(\theta^*)$$

In particular,

$$\tilde{C} \to^p C^*.$$

Since the structure of C (or Φ) determine the model, define the pLSA estimator $\hat{C}$ by

$$\hat{C} \in \mathrm{argmin}_C \hat{G}[(C - \tilde{C})^{\otimes 2}] + \sum_{\alpha,\beta} \kappa_T^{\alpha\beta} |c_{\alpha\beta}|^q,$$

where $\kappa_T^{\alpha\beta} = \alpha_T(|c_{\alpha\beta}| + \eta_T)^{-\gamma}, \eta_T \to 0 (T \to \infty)$ and $\hat{G} = -\frac{1}{T}\partial_C^2 \ell_T(\theta)$. Then by the Theorem 2.2, we obtain the selection consistency of the pLSA estimator $\hat{C}$.

Remark. By the Theorem 4.6. of [12], the exponential Hawkes model verifies the Assumptions 6-9 in the usual situation. By using the discussion in Appendix, same results are possibly derived in the situation where the true value of parameter $\theta^* \in \partial\Theta$. However, since it is difficult to treat that case, we omit the detailed description here.

3.2 Diffusion process

3.2.1 Ergodic case

Given a stochastic basis $(\Omega, \mathcal{F}, \mathbf{F}, P)$, $\mathbf{F} = (\mathcal{F}_t)_{t\in\mathbb{R}_+}$, we consider a d-dimensional process $X = (X_t)_{t\in\mathbb{R}_+}$ adapted to the filtration $\mathbf{F} = (\mathcal{F}_t)_{t\in\mathbb{R}_+}$ and satisfying the following stochastic integral equation

$$X_t = X_0 + \int_0^t a(X_s, \theta_2) ds + \int_0^t b(X_s, \theta_1) dW_s, \quad t \in \mathbb{R}_+$$

where W is an r-dimensional standard $\mathbf{F}$-Wiener process, $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2 = \Theta$ with Θ_1 and Θ_2 being bounded domains of $\mathbb{R}^{\mathsf{p}_1}$ and $\mathbb{R}^{\mathsf{p}_2}$, respectively, moreover $b : \mathbb{R}^d \times \Theta_1 \to \mathbb{R}^d \otimes \mathbb{R}^r$ and $a : \mathbb{R}^d \times \Theta_2 \to \mathbb{R}^d$. We define the function B by $B(x, \theta_1) = b(x, \theta_1)b(x, \theta_1)'$ and assume that $B(x, \theta_1)$ is invertible. We denote the true value of $\theta = (\theta_1, \theta_2)$ by $\theta^* = (\theta_1^*, \theta_2^*)$ and the number of active parameters of θ_k^* by p_k^0 for $k = 1, 2$. We assume that each parameter space have a locally Lipschitz boundary.

In this subsection, we assume that the process X is ergodic. That is, there exists a unique invariant probability measure $\mu = \mu_{\theta^*}$ such that for any bounded measurable function $g : \mathbb{R}^d \to \mathbb{R}$, the convergence

$$\frac{1}{T}\int_0^T g(X_t)dt \to^p \int_{\mathbb{R}^d} g(x)\mu(dx)$$

holds.

We suppose that $0 \in \mathbb{R}^{\mathsf{p}_1+\mathsf{p}_2}$ is in Θ. Here we have the discrete-time observations $(X_{t_i}, Y_{t_i})_{i=0}^n$ where $t_i = ih$ with $h = h_n$ depending on n. We will consider the situation when $h_n \to 0$ and $nh_n^p \to 0$ as $n \to \infty$, and there exists $\epsilon_0 \in (0, \frac{p-1}{p})$ such that $n^{\epsilon_0} \le nh_n$ for large n.

Here, we assume the following properties of an initial estimator $\tilde{\theta} = (\tilde{\theta}_{1,n}, \tilde{\theta}_{2,n})$:

$$\begin{aligned}&(\sqrt{n}(\tilde{\theta}_{1,n} - \theta_1^*), \sqrt{nh}(\tilde{\theta}_{2,n} - \theta_2^*))\\ &\qquad \to^d (\Gamma_1^{-\frac{1}{2}}\zeta_1, \Gamma_2^{-\frac{1}{2}}\zeta_2) \sim N_{\mathsf{p}_1+\mathsf{p}_2}(0, \mathrm{diag}(\Gamma_1^{-1}, \Gamma_2^{-1}))\end{aligned}$$

and

$$\sup_n \left(\left\| \sqrt{n}(\tilde{\theta}_{1,n} - \theta_1^*) \right\|_p + \left\| \sqrt{nh}(\tilde{\theta}_{2,n} - \theta_2^*) \right\|_p \right) < \infty$$

for every $p > 1$, where ζ_1 and ζ_2 are p_1 and p_2-dimensional standard normal variables respectively, and $\Gamma_1 = (\Gamma_1^{ij})_{i,j=1,\dots,\mathsf{p}_1}$ and $\Gamma_2 = (\Gamma_2^{ij})_{i,j=1,\dots,\mathsf{p}_2}$ with

$$\begin{aligned}\Gamma_1^{ij} &= \frac{1}{2}\int_{\mathbb{R}^d} \mathrm{Tr}\Big\{ \big(\partial_{\theta_1^i} B(x, \theta_1^*)\big) B^{-1}(x, \theta_1^*) \big(\partial_{\theta_1^j} B(x, \theta_1^*)\big) B^{-1}(x, \theta_1^*) \Big\} \mu(dx),\\ \Gamma_2^{ij} &= \int_{\mathbb{R}^d} (\partial_{\theta_2^i} a(x, \theta_2^*)' B(x, \theta_1^*)^{-1} \partial_{\theta_2^j} a(x, \theta_2^*)) \mu(dx).\end{aligned}$$

We assume integrability and non-degeneracy of Γ_1 and Γ_2. It is known that the quasi-maximum likelihood estimator, the quasi-Bayesian estimator and the hybrid type estimators possess these properties under certain mild conditions ([46], [37], [39], [25]). For instance, if we use the hybrid multistep estimator $\tilde{\theta}^H = (\tilde{\theta}_{1,n}^H, \tilde{\theta}_{2,n}^H)$ by Uchida and Kamatani ([25]) as an initial estimator $\tilde{\theta}$, then above conditions are satisfied by Theorem 1 of [25].

For $q_1, q_2 \in (0,1]$, we define the objective functions $Q_{1,n}^{(q_1)}$ and $Q_{2,n}^{(q_2)}$ by

$$Q_{1,n}^{(q_1)} = \hat{G}_{1,n}[(\theta_1 - \tilde{\theta}_{1,n})^{\otimes 2}] + \sum_{i_1=1}^{\mathsf{p}_1} \kappa_{1,n}^{i_1} |\theta_1^{i_1}|^{q_1}$$

and

$$Q_{2,n}^{(q_2)} = \hat{G}_{2,n}[(\theta_2 - \tilde{\theta}_{2,n})^{\otimes 2}] + \sum_{i_2=1}^{\mathsf{p}_2} \kappa_{2,n}^{i_2} |\theta_2^{i_2}|^{q_2},$$

respectively, where $\hat{G}_{k,n}$ $(k = 1, 2)$ are some $\mathsf{p}_k \times \mathsf{p}_k$ random matrices such that $\hat{G}_{k,n} \to^p \Gamma_k$ and that the family $\left\{|\hat{G}_{k,n}| + \left(\det \hat{G}_{k,n}\right)^{-1}\right\}_{k,n}$ is $L^{\infty-}$-bounded, and $\kappa_{k,n}^{i_k} = \alpha_{k,n}|\tilde{\theta}_{k,n}^{i_k}|^{-\gamma_k}$ for some numbers $\gamma_k > -(1-q_k)$ and some sequences $(\alpha_{1,n})_n$ and $(\alpha_{2,n})_n$ satisfying

$$(\sqrt{n})^{2-q_1+\gamma_1}\alpha_{1,n} \to \infty, \sqrt{n}\alpha_{1,n} \to 0$$

and

$$(\sqrt{nh})^{2-q_2+\gamma_2}\alpha_{2,n} \to \infty, \sqrt{nh}\alpha_{2,n} \to 0$$

respectively. Then we have the penalized LSA estimators $\hat{\theta}_{1,n}^{(q_1)}$ and $\hat{\theta}_{2,n}^{(q_2)}$ satisfying

$$\hat{\theta}_{1,n}^{(q_1)} \in \operatorname*{argmin}_{\theta_1 \in \bar{\Theta}_1} Q_{1,n}^{(q_1)}(\theta_1)$$

and

$$\hat{\theta}_{2,n}^{(q_2)} \in \operatorname*{argmin}_{\theta_2 \in \bar{\Theta}_2} Q_{2,n}^{(q_2)}(\theta_2).$$

For these estimators $\hat{\theta}_{1,n}^{(q_1)}$ and $\hat{\theta}_{2,n}^{(q_2)}$, Theorems 2.1-2.5 hold respectively. Additionally, we consider the limit distribution of the joint variable $((\hat{\theta}_{1,n}^{(q_1)})_{\mathcal{J}_1^1}, (\hat{\theta}_{2,n}^{(q_2)})_{\mathcal{J}_2^1})$. Here, $\mathcal{J}_k^1, \mathcal{J}_k^{11}$ and $\mathcal{J}_k^{10}$ are defined similarly to $\mathcal{J}^1, \mathcal{J}^{11}$ and $\mathcal{J}^{10}$, respectively, for each $k = 1, 2$.

Now we can rephrase Theorems 1-5 in the present situation. In particular,

Proposition 3.6. The convergence

$$\begin{aligned}\left(\sqrt{n}(\hat{\theta}_{1,n}^{(q_1)} - \theta_1^*)_{\mathcal{J}_1^1}, \sqrt{nh}(\hat{\theta}_{2,n}^{(q_2)} - \theta_2^*)_{\mathcal{J}_2^1}\right) &\to^d \left(\mathfrak{G}_1 {\Gamma_1}^{-\frac{1}{2}}\zeta_1, \mathfrak{G}_2 {\Gamma_2}^{-\frac{1}{2}}\zeta_2\right) \\ &\sim N_{\mathsf{p}_1^0+\mathsf{p}_2^0}\left(0, \operatorname{diag}\left(((\Gamma_1)_{\mathcal{J}_1^{11}})^{-1}, ((\Gamma_2)_{\mathcal{J}_2^{11}})^{-1}\right)\right)\end{aligned}$$

holds, where $\mathfrak{G}_k = \left[I_{\mathsf{p}_k^0} \quad ((\Gamma_k)_{\mathcal{J}_k^{11}})^{-1}(\Gamma_k)_{\mathcal{J}_k^{10}}\right], k = 1, 2.$

Proof. By Theorem 2.3, we have

$$\sqrt{n}(\hat{\theta}^{(q_1)}_{1,n} - \theta^*_1)_{\mathcal{J}^1_1} - \mathfrak{G}_1\big\{\sqrt{n}(\tilde{\theta}_{1,n} - \theta^*_1)_{\mathcal{J}^1_1}\big\} \to^p 0$$

and

$$\sqrt{nh}(\hat{\theta}^{(q_2)}_{2,n} - \theta^*_2)_{\mathcal{J}^1_2} - \mathfrak{G}_2\big\{\sqrt{nh}(\tilde{\theta}_{2,n} - \theta^*_2)_{\mathcal{J}^1_2}\big\} \to^p 0.$$

Therefore,

$$\begin{bmatrix} \sqrt{n}(\hat{\theta}^{(q_1)}_{1,n} - \theta^*_1)_{\mathcal{J}^1_1} \\ \sqrt{nh}(\hat{\theta}^{(q_2)}_{2,n} - \theta^*_2)_{\mathcal{J}^1_2} \end{bmatrix} \to^d \begin{bmatrix} \mathfrak{G}_1 \Gamma_1^{-\frac{1}{2}} \zeta_1 \\ \mathfrak{G}_2 \Gamma_2^{-\frac{1}{2}} \zeta_2 \end{bmatrix}$$
$$\sim N_{\mathsf{p}^0_1+\mathsf{p}^0_2}\bigg(0, \operatorname{diag}\Big(((\Gamma_1)_{\mathcal{J}^{11}_1})^{-1}, ((\Gamma_2)_{\mathcal{J}^{11}_2})^{-1}\Big)\bigg)$$

□

3.2.2 Non-ergodic case

In this subsection, we will deal with the case where the Fisher information matrix is not deterministic. We consider the following stochastic regression model

$$Y_t = Y_0 + \int_0^t b_s ds + \int_0^t \sigma(X_s, \theta) dW_s, \quad t \in [0, T], \tag{3.9}$$

where W is an r-dimensional standard Wiener process independent of the initial value of Y_0, X and b are progressively measurable processes with values in $\mathbb{R}^{\mathsf{d}}$ and $\mathbb{R}^{\mathsf{m}}$, respectively. σ is an $\mathbb{R}^{\mathsf{m}} \otimes \mathbb{R}^{\mathsf{r}}$-valued measurable function defined on $\mathbb{R}^{\mathsf{d}} \times \Theta$, and Θ is a bounded domain in $\mathbb{R}^{\mathsf{p}}$ with a locally Lipschitz boundary. Additionally, we define $S = \sigma^{\otimes 2} = \sigma\sigma'$. The data set consists of discrete observations $(X_{t_j}, Y_{t_j})_{j=0}^n$ with $t_j = jT/n$ and T is fixed.

Here, we assume that there exists an estimator $\tilde{\theta}_n$ of θ^* such that

$$\sqrt{n}(\tilde{\theta}_n - \theta^*) \to^{d_s} \Gamma^{-\frac{1}{2}} \zeta$$

as $n \to \infty$, and for any continuous functions $\mathsf{f} : \mathbb{R}^{\mathsf{p}} \to \mathbb{R}$ of at most polynomial growth,

$$E[\mathsf{f}(\sqrt{n}(\tilde{\theta}_n - \theta^*))] \to E[\mathsf{f}(\Gamma^{-\frac{1}{2}}\zeta)],$$

where Γ is the Fisher information matrix given by

$$\Gamma = \frac{1}{2T} \int_0^T \operatorname{Tr}\Big((\partial_\theta S) S^{-1} (\partial_\theta S) S^{-1}(X_t, \theta^*)\Big) dt,$$

ζ is a p-dimensional standard normal random variable independent of Γ and $\to^{d_s}$ means the $\sigma(\Gamma)$-stable convergence in distribution. Here we remark that the Fisher information matrix Γ is not necessarily deterministic. In fact, Uchida and Yoshida [38] proved that the quasi-maximum likelihood estimator and the quasi-Bayesian estimator have these properties under mild regularity conditions. An essential condition in their argument is the non-degeneracy of a key index χ_0:

Assumption 16. For every $L > 0$, there exists $c_L > 0$ such that

$$P[\chi_0 \leq r^{-1}] \leq \frac{c_L}{r^L} \qquad (r > 0)$$

where

$$\chi_0 = \inf_{\theta \neq \theta^*} \frac{-\mathbb{Y}(\theta)}{|\theta - \theta^*|^2}$$

with

$$\mathbb{Y}(\theta) = -\frac{1}{2T} \int_0^T \left\{ \log\Big(\frac{\det S(X_t, \theta)}{\det S(X_t, \theta^*)} \Big) + \mathrm{Tr}\Big(S^{-1}(X_t, \theta) S(X_t, \theta^*) - I_d \Big) \right\} dt.$$

For the initial estimator, for example, we can take the maximum likelihood type estimator $\tilde{\theta}_n^M$ that satisfies

$$\mathbb{H}_n(\tilde{\theta}_n^M) = \sup_{\theta \in \Theta} \mathbb{H}_n(\theta),$$

or the Bayes type estimator $\tilde{\theta}_n^B$ for a prior density $\pi : \Theta \to \mathbb{R}_+$ with respect to the quadratic loss defined by

$$\tilde{\theta}_n^B = \Big(\int_\Theta \exp(\mathbb{H}_n(\theta)) \pi(\theta) d\theta \Big)^{-1} \int_\Theta \theta \exp(\mathbb{H}_n(\theta)) \pi(\theta) d\theta,$$

where $\mathbb{H}_n(\theta)$ is a quasi-log likelihood function defined by

$$\mathbb{H}_n(\theta) = -\frac{1}{2} \sum_{i=1}^n \Big\{ \log \det S(X_{t_{i-1}}, \theta) + \frac{1}{h} S(X_{t_{i-1}}, \theta)^{-1} [(\Delta_i Y)^{\otimes 2}] \Big\}. \quad (3.10)$$

Then we can use the QLA method in [38] to show the stable convergence and the L^p-boundedness of the estimators. In order to verify Assumption 16 in practice, we may apply one of criteria given in [38].

Here we define the objective function

$$Q_n^{(q)}(\theta) = \hat{G}_n[(\theta - \tilde{\theta}_n)^{\otimes 2}] + \sum_{i=1}^{\mathsf{p}} \kappa_n^i |\theta_i|^q$$

and penalized LSA estimator $\hat{\theta}_n^{(q)} \in \operatorname*{argmin}_{\theta \in \bar{\Theta}} Q_n^{(q)}(\theta)$. We can take

$$\hat{G}_n = -\frac{1}{n}\partial_\theta^2 \mathbb{H}_n(\tilde{\theta}^M) 1_{\left\{-\partial_\theta^2 \mathbb{H}_n(\tilde{\theta}^M) \in \mathcal{S}_+\right\}} + \frac{1}{n} I_{\mathsf{p}} \tag{3.11}$$

when we use the QMLE as an initial estimator. Let $\kappa_n^i = \alpha_n |\tilde{\theta}_{n,i}|^{-\gamma}$ for the number $\gamma > -(1-q)$ and the sequence α_n satisfying the conditions

$$(\sqrt{n})^{2-q+\gamma}\alpha_n \to \infty, \quad \sqrt{n}\alpha_n \to 0.$$

Similarly to the previous sections, we can show that Theorems 2.1-4 hold for this penalized LSA estimator. On the other hand, we should choose $\hat{G}_n = I_{\mathsf{p}}$, in place of (3.11), for the P-O estimator.

Chapter 4

Simulations

In this section we report three simulations. The first one is the Cox process in Section 3.1.1, the second one is the Hawkes process in Section 3.1.2 and the third one is the non-ergodic diffusion type process in Section 3.2.2. For each simulation we perform 1000 Monte Carlo replications. %(*method*) in this chapter denotes the number of times, in percentage over 1000 iterations, that the estimator obtained by the *method* chooses the true model. In the first and third cases, for comparison we use the unified LASSO type estimator and the Bridge type estimator. The unified LASSO type estimator is the special case of penalized LSA estimator where $q = 1$. The Bridge estimator is not the special case of penalized LSA estimator, but we use this phraseology when the penalty has the form $r_T^{-1} \sum_i |\theta_i|^q$, i.e., $r = 1, \gamma = 0$ and $q < 1$.

For the convenience of calculation, in the first and third cases, we use identity matrix as a coefficient matrix $\hat{G}$. Thus, the penalized LSA estimator is not efficient, and we use the P-O estimator in order to obtain the efficient estimator.

It has been shown through the simulation studies that the penalized LSA estimator can select the correct model if we choose appropriate tuning parameters and the P-O estimator has good performance for the active parameters. It is an important thing to give a "good" tuning parameter, however, we do not refer to how to select the tuning parameter.

4.1 Simulation for the Cox model

We consider the Cox model (3.3) in Section 3.1.1 with $\alpha = 1$. Let $\mathsf{p} = 20$, then the parameter space Θ is $[-10, 10]^{20}$. The covariate process $X = (X_t)_{t\in[0,T]}$ is a 20-dimensional OU process satisfying the following stochastic differential

equations

$$dX_t^i = -\mathsf{a}_i X_t^i dt + 0.4 dW_t^i, \quad X_0 = 0, \quad t \in [0, T]$$

where a_i $(i = 1, \ldots, 20)$ are constants given by

$$\begin{aligned}
\mathsf{a}_1 = \mathsf{a}_6 =& \mathsf{a}_{11} = \mathsf{a}_{16} = 0.15, \\
\mathsf{a}_2 = \mathsf{a}_7 =& \mathsf{a}_{12} = \mathsf{a}_{17} = 0.2, \\
\mathsf{a}_3 = \mathsf{a}_8 =& \mathsf{a}_{13} = \mathsf{a}_{18} = 0.25, \\
\mathsf{a}_4 = \mathsf{a}_9 =& \mathsf{a}_{14} = \mathsf{a}_{19} = 0.3, \\
\mathsf{a}_5 = \mathsf{a}_{10} =& \mathsf{a}_{15} = \mathsf{a}_{20} = 0.35.
\end{aligned}$$

and $W = (W^i)_{i=1,\ldots,20}$ is a 20-dimensional standard Wiener process. Figure 4.1 shows a sample path of the covariate process. Data $N = (N_t)_{t\in[0,T]}$ is a sample path of the point process with intensity $\lambda(t, \theta^*)$ in (3.3), where the true values θ^* of the parameter is

$$\theta^* = [2, -1, 1, -0.5, -1.5, 1.5, 0.5, 0.75, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]'.$$

Figure 4.2 and 4.3 show sample paths of intensity process $\lambda(t, \theta^*)$ and counting process N_t on $[0, 200]$ and $[0, 3]$ respectively.

Let $\mathcal{L}_T(\theta) = \mathbb{L}_T(\theta) = -\ell_T(\theta)$ in (3.1) and we use QMLE for the initial estimator $\tilde{\theta}$. Then the objective function is denoted by

$$Q_T^{(q)}(\theta) = (\theta - \tilde{\theta})'(\theta - \tilde{\theta}) + \sum_{j=1}^{20} \kappa_T^j |\theta_j|^q.$$

where $\kappa_T^j = \alpha_T |\tilde{\theta}_j|^{-\gamma}, \alpha_T = (\frac{1}{\sqrt{T}})^r$, $1 < r < 2 - q + \gamma$. Let the triplet of tuning parameters $(\gamma, r, q) = (1, 1.2, 0.3)$. We will consider the cases $T = 50, 100, 200$ and 400. Table 1 compares the results of the variable selection of the penalized LSA estimator, the unified LASSO type estimator and the Bridge type estimator. Here, we define the unified LASSO type estimator and the Bridge type estimator as the penalized LSA estimator with tuning parameter $(\gamma, r, q) = (1, 1.2, 1)$ and $(\gamma, r, q) = (0, 1, 0.3)$, respectively.

Table 2 compares the means and standard deviations (parentheses) for the three estimators (initial estimator, penalized LSA estimator and P-O estimator) and shows the results of the variable selection for penalized LSA estimator in the case $T = 200$.

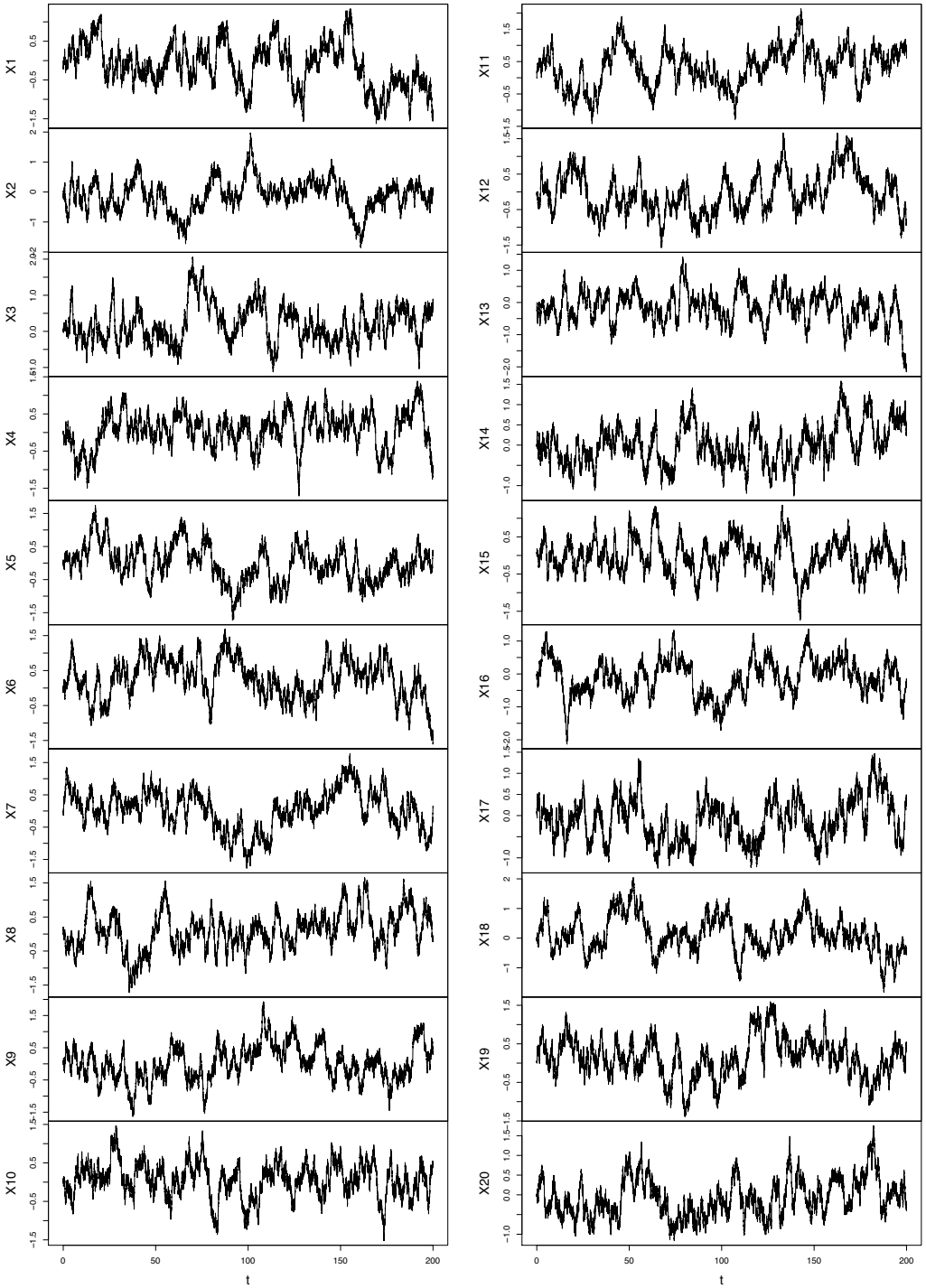

Figure 4.1: Sample path of covariate processes X

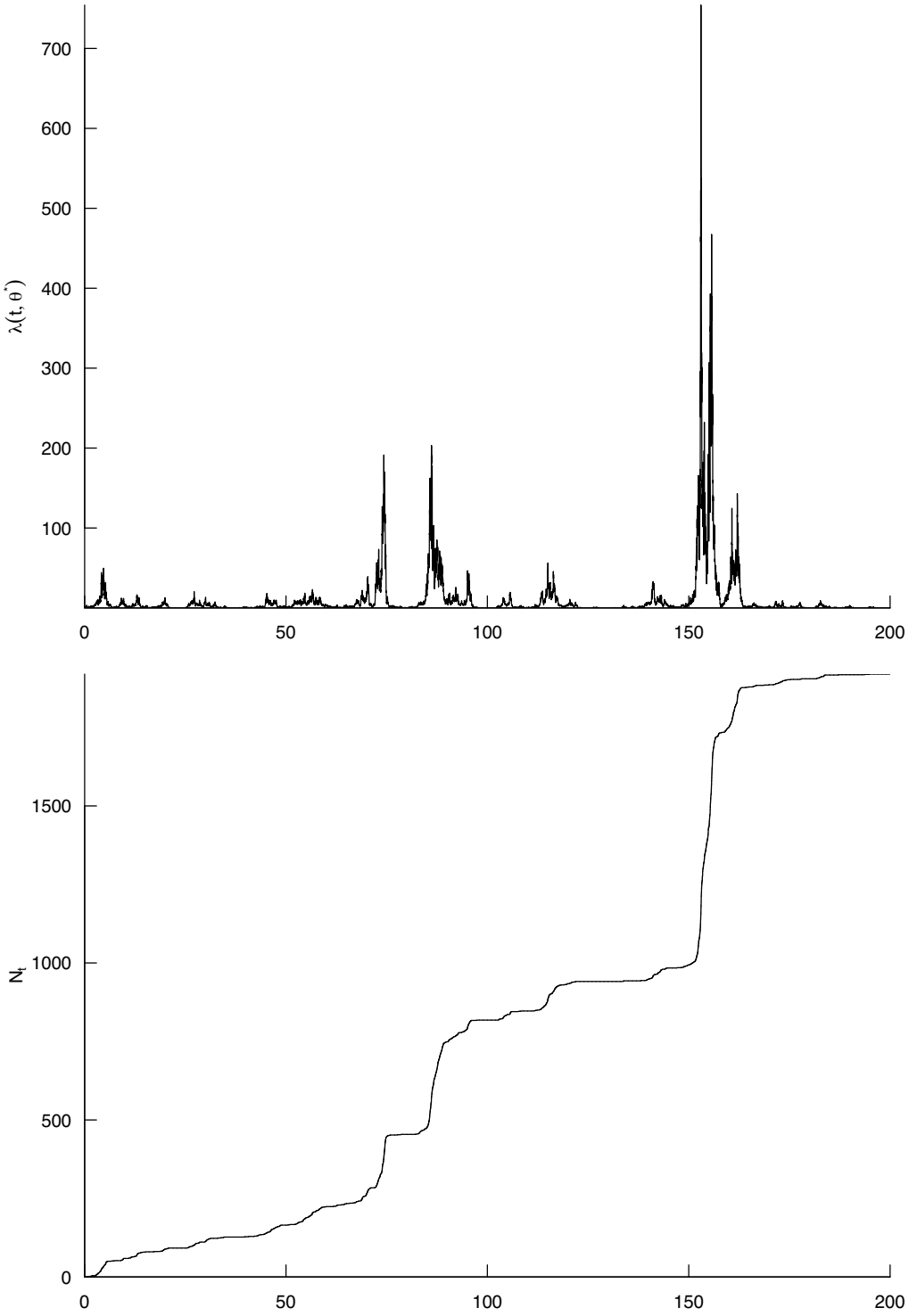

Figure 4.2: Sample paths of intensity $\lambda(t)$ and counting processes N (global)

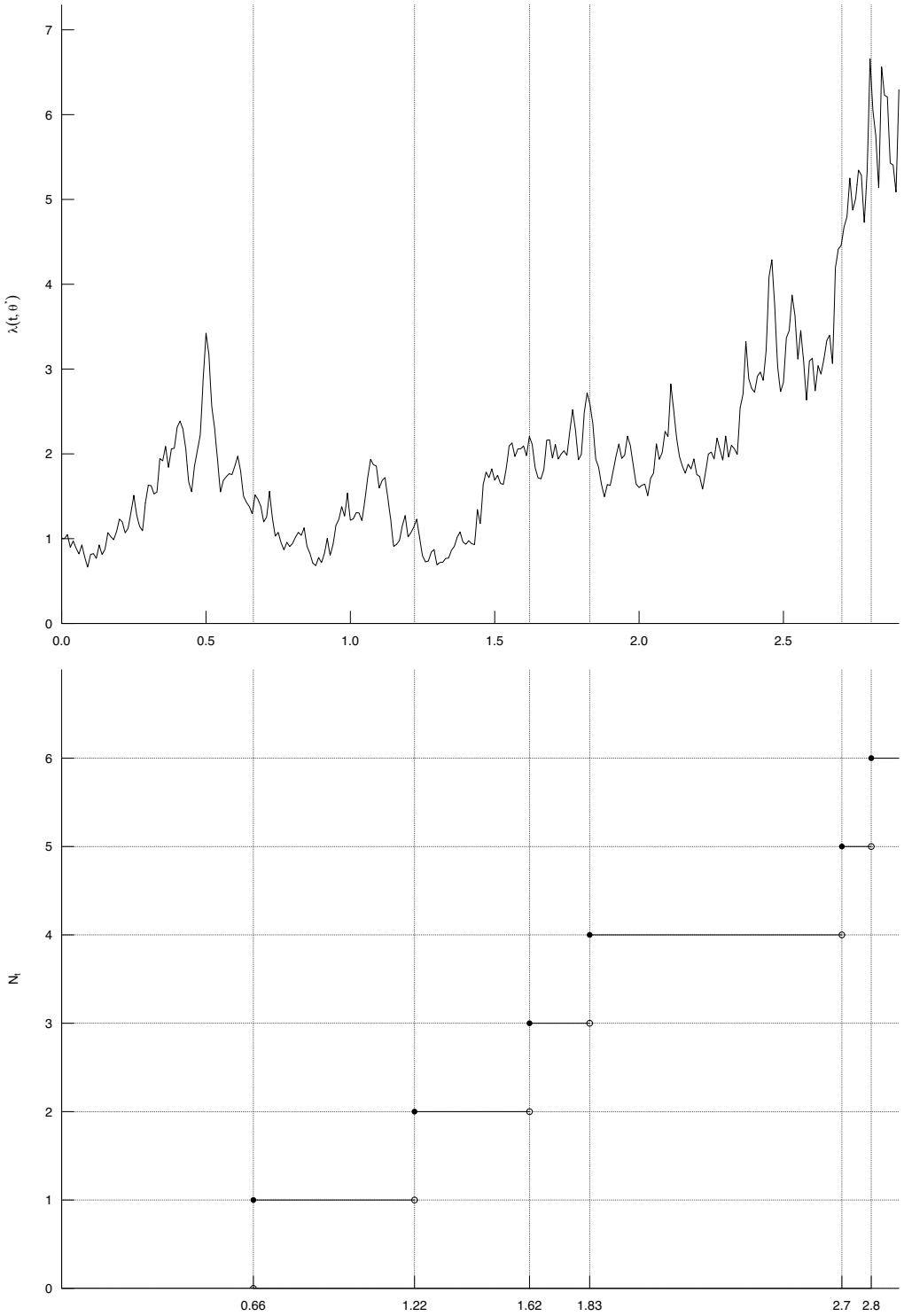

Figure 4.3: Sample paths of intensity $\lambda(t)$ and counting processes N (local)

Table 4.1. Results of the variable selection under T=50,100,200,400.

	(γ, r, q)	$T = 50$	$T = 100$	$T = 200$	$T = 400$
%(p-LSA)	(1, 1.2, 0.3)	32.1	70.4	96.8	99.9
%(unified LASSO)	(1, 1.2, 1)	5.9	17.4	47.1	79.6
%(Bridge type)	(0, 1, 0.3)	8.9	21.9	52.3	80.3

Table 4.2. The summary of results for the simulation under $T = 200$.

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
true	2	-1	1	-0.5	-1.5	1.5	0.5	0.75	0	0
initial	1.9938	-0.9936	0.9941	-0.5003	-1.4909	1.4978	0.5009	0.7490	0.0001	-0.0002
	(0.0722)	(0.0830)	(0.0858)	(0.0889)	(0.0962)	(0.0745)	(0.0851)	(0.0880)	(0.0908)	(0.0974)
p-LSA	1.9918	-0.9872	0.9877	-0.4758	-1.4877	1.4946	0.4750	0.7383	-0.0001	-0.0008
	(0.0728)	(0.0840)	(0.0868)	(0.1035)	(0.0965)	(0.0748)	(0.1049)	(0.0904)	(0.0265)	(0.0340)
P-O	1.9959	-0.9971	0.9964	-0.4979	-1.4931	1.4998	0.4946	0.7523	-0.0003	-0.0007
	(0.0565)	(0.0682)	(0.0713)	(0.0892)	(0.0823)	(0.0602)	(0.0915)	(0.0730)	(0.0228)	(0.0307)
%(p-LSA)	100.0	100.0	100.0	98.8	100.0	100.0	98.4	100.0	99.4	99.2
	θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}	θ_{18}	θ_{19}	θ_{20}
true	0	0	0	0	0	0	0	0	0	0
initial	0.0014	-0.0038	0.0007	0.0026	0.0063	-0.0015	0.0042	0.0013	-0.0048	-0.0014
	(0.0760)	(0.0794)	(0.0872)	(0.0947)	(0.0971)	(0.0701)	(0.0802)	(0.0896)	(0.0914)	(0.0941)
p-LSA	0.0000	0.0003	0.0007	0.0002	-0.0004	0.0000	0.0011	-0.0003	0.0013	0.0004
	(0.0000)	(0.0185)	(0.0252)	(0.0240)	(0.0248)	(0.0000)	(0.0194)	(00239)	(0.0247)	(0.0208)
P-O	0.0000	-0.0003	0.0007	0.0000	-0.0003	0.0000	0.0007	-0.0004	0.0010	0.0004
	(0.0000)	(0.0152)	(0.0221)	(0.0188)	(0.0259)	(0.0000)	(0.0134)	(0.0214)	(0.0199)	(0.0213)
%(p-LSA)	100.0	99.7	99.4	99.4	99.5	100.0	99.7	99.5	99.7	99.5

4.2 Simulations for the Hawkes process

We consider the Hawkes model in Section 3.1.2. Let $\mathsf{d} = 3$ and the parameter space

$$\Theta = \{(\nu, C, A) \in \mathbb{R}_+ \times \mathbb{R}_+^{3\times 3} \times \mathbb{R}_+^{3\times 3}; \rho(\Phi) < 1, \nu_\alpha > 0, a_{\alpha\beta} > 0, \forall(\alpha, \beta)\}.$$

For $\alpha \in \mathbf{I}$ we generated the data $(t_i^\alpha)_{i=1,2,\dots}$ which is jump point of the counting process $(N_t^\alpha)_t$ with the intensity

$$\lambda^\alpha(t) = \nu_\alpha + \sum_{\beta\in\mathbf{I}} \int_0^{t-} c_{\alpha\beta} e^{-a_{\alpha\beta}(t-s)} dN_s^\beta,$$

where $\theta^* = (\nu^*, C^*, A^*)$ is the true value of the parameter;

$$\nu^* = \begin{bmatrix} 1 \\ 0.5 \\ 1 \end{bmatrix}, \quad C^* = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0.5 & 1 \end{bmatrix}, \quad A^* = \begin{bmatrix} 2 & * & * \\ 2 & 3 & * \\ * & 3 & 4 \end{bmatrix}.$$

In this simulation, we take $\xi = 1$ (in subsection 3.1.2), i.e.,

$$A^* = \begin{bmatrix} 2 & 1 & 1 \\ 2 & 3 & 1 \\ 1 & 3 & 4 \end{bmatrix}.$$

Then the matrix Φ^* is

$$\Phi^* = \begin{bmatrix} 1/2 & 0 & 0 \\ 1/4 & 1/3 & 0 \\ 0 & 1/6 & 1/4 \end{bmatrix}.$$

Figure 4.4 shows the graph structure of this model.

Let $\mathcal{L}_T(\theta) = -\ell_T(\theta)$ in (3.1) and we use QMLE for the initial estimator $\tilde{\theta} = (\tilde{\nu}, \tilde{C}, \tilde{A})$. Then the objective function is denoted by

$$Q_T^{(q)}(C) = \hat{G}[(C - \tilde{C})^{\otimes 2}] + \sum_{\alpha,\beta\in\mathbf{I}} \kappa_T^{\alpha\beta} |c_{\alpha\beta}|^q,$$

where $\kappa_T^{\alpha\beta} = \alpha_T(|c_{\alpha\beta}| + 1/T)^{-\gamma}, \alpha_T = T^{-r/2}, 1 < r < 2 - q + \gamma$ and $\hat{G} = \partial_C^2 \ell_T(\theta)$. In this simulation, we take a tuning parameter $(\gamma, r, q) = (3, 1.2, 1)$.

Tables 4.3 and 4.4 show the means and standard deviations (parentheses) for the QMLE (with $T = 100$, $T = 200$, $T = 400$ and $T = 600$) and Table 4.5 shows the results of the variable selection for the QMLE and the penalized LSA estimator.

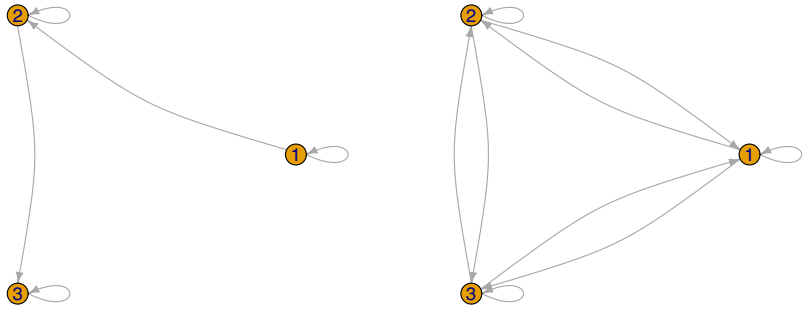

Figure 4.4: Graph structure; true model (left) and full model (right)

Here, we refer to the algorithm for generating the sample path of Hawkes process. Ogata [31] gave a proposition that states the simulation of multivariate point processes, by distributing accepted points to each dimension with probabilities proportional to their intensities.

Lemma 4.1 (Ogata [31])**.** Let $N_t = (N_t^1, \ldots, N_t^d)$ be a d-variate point process on an interval $[0, T]$ with stochastic intensities $\lambda^\alpha(t) = \lambda^\alpha(t|\mathcal{F}_{t-}^N)$ for $\alpha \in \mathbf{I}$. Suppose there is a one-dimensional $\mathcal{F}^N$-predictable process $\bar{\lambda}(t)$ which is defined path-wisely satisfying

$$\sum_{\alpha \in \mathbf{I}} \lambda^\alpha(t) \leq \bar{\lambda}(t), \quad 0 < t \leq T$$

and set

$$\lambda^0(t) = \bar{\lambda}(t) - \sum_{\alpha \in \mathbf{I}} \lambda^\alpha(t).$$

Let $\bar{t}_1, \ldots, \bar{t}_{\bar{N}_T} \in (0, T]$ be the points of the process $\bar{N}_t$ with stochastic intensity $\bar{\lambda}(t)$. For each of the points $\bar{t}_k, k = 1, \ldots, \bar{N}_T$, attach a mark $\alpha = 0, 1, \ldots, d$ with probability $\lambda^\alpha(\bar{t}_k)/\bar{\lambda}(\bar{t}_k)$, respectively. Then the points with marks $\alpha \in \mathbf{I}$, provide a d-variate point process with stochastic intensities $\lambda^\alpha(t)$.

The simulation of a multivariate Hawkes process with exponential decays on a fixed interval is similar to the univariate case, with only one extra step

that decides which dimension an accepted point belongs to. By Lemma 4.1, given that a point is accepted at time s, it should be distributed to dimension α with probability $\lambda^\alpha(s)/\sum_{\alpha\in\mathbf{I}}\lambda^\alpha(s)$, for $\alpha \in \mathbf{I}$. The procedure for simulating a d-variate Hawkes process on the interval $[0, T]$ is summarized below, where $\mathcal{T}^\alpha$ represents the ordered set of accepted points in dimension α and n^α counts the number of points $\mathcal{T}^\alpha$, for $\alpha \in \mathbf{I}$. As before, s is always the newest candidate point generated.

(1) Set $\mathcal{T}^1 = \cdots = \mathcal{T}^d = \varnothing, s = 0$ and $n^1 = \cdots = n^d = 0$.

(2) Repeat the following until $s > T$.

- **(i)** Set $\bar{\lambda} = \sum_{\alpha\in\mathbf{I}}\lambda^\alpha(s) = \sum_{\alpha\in\mathbf{I}}\left(\nu_\alpha + \sum_{\beta\in\mathbf{I}}\sum_{\tau\in\mathcal{T}^\beta} c_{\alpha\beta}e^{-a_{\alpha\beta}(s-\tau)}\right)$.
- **(ii)** Generate u from a uniform distribution on $[0, 1]$.
- **(iii)** Generate $w = -\log u/\bar{\lambda}$ as the interarrival to the next candidate point.
- **(iv)** Set the new candidate point $s = s + w$.
- **(v)** Generate D from a uniform distribution on $[0, 1]$.
  - **(a)** If $D \le \sum_{\alpha\in\mathbf{I}}\lambda^\alpha(s)/\bar{\lambda}$, then do the following:
    - **(a1)** Find $\beta \in \mathbf{I}$ such that $\sum_{\alpha=1}^{\beta-1}\lambda^\alpha(s) < D\bar{\lambda} \le \sum_{\alpha=1}^{\beta}\lambda^\alpha(s)$.
    - **(a2)** Assign candidate point s to dimension β by setting $n^\beta = n^\beta + 1, t^\beta_{n^\beta} = s$ and $\mathcal{T}^\beta = \mathcal{T}^\beta \bigcup \{t^\beta_{n^\beta}\}$.
  - **(b)** else, do nothing.

(3) If $t^\beta_{n^\beta} > T$, then $\mathcal{T}^1, \ldots, \mathcal{T}^\beta - \{t^\beta_{n^\beta}\}, \ldots, \mathcal{T}^d$ contain the simulated points for each dimension;

(4) else $\mathcal{T}^\alpha, \alpha \in \mathbf{I}$ contain the simulated points.

Figures 4.5-4.7 show an example sample path of intensity $\lambda^1(t), \lambda^2(t)$ and $\lambda^3(t)$ and associated counting process $N = (N^\alpha)_{\alpha\in\mathbf{I}}$.

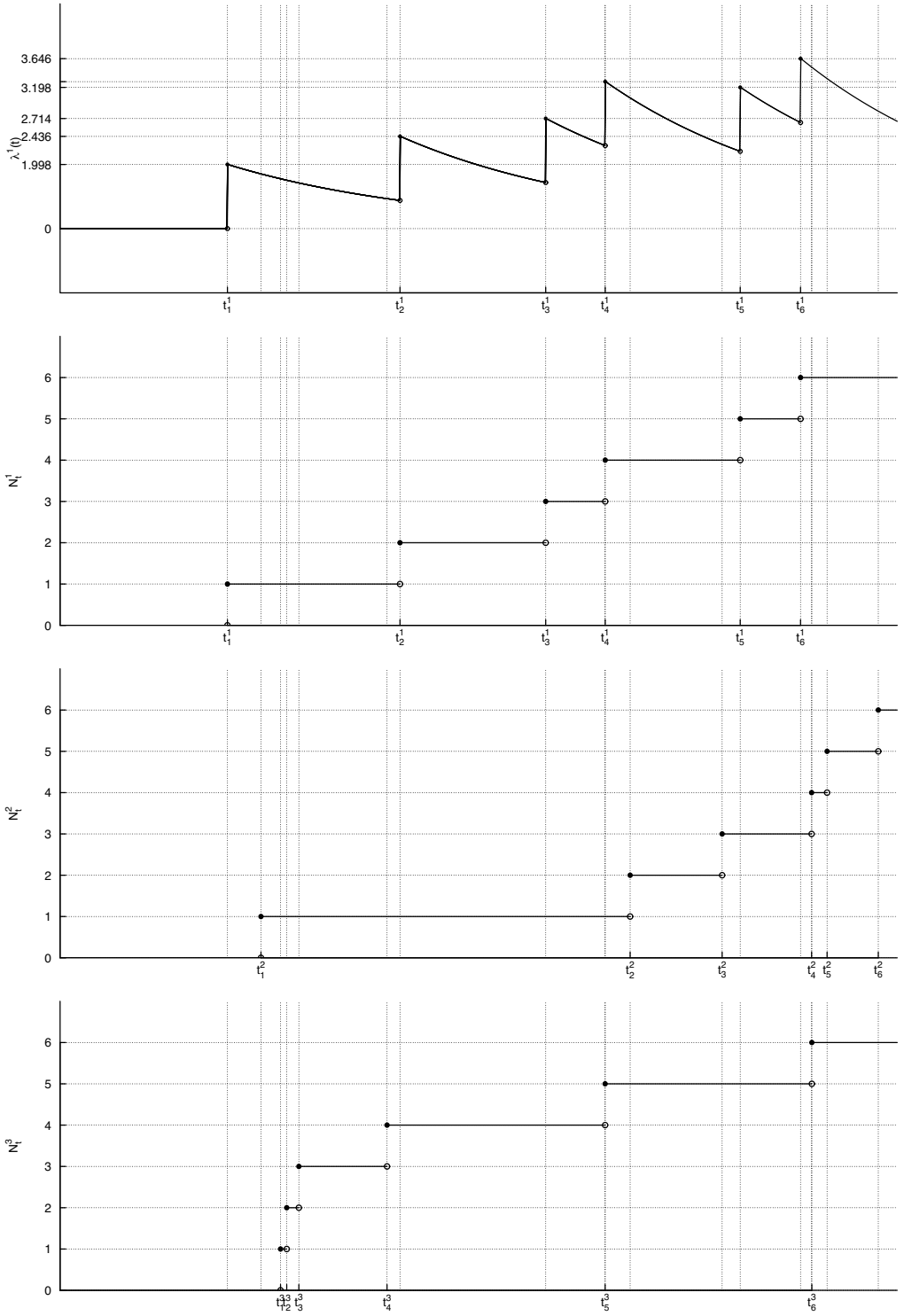

Figure 4.5: Sample paths of intensity $\lambda^1(t)$ and counting processes N

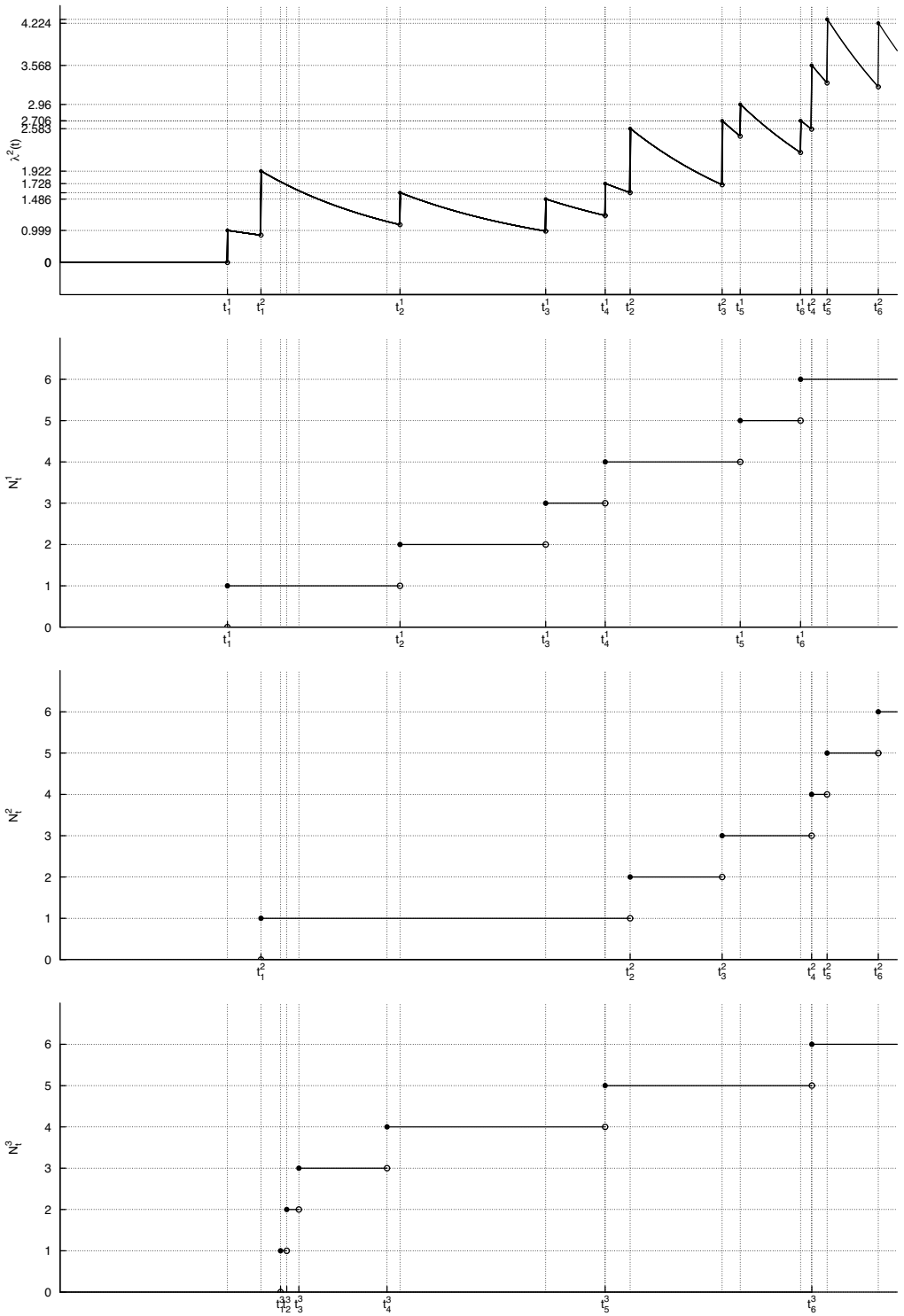

Figure 4.6: Sample paths of intensity $\lambda^2(t)$ and counting processes N

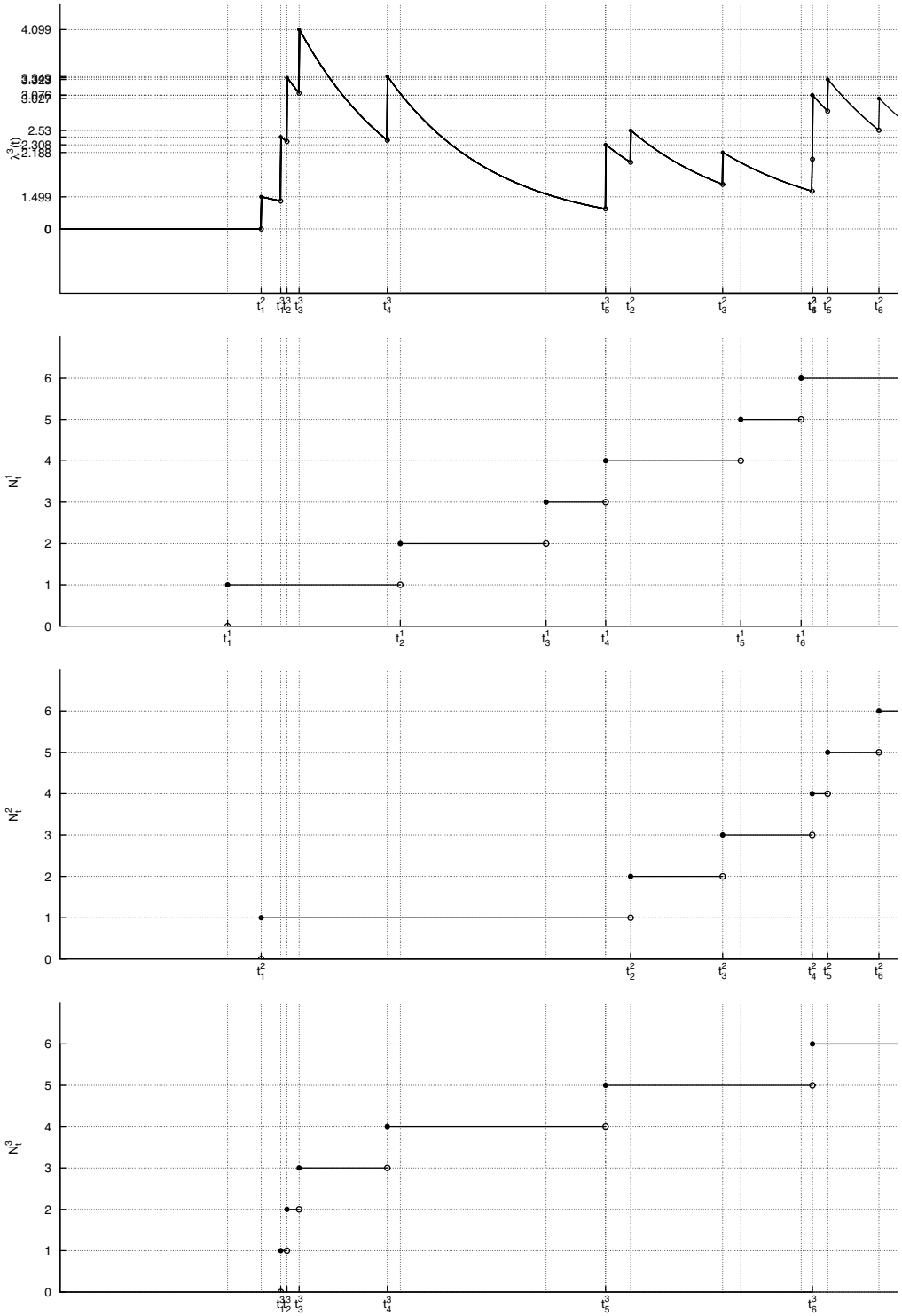

Figure 4.7: Sample paths of intensity $\lambda^3(t)$ and counting processes N

Table 4.3. normal methods (ν, QMLE) ; $T = 100, 200, 400, 600$.

	ν_1	ν_2	ν_3
true	1	0.5	1
$T = 100$	0.9404	0.4581	0.9375
	(0.2638)	(0.2098)	(0.2438)
$T = 200$	0.9622	0.4806	0.9828
	(0.1748)	(0.1547)	(0.1710)
$T = 400$	0.9785	0.5008	1.0037
	(0.1252)	(0.1013)	(0.1132)
$T = 600$	0.9837	0.5074	1.0122
	(0.1040)	(0.0808)	(0.0901)

Table 4.4. normal methods (C and A, QMLE) ; $T = 100, 200, 400, 600$.

	C_{11}	C_{12}	C_{13}	C_{21}	C_{22}	C_{23}	C_{31}	C_{32}	C_{33}
true	1	0	0	0.5	1	0	0	0.5	1
$T = 100$	0.9869	0.1437	0.1174	0.5395	1.0300	0.1003	0.0995	0.5598	1.0266
	(0.3206)	(0.3375)	(0.2509)	(0.3024)	(0.4968)	(0.2224)	(0.2369)	(0.4454)	(0.5151)
$T = 200$	0.9737	0.0871	0.0625	0.4984	0.9841	0.0522	0.0558	0.5248	0.9848
	(0.2131)	(0.1781)	(0.1480)	(0.1783)	(0.3004)	(0.1126)	(0.1406)	(0.3173)	(0.3592)
$T = 400$	0.9736	0.0467	0.0414	0.4760	0.9609	0.0294	0.0279	0.4976	0.9515
	(0.1493)	(0.0962)	(0.0862)	(0.1136)	(0.2003)	(0.0723)	(0.0711)	(0.2069)	(0.2300)
$T = 600$	0.9721	0.0305	0.0300	0.4725	0.9557	0.0194	0.0202	0.4934	0.9479
	(0.1170)	(0.0670)	(0.0644)	(0.0954)	(0.1628)	(0.0526)	(0.0507)	(0.1739)	(0.1802)
	A_{11}	A_{12}	A_{13}	A_{21}	A_{22}	A_{23}	A_{31}	A_{32}	A_{33}
true	2	***	***	2	3	***	***	3	4
$T = 100$	2.3240	3.4120	3.2599	2.3013	3.9122	3.3814	3.4714	4.0366	4.9993
	(1.2544)	(3.7475)	(3.8204)	(2.5430)	(3.0128)	(3.4455)	(4.0253)	(4.3408)	(3.659)
$T = 200$	2.0683	3.1594	2.8740	2.0451	3.2604	2.9903	3.1011	3.7092	4.3864
	(0.5400)	(2.9022)	(2.3451)	(1.1171)	(1.3562)	(1.9802)	(2.5508)	(3.0898)	(2.5230)
$T = 400$	2.0024	2.8727	2.8210	1.9237	3.0091	2.8593	2.8526	3.3250	3.9466
	(0.3594)	(1.8543)	(1.7061)	(0.5999)	(0.8115)	(1.7341)	(1.7710)	(1.9000)	(1.3386)
$T = 600$	1.9713	2.7551	2.6996	1.8992	2.9361	2.7922	2.8203	3.2191	3.8592
	(0.2776)	(1.6391)	(1.4824)	(0.4821)	(0.6134)	(1.5474)	(1.4925)	(1.5721)	(0.9821)

Table 4.5. The number of times, in percentage over 1000 iterations, that the estimator chooses the true model (coordinatewise); $T = 100, 200, 400, 600$.

		Φ_{11}	Φ_{12}	Φ_{13}	Φ_{21}	Φ_{22}	Φ_{23}	Φ_{31}	Φ_{32}	Φ_{33}	all
true	T	active	0	0	active	active	0	0	active	active	
initial	100	100.0	54.4	54.6	100.0	99.9	55.9	55.9	97.8	99.8	9.4
	200	100.0	56.1	59.9	100.0	100.0	60.1	60.4	99.6	100.0	11.9
	400	100.0	58.7	59.7	100.0	100.0	64.2	64.2	100.0	100.0	12.9
	600	100.0	63.8	61.1	100.0	100.0	66.8	68.3	100.0	100.0	17.3
pLSA	100	99.7	78.6	81.3	93.0	99.1	82.6	81.7	80.1	97.0	31.0
	200	100.0	84.8	88.5	97.8	99.9	90.1	91.2	86.1	99.4	49.6
	400	100.0	90.8	91.3	99.7	100.0	94.1	95.3	93.4	100.0	69.5
	600	100.0	94.0	93.6	99.8	100.0	96.1	95.4	96.7	100.0	77.8

4.3 Simulation for the diffusion type process

We consider the model (3.9) in Section 3.2.2. Let $\mathsf{p} = 10$, i.e., the parameter space Θ is $[-10, 10]^{10}$. The process Y is defined by

$$Y_t = \int_0^t \sigma(X_s, \theta) dW_s, \quad t \in [0, 1],$$

where X is a 10-dimensional OU process satisfying the following stochastic differential equation

$$dX_t = -0.2X_t dt + 0.5 I_{10} dw_t, \quad X_0 = 0, \quad t \in [0, 1]$$

and $\sigma(x, \theta) = \exp(\sum_{j=1}^{10} \theta_j x_j) \wedge M_0, M_0 = 10^5$. Here w is a 10-dimensional standard Wiener process independent of W. We generated the data $(Y_{t_i}, X_{t_i})_{i=0,1,\ldots,n}, t_i = \frac{i}{n}$ with

$$\theta^* = [1, 1, -1, -1, 0.5, 0, 0, 0, 0, 0]'.$$

Let $\mathcal{L}_n(\theta) = \mathbb{L}_n(\theta) = -\mathbb{H}_n(\theta)$ in (3.10). We used the QMLE for the initial estimator $\tilde{\theta}$.

In order to apply our methods to this model, we use Theorem 6 in ([38]). By the definition of $\sigma(x, \theta)$, [A1] holds. [B2] is satisfied if we choose the stopping time $\tau \equiv 0$. Now we need to check [A3′]. Since $\mathrm{supp}\mathcal{L}\{X_0\} = \{0\}$, we can take $0 \in U \subset \{x \in \mathbb{R}^{10}; \sigma(x, \theta) < M, \forall\theta\}$. If we define $f(x, \theta) = (m_0 \sum_j (\theta_j - \theta_j^*) x_j)/|\theta - \theta^*|$ for sufficiently small m_0 when $\theta \neq 0$ and $f(x, 0) = \epsilon_0$ for some positive number ϵ_0, then (i) is satisfied for $\varrho = 2$. Next we take a covering $\{\Theta_k\}_{k=1,\ldots,11}$ such that $\Theta_k = \{\theta \in \Theta - \{\theta^*\}; |\theta_k - \theta_k^*| \geq |\theta_j - \theta_j^*|, \forall j\}$ for $k = 1, \ldots, 10$ and $\Theta_{11} = \{\theta^*\} \subset \Theta$. For $\Theta_k (k = 1, \ldots, 10)$ if we take $\xi_0 = e_k$ and $\Psi(P_{\xi_0}^{\perp} x, \theta) = (\sum_{j \neq k} (\theta_j - \theta_j^*) x_j)/(\theta_k - \theta_k^*)$, then $|f(x, \theta)| \geq \frac{1}{\sqrt{10}} (\xi_0 \cdot x + \Psi(P_{\xi_0}^{\perp} x, \theta))$, and (ii) holds.

Then the objective function is denoted by

$$Q_n^{(q)}(\theta) = (\theta - \tilde{\theta})'(\theta - \tilde{\theta}) + \sum_{j=1}^{10} \kappa_n^j |\theta_j|^q$$

where $\kappa_n^j = \alpha_n |\tilde{\theta}_j|^{-\gamma}$, $\alpha_n = (\frac{1}{\sqrt{n}})^r, 1 < r < 2 - q - \gamma$. We considered the cases where $n = 2500, 50000, 10000, 20000$ and the triplet of tuning parameters $(\gamma, r, q) = (3.2, 1.2, 0.3)$. In the same way as Table 1, Table 3 compares the results of the variable selection of the penalized LSA estimator, the unified LASSO type estimator and the Bridge type estimator.

Table 4 compares the means and the standard deviations (parentheses) for the three estimators (initial estimator, penalized LSA estimator and P-O estimator) in the case $n = 10000$.

Table 4.6. Results for the variable selection under n=2500,5000,10000,20000.

	(γ, r, q)	$n = 2500$	$n = 5000$	$n = 10000$	$n = 20000$
% (p-LSA)	(3.2, 1.2, 0.3)	64.8	86.3	97.8	99.9
% (unified LASSO)	(3.2, 1.2, 1)	54.0	77.2	93.8	98.6
% (Bridge type)	(0, 1, 0.3)	0.7	0.9	1.0	1.7

Table 4.7. The summary of results for the simulation under $n = 10000$.

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
true	1	1	-1	-1	0.5	0	0	0	0	0
initial	1.0020	0.9952	-0.9981	-1.0017	0.5021	-0.0040	-0.0006	-0.0027	0.0046	0.0019
	(0.0879)	(0.0896)	(0.0835)	(0.0847)	(0.0842)	(0.0885)	(0.0896)	(0.0849)	(0.0839)	(0.0896)
p-LSA	1.0013	0.9945	-0.9974	-1.0010	0.4856	-0.0003	0.0000	0.0000	0.0000	0.0003
	(0.0881)	(0.0899)	(0.0838)	(0.0849)	(0.1097)	(0.0107)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
P-O	1.0017	0.9980	-0.9960	-0.9978	0.4936	-0.0002	0.0000	0.0000	0.0000	0.0002
	(0.0663)	(0.0694)	(0.0731)	(0.0676)	(0.0916)	(0.0058)	(0.0000)	(0.0000)	(0.0000)	(0.0049)
%(p-LSA)	100.0	100.0	100.0	100.0	98.0	99.9	100.0	100.0	100.0	99.9

Bibliography

[1] Akritas, M.G., Johnson, R.A.: Asymptotic inference in lévy processes of the discontinuous type. The Annals of Statistics pp. 604–614 (1981)

[2] Athreya, K.B., Keiding, N.: Estimation theory for continuous-time branching processes. Sankhyā: The Indian Journal of Statistics, Series A pp. 101–123 (1977)

[3] Bacry, E., Delattre, S., Hoffmann, M., Muzy, J.F.: Modelling microstructure noise with mutually exciting point processes. Quantitative Finance **13**(1), 65–77 (2013)

[4] Basawa, I.V., Scott, D.J.: Asymptotic optimal inference for non-ergodic models, vol. 17. Springer Science & Business Media (2012)

[5] Bickel, P.J., Ritov, Y., Tsybakov, A.B., et al.: Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics **37**(4), 1705–1732 (2009)

[6] Billingsley, P.: Statistical inference for Markov processes, vol. 2. University of Chicago Press (1961)

[7] Blundell, C., Beck, J., Heller, K.A.: Modelling reciprocating relationships with hawkes processes. In: Advances in Neural Information Processing Systems, pp. 2600–2608 (2012)

[8] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning **3**(1), 1–122 (2011)

[9] Bradic, J., Fan, J., Jiang, J.: Regularization for Cox ' s proportional hazards model with np-dimensionality. Annals of Statistics **39**(6), 3092 (2011)

[10] Brillinger, D.R.: Statistical inference for stationary point processes. In: Selected Works of David Brillinger, pp. 499–543. Springer (2012)

[11] Brown, M.: Statistical analysis of non-homogeneous poisson processes. Stochastic point processes pp. 67–89 (1972)

[12] Clinet, S., Yoshida, N.: Statistical inference for ergodic point processes and application to limit order book. Stochastic Processes and their Applications **127**(6), 1800–1839 (2017)

[13] Crane, R., Sornette, D.: Robust dynamic classes revealed by measuring the response function of a social system. Proceedings of the National Academy of Sciences **105**(41), 15,649–15,653 (2008)

[14] De Gregorio, A., Iacus, S.M.: Adaptive lasso-type estimation for multivariate diffusion processes. Econometric Theory **28**(4), 838–860 (2012)

[15] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al.: Least angle regression. The Annals of Statistics **32**(2), 407–499 (2004)

[16] Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association **96**(456), 1348–1360 (2001)

[17] Feigin, P.D.: A note on maximum likelihood estimation for simple branching processes. Australian Journal of Statistics **19**(2), 152–154 (1977)

[18] Frank, L.E., Friedman, J.H.: A statistical view of some chemometrics regression tools. Technometrics **35**(2), 109–135 (1993)

[19] Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)

[20] Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software **33**(1), 1 (2010)

[21] Hawkes, A.G.: Point spectra of some mutually exciting point processes. Journal of the Royal Statistical Society: Series B (Methodological) **33**(3), 438–443 (1971)

[22] Hawkes, A.G.: Spectra of some self-exciting and mutually exciting point processes. Biometrika **58**(1), 83–90 (1971)

[23] Huang, J., Sun, T., Ying, Z., Yu, Y., Zhang, C.H.: Oracle inequalities for the lasso in the cox model. Annals of Statistics **41**(3), 1142 (2013)

[24] Jeganathan, P.: On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal. Sankhyā Ser. A **44**(2), 173–212 (1982)

[25] Kamatani, K., Uchida, M.: Hybrid multi-step estimators for stochastic differential equations based on sampled data. Statistical Inference for Stochastic Processes **18**(2), 177–204 (2015)

[26] Knight, K., Fu, W.: Asymptotics for lasso-type estimators. Annals of Statistics pp. 1356–1378 (2000)

[27] Luschgy, H.: Asymptotic inference for semimartingale models with singular parameter points. Journal of Statistical Planning and Inference **39**(2), 155–186 (1994)

[28] Masuda, H., Shimizu, Y.: Moment convergence in regularized estimation under multiple and mixed-rates asymptotics. Mathematical Methods of Statistics **26**(2), 81–110 (2017)

[29] Meinshausen, N., Bühlmann, P., et al.: High-dimensional graphs and variable selection with the lasso. The Annals of Statistics **34**(3), 1436–1462 (2006)

[30] Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P., Tita, G.E.: Self-exciting point process modeling of crime. Journal of the American Statistical Association **106**(493), 100–108 (2011)

[31] Ogata, Y.: On lewis' simulation method for point processes. IEEE Transactions on Information Theory **27**(1), 23–31 (1981)

[32] Ogata, Y.: Seismicity analysis through point-process modeling: A review. In: Seismicity Patterns, their Statistical Significance and Physical Meaning, pp. 471–507. Springer (1999)

[33] Park, M.Y., Hastie, T.: L1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **69**(4), 659–677 (2007)

[34] Suzuki, T., Yoshida, N.: Penalized least squares approximation methods and their applications to stochastic processes. Japanese Journal of Statistics and Data Science pp. 1–29 (2020)

[35] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996)

[36] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(1), 91–108 (2005)

[37] Uchida, M., Yoshida, N.: Adaptive estimation of an ergodic diffusion process based on sampled data. Stochastic Processes and their Applications **122**(8), 2885–2924 (2012)

[38] Uchida, M., Yoshida, N.: Quasi likelihood analysis of volatility and nondegeneracy of statistical random field. Stochastic Processes and their Applications **123**(7), 2851–2876 (2013)

[39] Uchida, M., Yoshida, N.: Adaptive bayes type estimators of ergodic diffusion processes from discrete observations. Statistical Inference for Stochastic Processes **17**(2), 181–219 (2014)

[40] Vere-Jones, D.: Stochastic models for earthquake occurrence. Journal of the Royal Statistical Society: Series B (Methodological) **32**(1), 1–45 (1970)

[41] Vere-Jones, D., Ozaki, T.: Some examples of statistical estimation applied to earthquake data. Annals of the Institute of Statistical Mathematics **34**(1), 189–207 (1982)

[42] Wainwright, M.J.: Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). IEEE Transactions on Information Theory **55**(5), 2183–2202 (2009)

[43] Wang, H., Leng, C.: Unified lasso estimation by least squares approximation. Journal of the American Statistical Association **102**(479), 1039–1048 (2007)

[44] Witten, D.M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10**(3), 515–534 (2009)

[45] Yang, S.H., Zha, H.: Mixture of mutually exciting processes for viral diffusion. In: International Conference on Machine Learning, pp. 1–9 (2013)

[46] Yoshida, N.: Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. Annals of the Institute of Statistical Mathematics **63**(3), 431–479 (2011)

[47] Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**(1), 49–67 (2006)

[48] Yuan, M., Lin, Y.: Model selection and estimation in the gaussian graphical model. Biometrika **94**(1), 19–35 (2007)

[49] Zhang, C.H., et al.: Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics **38**(2), 894–942 (2010)

[50] Zhao, P., Yu, B.: On model selection consistency of lasso. Journal of Machine Learning Research **7**(Nov), 2541–2563 (2006)

[51] Zhou, K., Zha, H., Song, L.: Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In: Artificial Intelligence and Statistics, pp. 641–649 (2013)

[52] Zou, H.: The adaptive lasso and its oracle properties. Journal of the American Statistical Association **101**(476), 1418–1429 (2006)

[53] Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. Journal of Computational and Graphical Statistics **15**(2), 265–286 (2006)

appendix

A1. Polynomial type large deviation for sparse estimation

Polynomial type large deviation inequality

Let Θ be a bounded open set in $\mathbb{R}^{\mathsf{d}}$. The closure of Θ in $\mathbb{R}^{\mathsf{d}}$ is denoted by $\bar{\Theta}$. Our interest will be on inference for the parameter θ in Θ. Given a probability space $(\Omega, \mathcal{F}, P)$, we consider a sequence of random fields $\mathbb{H}_T : \Omega \times \bar{\Theta} \to \mathbb{R}$, where $\mathbb{T}$ is a subset in $(0, \infty)$ such that $\sup \mathbb{T} = \infty$; e.g. $\mathbb{T}$ is $\mathbb{R}_+ = [0, \infty), \mathbb{Z}_+ = \{0, 1, 2, \ldots\}, \mathbb{N} = \{1, 2, \ldots\}$ etc. We suppose that the mapping $\bar{\Theta} \ni \theta \mapsto \mathbb{H}_T(\omega, \theta) \in \mathbb{R}$ is continuous for all $\omega \in \Omega$.

Let $\theta^* \in \bar{\Theta}$. Let $a_T \in GL(\mathsf{p}, \mathbb{R})$ $(T \in \mathbb{T})$ such that $\lim_{T\to\infty} |a_T| = 0$, where $|a_T| = \{\mathrm{Tr}(a_T a_T')\}^{1/2}$. Let $\mathbb{U}_T = \{u \in \mathbb{R}^{\mathsf{d}}; \theta^* + a_T u \in \bar{\Theta}\}$. Define a statistical random field $\mathbb{Z}_T : \Omega \times \mathbb{U}_T \to (0, \infty)$ by

$$\mathbb{Z}_T(u) = \exp\Big\{\mathbb{H}_T(\theta_T^\dagger(u) - \mathbb{H}_T(\theta^*)\Big\} \quad (u \in \mathbb{U}_T),$$

where $\theta_T^\dagger(u) = \theta^* + a_T u$.

Some notation will be necessary. We write

$$T[u_1, \ldots, u_k] = T[u_1 \otimes \cdots \otimes u_k] = \sum_{i_1,\ldots,i_k} T_{i_1,\ldots,i_k} u_1^{i_1} \cdots u_k^{i_k}$$

for a tensor-valued tensor $T = (T_{i_1,\ldots,i_k})_{i_1,\ldots,i_k}$ and $u_1 = (u_1^{i_k})_{i_k}$. The brackets $[\ ]$ will often be used for multilinear mappings. We simply denote the r times product of u by $u^{\otimes r} = u \otimes \cdots \otimes u$.

It is useful to consider the situation where $\mathbb{Z}$ is locally asymptotically quadratic (LAQ). Let Δ_T be a d-dimensional random variable and let Γ be a $\mathsf{d} \times \mathsf{d}$ nonnegative-definite symmetric random matrix. We defined the random

field $r_T : \Omega \times \mathbb{U}_T \to \mathbb{R}$ by

$$\mathbb{Z}_T(u) = \exp\left(\Delta_T[u] - \frac{1}{2}\Gamma[u^{\otimes 2}] + r_T(u) \right) \quad (u \in \mathbb{U}_T).$$

It should be remarked that $0 \in \mathbb{R}^{\mathsf{d}}$ is in $\mathbb{U}_T$ but any d-dimensional ball centered at 0 may not included in $\mathbb{U}_T$ even for large T. The restriction on $\mathbb{U}_T$ is common if u is banned from moving in all directions. Even in such a case, the LAN property makes sense if the random field $\mathbb{H}_T$ is smooth in θ in restricted direction, as we will later see in application.

Though θ^* may be in Θ, we are interested in the case where $\theta^* \in \partial\Theta$, since the former case was treated in Yoshida [46]. The aim of this note is to remark that the same polynomial type large deviation as [46] holds, and to illustrate an application in the context of sparse estimation. In what follows, for simplicity, we will treat a single probability measure P, while a uniform estimate in ξ is considered in [46]. Besides, we will not treat the random field with a nuisance parameter τ in [46].

Let $b_T = \lambda_{min}(a_T' a_T)^{-1}$, where λ_{min} denotes the minimum eigenvalue of a real symmetric matrix. Since $|a_T| \to 0, 0 < b_T \to \infty$ as $T \to \infty$. We assume that

$$\sup_T \big\{ b_T \lambda_{max}(a_T' a_T) \big\} < \infty,$$

where λ_{max} is the maximum eigenvalue of a real symmetric matrix.

Let

$$\mathbb{Y}_T(\theta) = b_T^{-1} \big\{ \mathbb{H}_T(\theta) - \mathbb{H}_T(\theta^*) \big\}.$$

Fix a constant $\alpha \in (0,1)$. Let

$$\mathbb{U}_T(r) = \big\{ u \in \mathbb{U}_T; r \le |u| \le b_T^{(1-\alpha)/2} \big\}.$$

Let ρ, ρ_1, ρ_2 and β_2 be real constants. Let $L > 0$. The following conditions were assumed in [46].

[A1] There exists a constant C_L such that

$$\sup_{T \in \mathbb{T}} P\left[\sup_{u \in \mathbb{U}_T(r)} (1 + |u|^2)^{-1} |r_T(u)| \ge r^{-\rho_1} \right] \le C_L r^{-L} \quad (r > 0)$$

The supremum on the empty set reads $-\infty$ by convention.

[A2] There exists a constant C_L such that

$$P[\lambda_{min}(\Gamma) < 4r^{-\rho_1}] \leq C_L r^{-L} \quad (r > 0)$$

Suppose that a random field $\mathbb{Y} : \Omega \times \bar{\Theta} \to \mathbb{R}$ is given.

[A3] There exists a positive random variable χ_0 such that

$$\mathbb{Y}(\theta) = \mathbb{Y}(\theta) - \mathbb{Y}(\theta^*) \leq -\chi_0 |\theta - \theta^*|^{\rho} \quad (\theta \in \bar{\Theta}).$$

[A4] The following inequalities hold:

$$0 < \rho_1 < 1, \quad \alpha\rho < \rho_2, \quad \beta_2 \geq 0, \quad 1 - 2\beta_2 - \rho_2 > 0.$$

[A5] There exists a constant C_L such that

$$P[\chi_0 \leq r^{-(\rho_2 - \alpha\rho)}] \leq C_L r^{-L} (r > 0)$$

[A6] $\sup_{T \in \mathbb{T}} E[|\Delta_T|^{M_1}] < \infty$ for $M_1 = L(1 - \rho_1)^{-1}$. Moreover,

$$\sup_{T \in \mathbb{T}} E\left[\left\{ \sup_{\theta \in \bar{\Theta}, |\theta - \theta^*| \geq b_T^{-\alpha/2}} b_T^{\frac{1}{2} - \beta_2} |\mathbb{Y}_T(\theta) - \mathbb{Y}(\theta)| \right\}^{M_2} \right] < \infty$$

for $M_2 = L(1 - 2\beta_2 - \rho_2)^{-1}$.

Even in the case where $\theta^* \in \partial\Theta$, the same proof as Theorem 1 of [46] can be applied to the following theorem.

Theorem 4.2. Suppose that Conditions [A1]-[A6] are satisfied for a given positive number L. Then there exists a constant C_L such that

$$P\left[\sup_{u \in \mathbb{V}_T(r)} \mathbb{Z}_T(u) \geq \exp\Big(-\frac{1}{2} r^{2 - (\rho_1 \vee \rho_2)} \Big) \right] \leq C_L r^{-L} \quad (r > 0, T \in \mathbb{T}), \quad (4.1)$$

where $\mathbb{V}_T(r) = \{u \in \mathbb{U}_T; |u| \geq r\}$.

We will consider simplified version of Theorem 2.1. Among many possibilities, we only treat $\mathbb{H}_T$ of class C^3. In what follows, we only consider a bounded convex open set Θ. An extension of the results in this appendix to a finite sum of bounded convex open sets is straightforward. Let $C^k(\bar{\Theta})$ be the space of functions $f : \bar{\Theta} \to \mathbb{R}$ such that $f|_{\Theta} \in C^k(\Theta)$

and the derivatives $\partial_\theta^i f \in C(\Theta; (\mathbb{R}^{\mathsf{d}})^{\otimes i})$ are continuously extended to $\partial\Theta$ for $i \in \{0, 1, \ldots, k\}, (\mathbb{R}^{\mathsf{d}})^{\otimes 0} = \mathbb{R}$. For $f \in C^k(\bar{\Theta})$, Taylor's formula is valid:

$$f(\theta) = \sum_{i=0}^{k-1} \frac{1}{k!} \partial_\theta^i f(\theta_0)[(\theta - \theta_0)^{\otimes i}] + \frac{1}{(k-1)!} \int_0^1 (1-s)^{k-1} \partial_\theta^k f(\theta_0 + s(\theta - \theta_0)) ds [(\theta - \theta_0)^{\otimes k}]$$

for $\theta, \theta_0 \in \bar{\Theta}$ since $\bar{\Theta}$ is convex and the derivatives $\partial_\theta^i f$ are continuously extended to $\partial\Theta$.

We will assume that the mapping $\mathbb{H}_T(\omega, \cdot) \in C^3(\bar{\Theta})$ for all T and $\omega \in \Omega$. Set $\beta = \alpha/(1-\alpha)$. We replace Condition [A4] by

[A4'] Parameters β_1, ρ_1, ρ_2 and β_2 satisfy the following inequalities: $0 < \beta_1 < 1/2, 0 < \rho_1 < \min\{1, \beta, 2\beta_1/(1-\alpha)\}, \alpha\rho < \rho_2, \beta_2 \geq 0$ and $1 - 2\beta_2 - \rho_2 > 0$.

Furthermore, we assume the following additional conditions:

[A1'] For $M_3 = L(\beta - \rho_1)^{-1}$,

$$\sup_{T \in \mathbb{T}} E\left[\left(b_T^{-1} \sup_{\theta \in \Theta} |\partial_\theta^3 \mathbb{H}_T(\theta)| \right)^{M_3} \right] < \infty.$$

Moreover, for $M_4 = L(\frac{2\beta_1}{1-\alpha} - \rho_1)^{-1}$,

$$\sup_{T \in \mathbb{T}} E\left[b_T^{\beta_1} |\Gamma_T(\theta^*) - \Gamma| \right] < \infty,$$

where a random matrix $\Gamma_T(\theta)$ is defined by

$$\Gamma_T(\theta)[u, u] = -\partial_\theta^2 \mathbb{H}_T(\theta)[a_T u, a_T u]$$

for $u \in \mathbb{R}^{\mathsf{d}}$.

Now it is easy to obtain a counterpart of Theorems 2 of [46].

Theorem 4.3. Let $L > 0$. Suppose that Conditions [A1'], [A2], [A3], [A4'], [A5] and [A6] are satisfied. Then there exists a constant C_L such that Inequality (4.1) holds for all $T > 0$ and $r > 0$.

Restricted tangents

For limit theorems, for simplicity, we will only consider the situation where $\mathbb{U}_T$ is asymptotically locally similar in the following sense. Let $\mathbb{S}$ be a subset of $\mathbb{R}^{\mathsf{d}}$.

[S1] For every $R > 0$, there exists $T_R \in \mathbb{T}$ such that

$$\{u \in \mathbb{U}_T; |u| \leq R\} = \mathbb{S} \cap \{u \in \mathbb{R}^{\mathsf{d}}; |u| \leq R\}$$

for all $T \in \mathbb{T}$ with $T \geq T_R$.

Write $\overline{B(0,R)} = \{u \in \mathbb{R}^{\mathsf{d}}; |u| \leq R\}$. Assume [S1]. Define $\mathbb{S}(R)$ by

$$\mathbb{S}(R) = \{u \in \mathbb{U}_{T_R}; |u| \leq R\} = \mathbb{S} \cup \overline{B(0,R)}.$$

Then

$$\mathbb{S} = \bigcup_{R>0} \mathbb{S}(R) = \bigcup_{R>0} \{u \in \mathbb{U}_{T_R}; |u| \leq R\} = \bigcup_{R>0} \bigcup_{T \geq T_R} \{u \in \mathbb{U}_T; |u| \leq R\}.$$

Since $\mathbb{S}(R) = \mathbb{U}_{T_R} \cap \overline{B(0,R)}$ is convex and increasing in R, the set $\mathbb{S}$ is convex. BY definition, $0 \in \mathbb{S}$. If $\theta^* \in \Theta$, then $\mathbb{S} = \mathbb{R}^{\mathsf{d}}$. Conditions [S1] imposes restrictions on the shape of Θ and a_T when $\theta^* \in \partial\Theta$.

Typically, the random field $\mathbb{Z}_T$ is locally asymptotically mixed normal. To simplify description, we set $r_T(u) = 0$ for $u \in \mathbb{R}^{\mathsf{d}} - \mathbb{U}_T$. Let $\mathcal{G}$ be a sum σ-field of $\mathcal{F}$ such that Γ is $\mathcal{G}$-measurable. Let Δ be a d-dimensional random variable defined on an extension $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{P})$ of $(\Omega, \mathcal{F}, P)$. Random variables on Ω are naturally extended on $\bar{\Omega}$. The $\mathcal{G}$-stable convergence is denoted by $d_s(\mathcal{G})$. Let

$$\mathbb{Z}(u) = \exp\Big(\Delta[u] - \frac{1}{2}\Gamma[u^{\otimes 2}]\Big) \quad (u \in \mathbb{R}^{\mathsf{d}}).$$

We denote $\mathbb{C}(R) = C(\mathbb{S}(R))$ for $R > 0$ and equip $\mathbb{C}(R)$ with the supremum norm. Then $(\mathbb{C}(R), \mathbb{B}[\mathbb{C}(R)])$ is a measurable space with the Borel σ-field $\mathbb{B}[\mathbb{C}(R)]$.

Proposition 4.4. Suppose that [S1] and the following conditions are satisfied.

(i) For every $R > 0$, $\sup_{u \in \mathbb{S}(R)} |r_T(u)| \to^p 0$ as $T \to \infty$.

(ii) $\Delta_T \to^{d_s(\mathcal{G})} \Delta$ as $T \to \infty$.

Then

$$\mathbb{Z}_T|_{\mathbb{C}(R)} \to^{d_s(\mathcal{G})} \mathbb{Z}|_{\mathbb{C}(R)} \quad (T \to \infty)$$

for every $R > 0$.

quasi-maximum likelihood estimator

A measurable mapping $\tilde{\theta}_T^M : \Omega \to \bar{\Theta}$ is called a quasi-maximum likelihood estimator (QMLE) with respect to $\mathbb{H}_T$ if

$$\mathbb{H}_T(\tilde{\theta}_T^M) = \max_{\theta \in \bar{\Theta}} \mathbb{H}_T(\theta).$$

Such a mapping exists according to the measurable selection theorem.

If Γ is positive definite a.s., then by convexity, $\mathbb{Z}_T$ takes its maximum in $\mathbb{S}$ at a single point $\tilde{u}_{\mathbb{S}}^M \in \mathbb{S}$ a.s.:

$$\tilde{u}_{\mathbb{S}}^M = \mathrm{argmax}_{u \in \mathbb{S}} \mathbb{Z}(u).$$

Let $\tilde{u}_T^M = a_T^{-1}(\tilde{\theta}_T^M - \theta^*)$.

Proposition 4.5. Suppose that [S1] and the following conditions are fulfilled.

(i) $\mathbb{Z}_T|_{\mathbb{C}(R)} \to^{d_s(\})} \mathbb{Z}|_{\mathbb{C}(R)}$ as $T \to \infty$ for every $R > 0$.

(ii) $\{\tilde{u}_T^M\}_T$ is tight.

Then $\tilde{u}_T^M \to^{d_s(\mathcal{F})} \tilde{u}_{\mathbb{S}}^M$ as $T \to \infty$.

Combining Theorem 3.2 with the PLD inequality, we obtain convergence of moments of $\tilde{u}_T^M$ as well as its tightness.

Theorem 4.6. Let $L > p > 0$. Suppose that Condition [S1] and the following conditions are fulfilled.

(i) $\Delta_T \to^{d_s(\mathcal{G})} \Delta$ as $T \to \infty$.

(ii) For every $R > 0$, $\sup_{u \in \mathbb{S}} |r_T(u)| \to^p 0$ as $T \to \infty$.

(iii) There exists a constant C_L such that

$$P\left[\sup_{u \in \mathbb{V}_T(r)} \mathbb{Z}_T(u) \geq 1\right] \geq C_L r^{-L} \quad (r > 0, T \in \mathbb{T})$$

Then

$$E[f(\tilde{u}_T^M)Y] \to E[f(\tilde{u}_{\mathbb{S}}^M)] \quad (T \to \infty).$$

for every continuous function $f : \mathbb{S} \to \mathbb{R}$ satisfying $\sup_{u \in \mathbb{S}}(1 + |u|)^{-p}|f(u)| < \infty$ and every bounded $\mathcal{G}$-measurable random variable Y.

Proof. We apply Propositions 3.1 and 3.2. Condition (iii) gives L^p boundedness of $\{\tilde{u}_T^M\}_{T \in \mathbb{T}}$ for $p < L$. □