Asymptotic properties of estimators and information criteria for random fields

（確率場に対する推定量および情報量規準の漸近的性質）

Tomonari SEI

清 智也

Abstract

This thesis studies the asymptotic analysis of statistical inference for stochastic processes. The stochastic processes considered here are mainly real-valued processes whose time-parameter ranges over a bounded region of Euclidean space. In contrast to the ordinary asymptotics on stochastic processes, the ergodic theory can not be used. A concept of the fixed domain asymptotics is adopted here. This means that the set of observed points converges to a dense subset on the fixed region. In this context, three new results are obtained. The first one gives an asymptotic property of the quasi-likelihood estimator for the fractal index of Gaussian processes. The second one shows the local asymptotic mixed normality of a class of transformed Gaussian models. The last one gives an information criterion for the locally asymptotically mixed normal models. The above results contribute to the prediction problem on spatial statistics. Before describing the results, some basic concepts and known properties on the fixed domain asymptotics, the asymptotic decision theory and the model selection theory are reviewed.

Contents

Chapter 1

Introduction

This thesis studies the asymptotic analysis of statistical inference for stochastic processes. The stochastic processes considered here are mainly real-valued processes whose parameter space is a bounded region of Euclidean space. For example, we deal with a stochastic process

$$X = (X_t \mid t \in [0,1]^d), \quad d \geq 1. \tag{1.1}$$

In *geostatistics*, data are observed on a bounded region. Therefore, to deal with stochastic processes like (1.1) is natural in that area. In contrast to the ordinary asymptotics on stochastic processes, the ergodic theory can not be used since the observed region is bounded. Instead of the ergodicity, a concept of the fixed domain asymptotics, or micro-ergodicity, is adopted here. This means that the observed points converge to a dense subset in the fixed region. In this context, three new results are obtained. The first one shows consistency and asymptotic normality of the quasi-maximum likelihood estimator for the fractal index of Gaussian processes. The second one shows the local asymptotic mixed normality of a class of transformed Gaussian models. The last one gives an information criterion for the locally asymptotically mixed normal models. The above results contribute to the prediction problem on spatial statistics. Before describing the results, some basic concepts and known properties on the fixed domain asymptotics, the asymptotic decision theory and the model selection theory are reviewed.

The first three chapters are review. In Chapter 2, we briefly review some definitions and results on spatial statistics. In particular, the estimation problem of the fractal index under the fixed domain asymptotics is reviewed. A simulating method for numerical experiments is touched on there. In Chapter 3, we summarize the asymptotic decision theory for estimation. The local asymptotic mixed normality is defined there. We also give the asymptotic decision theory for prediction as an analogy. There seems to be no past research on this handling of prediction problem. In Chapter 4, several information criteria are briefly explained. The local asymptotic maximum risk and regret of information criteria are defined. As an example, the local asymptotic maximum risk of AIC and BIC is considered.

The next three chapters describe our new results. In Chapter 5, we prove the consistency and asymptotic normality of an estimating method of fractal index proposed by M. L. Stein. In Chapter 6, we show that a class of transformed Gaussian model has local asymptotic mixed normality under an assumption that the hidden Gaussian process has independent increments. The proof is involved but some examples are given. In Chapter 7, we propose an information criterion called Bayes-LAMN-IC and discuss its predictive performance together with numerical experiments. The remaining works are touched on. Finally we give some discussions in Chapter 8.

Roughly speaking, three pairs of the chapters are closely related, respectively, as indicated in Table 1.1.

Table 1.1: Relations between chapters (The symbol $\Longleftrightarrow$ denotes "closely related").

Chapter 2	$\Longleftrightarrow$	Chapter 5
Chapter 3	$\Longleftrightarrow$	Chapter 6
Chapter 4	$\Longleftrightarrow$	Chapter 7

Some notations used throughout the paper are listed below.

- A symbol $\mathbb{I}_{(A)}$ for a proposition A is defined by $\mathbb{I}_{(A)} = 1$ if A is true, 0 otherwise. If A is defined on a probability space, $\mathbb{I}_{(A)}$ becomes the indicator function of the event that A is true.

- A symbol $\overset{\theta}{\rightsquigarrow}$ denotes the convergence in distribution, where the parameter θ corresponds to the true probability distribution. If the true distribution is clearly specified, we simply use $\rightsquigarrow$ instead of $\overset{\theta}{\rightsquigarrow}$.

- A symbol α is used for two meanings. One is the fractal index (Chapter 2 and 5). Another one is the index of submodels (Chapter 4 and 7).

- The index parameter of the stochastic processes, that is, the quantity t of (X_t), is called *time* regardless of its dimension.

- The two terms, random fields and stochastic processes, are identically used.

- For vectors and matrices A, the symbol A' is used for transposition unless otherwise stated.

- All nonnegative integers, all rational numbers and all real numbers are denoted by $\mathbb{N}$, $\mathbb{Q}$ are $\mathbb{R}$, respectively.

- Let x and y be any real numbers. Then the symbol $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. The symbol $\lceil x \rceil$ denotes the least integer greater than or equal to x. The symbol $x \wedge y$ and $x \vee y$ denote $\min(x, y)$ and $\max(x, y)$, respectively.

Chapter 2

Statistical inference of spatial data

In this chapter, some definitions and results on statistical inference of spatial data are briefly summarized. In particular, the estimation problem of the fractal index of Gaussian processes is reviewed.

2.1 Basic terms and notations

Stochastic processes considered here are $\mathbb{R}$-valued processes with the time parameter space $\mathbb{R}^d$ unless otherwise stated.

Definition 2.1. A stochastic process $(X_t \mid t \in \mathbb{R}^d)$ is (strictly) *stationary* if all of its joint distributions are invariant under any translation in $\mathbb{R}^d$. A stochastic process $(X_t \mid t \in \mathbb{R}^d)$ is *intrinsic* if its difference $(X_{t+s} - X_t \mid t \in \mathbb{R}^d)$ for any fixed s is a stationary process (e.g. Chilès & Delfiner (1999)).

Definition 2.2 (Gaussian random field). A stochastic process $(X_t \mid t \in \mathbb{R}^d)$ is called a Gaussian random field if its any joint distribution is multivariate normal.

Definition 2.3 (Transformed Gaussian model). A statistical model P_θ induced by a stochastic process X_t is called *a transformed Gaussian model* if X_t is written by an equation

$$g(X_t; \theta) \quad = \quad Y_t \tag{2.1}$$

with a parametric family of one-to-one functions $\{g(\cdot; \theta) \mid \theta \in \Theta\}$ and a Gaussian process Y_t subject to a parametric family $(Y_t) \sim \{Q_\theta \mid \theta \in \Theta\}$.

Remark 2.4. De Oliveira et al. (1997) use a term *Bayesian Transformed Gaussian model* to emphasize the Bayesian prediction.

2.2 Fixed domain asymptotics

Let us consider a set of finite samples $(X_t \mid t \in D)$, where a finite subset D of $\mathbb{R}^d$ denotes sampling points. The cardinality of D is denoted by n. When n is assumed to increase, several types of asymptotics can be considered. In particular, if $d \geq 2$, asymptotic properties are usually complicate due to the variousness of the shape of domain. Even if we restrict D to be a regular lattice, there are several possibilities of asymptotics: D converges to a dense subset of a bounded region, D converges to a dense subset of $\mathbb{R}^d$, or D converges to a lattice in $\mathbb{R}^d$. We adopt the first one. This is called the *fixed domain asymptotics* (Stein, 1995, 1999). It is also called infill asymptotics (Cressie, 1993) or micro-ergodicity (Chilès & Delfiner, 1999).

Stein (1999) studied efficiency of the best unbiased prediction under fixed domain asymptotics. He considered misspecification problem. Namely, he assumed two probability measures on a fixed region and evaluated the performance of prediction based on one of the measure when the true measure is another. The prediction problem he treated is interpolation problem. Interpolation is important for practical purposes. However, he gave only few asymptotic results on estimation or model selection, as he noted. We try to solve this problem by using *contiguous* probability measures in Chapter 3, 4 and 7. A model selection procedure is constructed from this context. However, a theory that is consistent to Stein's one is still not developed in this paper. In fact, the prediction problem we consider is essentially extrapolation since the predicted stochastic process is assumed to be independent of the observed stochastic process.

In time series analysis, the fixed domain asymptotics is considered by several researchers. Dohnal (1987) studied the fixed domain asymptotics on estimation of the diffusion coefficient, where the assumed model is a model of one-dimensional diffusion processes. The local asymptotic mixed normality (LAMN; see Chapter 3) was proved there. A main difficulty on the model is that the probability density function is not explicitly expressed. Genon-Catalot & Jacod (1993, 1994) generalized Dohnal's result to multivariate diffusions (but essentially integrable case) and randomized observations. Furthermore, they showed a minimum contrast estimator is optimal in the sense of LAMN. A related work is given by Yoshida (1997). A survey on this topic is given by Prakasa Rao (1999).

Further results associated with the fixed domain asymptotics on diffusion-process models are obtained by Sørensen & Uchida (2003) and Uchida (2003). They dealt with the case that the diffusion coefficient ϵ decreases as the number of observation n increases. The asymptotics that ϵ decreases under continuous observation is studied by Kutoyants (1984, 1994), Yoshida (1992a,b) and other researchers. A formal asymptotic expansion and a second-order efficiency are discussed by Sei & Komaki (2003) from the viewpoint of information geometry (Amari, 1987).

2.3 Estimation of fractal dimension

In this section, we describe past works on estimation of fractal dimension of Gaussian processes. In particular, their properties under fixed domain asymptotics are summarized. We concentrate to the one-dimensional case. Multi-dimensional and Non-Gaussian cases are touched on in the last of the section.

The assumed model is semiparametric. The underlying process $(X_t \mid t \in [0,1])$ is an intrinsic Gaussian process with the constant mean μ and the variogram

$$\gamma(t) \;=\; \frac{1}{2}\mathrm{E}[(X_{s+t}-X_s)^2] \;=\; \nu|t|^\alpha + \mathrm{o}(|t|^\alpha) \quad \text{as } |t| \downarrow 0. \tag{2.2}$$

The condition for existence of such a process is that $\gamma(s)$ is conditionally negative definite: $\sum_{i,j} a_i a_j \gamma(t_i - t_j) \le 0$ for any finite set $\{t_i\}$ and $\{a_i\}$ satisfying $\sum_i a_i = 0$. We call α the fractal index. The Hausdorff dimension D of the sample path is given by

$$D \;=\; 2 - \alpha/2 \tag{2.3}$$

(Adler, 1981). Typical sample paths for several α's are shown in Fig. 2.1.

Our goal is to find a good estimator of the parameter α from the discrete observations $\{X_{i/n} \mid i = 0, 1, \cdots, n\}$. The parameters ν and the remainder term in (2.2) are considered as (infinite-dimensional) nuisance parameters. The estimators studied by the early researchers are surveyed in this section. Details on the regularity conditions are omitted.

Example 2.5. The fractional Brownian motion B_t^H with the Hurst index $H \in (0,1)$ is defined by

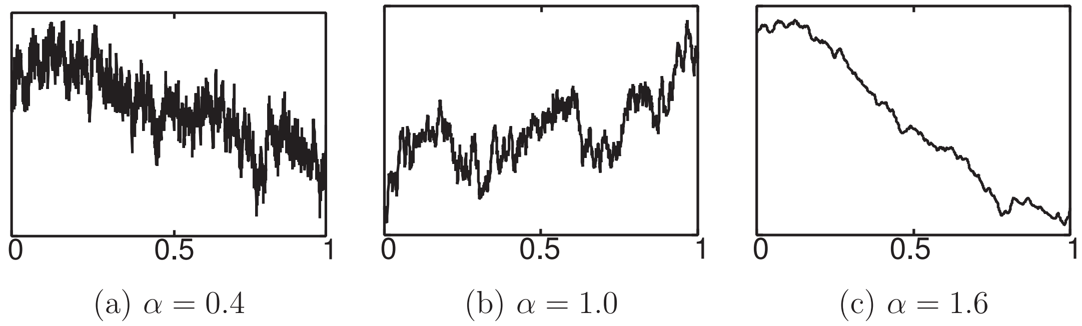
$$B_t^H \;=\; \frac{1}{\Gamma(H+1/2)}\left[\int_{-\infty}^0 [(t-s)^{H-1/2} - (-s)^{H-1/2}]\mathrm{d}B_s + \int_0^t (t-s)^{H-1/2}\mathrm{d}B_s\right],$$

where $(B_t \mid t \in (-\infty, \infty))$ is the standard Brownian motion (see e.g. Mandelbrot & van Ness (1968)). The variogram is given by

$$\frac{1}{2}\mathrm{E}\left[(B_t^H - B_s^H)^2\right] \;=\; \frac{1}{2}|t-s|^{2H}.$$

It satisfies (2.2) with $\alpha = 2H$ and $\nu = 1/2$. □

Many estimators are regression-type estimators. Fix an integer k. The basic idea is to prepare certain quantities y_l and x_l for $l = 1, \cdots, k$ such that $y_l \simeq C_l x_l^\alpha$ where C_l



(a) $\alpha = 0.4$ (b) $\alpha = 1.0$ (c) $\alpha = 1.6$

Figure 2.1: Sample paths of X_t for several α.

does not depend on α. Then an estimator $\hat{\alpha}$ can be defined as the ordinary least squares estimator of the regression coefficient:

$$\hat{\alpha} = \frac{\sum_{l=1}^{k}(\log x_l - k^{-1}\sum_{l'=1}^{k}\log x_{l'})\log y_l}{\sum_{l=1}^{k}(\log x_l - k^{-1}\sum_{l'=1}^{k}\log x_{l'})^2}. \tag{2.4}$$

The asymptotic properties of *box-counting estimator* are studied by Hall & Wood (1993). The estimator is a regression-based estimator (2.4) with

$$y_l = \left[\sum_{i=1}^{\lfloor n/ml \rfloor}(U_{il} - V_{il})\right]^2, \quad x_l = l, \tag{2.5}$$

$$U_{il} = \max_{t \in B(i,l)} X_t, \quad V_{il} = \min_{t \in B(i,l)} X_t, \tag{2.6}$$

$$B(i,l) = \{(i-1)ml/n, [(i-1)m+1]l/n, \cdots, iml/n\}, \tag{2.7}$$

where m is a fixed integer. The set $B(i,l)$ is i-th block of $m+1$ points when the data are sampled at every multiple of l/n. A quantity $(ml/n)(U_{il} - V_{il})$ is area of the rectangle covering the graph $\{(t, X_t) \mid t \in B(i,l)\}$ (see Fig. 2.2). A fact that $U_{il} - V_{il}$ is approximated by $(ml/n)^{\alpha/2}$ is used.

The *variogram-based estimator* is proposed by Constantine & Hall (1994). It is a regression-based estimator (2.4) with

$$y_l = \frac{1}{n-l+1}\sum_{j=0}^{n-l}(X_{(j+l)/n} - X_{j/n})^2, \quad x_l = l. \tag{2.8}$$

This is approximated by $\nu(l/n)^{\alpha}$ from the assumption.

The *increment-based estimator* is proposed by Kent & Wood (1997) and Istas & Lang (1997), independently. This estimator is a generalization of the variogram-based estimator. An increment of order $p \geq 0$ is a finite vector $a = (a_j \mid j = -J, \cdots, J)$ for some

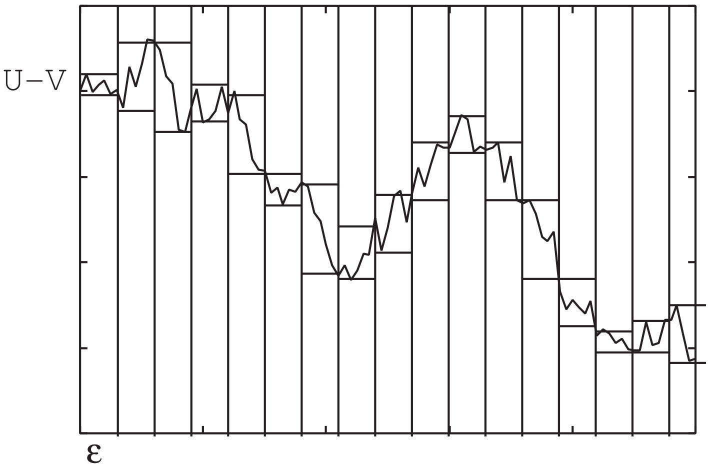


Figure 2.2: Covering by rectangles ($\epsilon = ml/n$).

$J > 0$ such that

$$\sum_{j=-J}^{J} j^r a_j \quad = \quad 0 \quad \text{if } 0 \leq r \leq p, \tag{2.9}$$

$$\sum_{j=-J}^{J} j^{p+1} a_j \quad \neq \quad 0. \tag{2.10}$$

For each $l = 1, \cdots, k$, let $Y^l(i)$ be a filtered process defined by

$$Y^l(i) \quad = \quad \sum_{j=-J}^{J} a_j X_{(i+jl)/n}, \tag{2.11}$$

for $i = Jl, Jl+1, \cdots, n - Jl$. Then the increment-based estimator is a regression-type estimator (2.4) with

$$y_l = \frac{1}{n - 2Jl + 1} \sum_{i=Jl}^{n-Jl} Y^l(i)^2, \quad x_l = l. \tag{2.12}$$

The generalized least squares method is also proposed by Kent & Wood (1997). As a generalization of their result, Chan & Wood (2004) shows that the same method as above works well for transformed Gaussian processes.

The *periodogram-based estimator* is proposed by Chan et al. (1995). The estimator is defined by a regression-type estimator (2.4) with

$$y_l = A(\omega_l)^2, \quad x_l = \omega_l^{-1}, \tag{2.13}$$

$$A(\omega) = \int_0^1 X_t \cos(\omega(2t-1)) \mathrm{d}t, \tag{2.14}$$

where $\omega_1, \dots, \omega_k$ are arbitrary frequencies.

An estimator based on the number of level crossing is proposed by Feuerverger et al. (1994). We call it the *level-crossing estimator*. Let $Y_h(t)$ is the smoothed process defined by

$$Y_h(t) \;=\; h^{-1}\int_{\mathbb{R}} K(u/h) X_{t+u} \mathrm{d}u, \tag{2.15}$$

where $K : \mathbb{R} \to \mathbb{R}$ is a kernel function. The number of the level crossing $N_h(u)$ is defined by

$$N_h(u) \;=\; \sharp\{t \in (0,1) \mid Y_h(t) = u\}. \tag{2.16}$$

The averaged number of the level crossing is given by

$$M(h) \;=\; \int_{\mathbb{R}} N_h(u)\mathrm{d}u \;=\; \int_0^1 |Y_h'(t)|\mathrm{d}t \tag{2.17}$$

Then the estimator is defined by a regression-type estimator (2.4) with

$$y_l = M(h_l), \quad x_l = h_l, \tag{2.18}$$

where $h_1, \dots, h_k$ are taken to decrease as $n \to \infty$.

The wavelet-shrinkage estimator is proposed by Wang (1997). They use a fact that the decay rate of wavelet coefficients is related to the fractal index. Let $\psi_{j,k}(t) = 2^{j/2}\psi(2^j x - k)$ $(j = 0, 1, \cdots; k = 0, 1, \cdots, 2^j - 1)$ with an appropriate wavelet $\psi(\cdot)$. Let $\hat{X}_{j,k}$ be the (j,k)-wavelet coefficient of the process X, that is,

$$\hat{X}_{j,k} \;=\; \sum_{i=1}^{n} X_{i/n}\psi_{j,k}(i/n), \tag{2.19}$$

$$X_t \;=\; \sum_{j=0}^{J_n}\sum_{k=0}^{2^j-1} \hat{X}_{j,k}\psi_{j,k}(t). \tag{2.20}$$

The wavelet-shrinkage method is a method that exchanges the empirical wavelet coefficients $\hat{X}_{j,k}$ by $\delta_{t_n}(\hat{X}_{j,k}) = \mathrm{sgn}(\hat{X}_{j,k})(|\hat{X}_{j,k}| - t_n)_+$. Then the wavelet-shrinkage estimator is defined by

$$\hat{\alpha} \;=\; 1 - \frac{2\log\sum_{k=1}^{2^{j_n}} \delta_{t_n}(\hat{X}_{j_n,k})}{j_n}. \tag{2.21}$$

Stein (1995) proposed an estimator based on Whittle's likelihood. We call it the *quasi-maximum likelihood estimator (QMLE)*. It minimizes an approximated expression of negative log-likelihood

$$\int_{-\pi}^{\pi} \frac{I_n(\lambda)}{f(\lambda|\alpha)}\mathrm{d}\lambda, \tag{2.22}$$

where $I_n(\lambda)$ is the periodogram of the difference $X_{i/n} - X_{(i-1)/n}$ and $f(\lambda|\alpha)$ is the spectral density of a long-memory process called the fractional Gaussian noise. The fractional Gaussian noise is reviewed in the next section. We elucidate the properties of QMLE in Chapter 5.

Asymptotic order of variance and bias for each estimator is summarized in Table 2.1. Here we assume that the variogram of the underlying process X is

$$\frac{1}{2}\mathrm{E}[(X_t - X_0)^2] \quad = \quad \nu|t|^{\alpha} + \nu_2|t|^{\alpha+\beta} + o(|t|^{\alpha+\beta}) \tag{2.23}$$

for some $\beta > 0$. Regularity conditions on differentiability of the variogram function are omitted. From the table, the box-counting and variogram-based estimators have larger order of variance than the increment-based estimator ($p \geq 1$) when $\alpha \in (\frac{3}{2}, 2)$. The periodogram-based and level-crossing estimators need non-trivial selection of k and h, respectively. Therefore we conclude that the best estimator except the wavelet-shrinkage and QMLE is the increment-based estimator.

Table 2.1: Asymptotic order of variance and bias for each estimator. The quantity β appears in (2.23). For the periodogram-based estimator, k denotes the number of frequencies for which the periodogram is calculated. For the level-crossing estimator, h denotes the bandwidth of smoothing. For QMLE, it is assumed that $\alpha > 1$ (Chapter 5). The symbol '-' denotes that the author is unaware of the order.

Estimator Name	Variance		Bias
	$\alpha \in (0, \frac{3}{2})$	$\alpha \in (\frac{3}{2}, 2)$	$\alpha \in (0, 2)$
Box-counting	n^{-1}	$n^{-(4-2\alpha)}$	$n^{-\beta}$
Variogram-based	n^{-1}	$n^{-(4-2\alpha)}$	$n^{-\beta}$
Increment-based ($p \geq 1$)	n^{-1}	n^{-1}	$n^{-\beta}$
Periodogram-based	k^{-1}	k^{-1}	$k^{-\beta}$
Level-crossing	h	$h^{4-2\alpha}$	h^{β}
Wavelet-shrinkage	-	-	-
QMLE	n^{-1}	n^{-1}	$n^{-\beta}$

We have treated so far an estimating problem of a fractal index for the observed process with time parameter space $\mathbb{R}$. The case of the multi-dimensional time is also needed from a practical point of view. The increment-based estimators are also discussed in multi-dimensional case: Davies & Hall (1999), Chan & Wood (2000), Zhu & Stein (2002) and Chan & Wood (2004). Stein's QMLE (Stein, 1995) is proposed also for multi-dimensional case but not theoretically analyzed. For the multi-dimensional case, the edge effects (Guyon, 1982) must be taken into account as Stein noted.

Another generalization is estimation problem of the fractal index of transformed Gaussian random fields. The results of Hall & Roy (1994) and Chan & Wood (2004) are in this direction.

2.4 Methods of simulation

Generating methods of a sample path from a given correlation function is briefly described here. A naive method is to use the Cholesky decomposition of the correlation matrix, but this method spends much computational time. If the observing points of the process are regular lattice, the fast Fourier transform (FFT) is useful. For periodic stochastic processes, the method is clear. Even if the underlying process is not periodic, one can usually embed it to a periodic process since observed region is bounded under fixed domain asymptotics.

Chilès & Delfiner (1999) gave the embedding method for the one-dimensional fractional Gaussian noise. Stein (2001, 2002) gave its generalization to multi-dimensional case, although a restriction to the parameter range exists. Gneiting (2002) proposed to change the correlation function by multiplying a correlation function with a compact support.

In Chapter 5, we adopt more practical method: one executes FFT anyway and checks whether the resulting Fourier coefficients are positive or not. If the coefficients are all positive, one can use FFT method. If some coefficients are negative, one must use the Cholesky decomposition.

2.5 Fractional Gaussian noise

We briefly review the fractional Gaussian noise and an estimation problem related to it. It is needed in Chapter 5. See Beran (1994) and Yajima (2003) for details.

Definition 2.6. The fractional Gaussian noise (FGN) for the discrete-time process $\{Z_i \mid i = 1, 2, \cdots\}$ is a stationary Gaussian process with mean $\mathrm{E}[Z_i] = 0$ and covariance $\mathrm{E}[Z_i Z_{i+h}] = (A/2)\{|h+1|^\alpha + |h-1|^\alpha - 2|h|^\alpha\}$, where $A > 0$ and $\alpha \in (0, 2)$. □

The FGN is introduced by Mandelbrot & van Ness (1968) for modeling self-affine and long-range phenomena. It is identified with the difference of the fractional Brownian motion B_t^H defined in Example 2.5:

$$Z_i \;=\; \sqrt{A}\{B_{i+1}^H - B_i^H\}, \quad i = 1, \cdots, n, \tag{2.24}$$

where $H = \alpha/2$. The spectral density of FGN is

$$f(\lambda) \;=\; f(\lambda|\alpha) \;=\; 2As(\alpha)\sin^2(\lambda/2)\sum_{j=-\infty}^{\infty}|2\pi j+\lambda|^{-\alpha-1}, \quad \lambda\in(-\pi,\pi], \quad (2.25)$$

where $s(\alpha) = \Gamma(\alpha+1)\sin(\pi\alpha/2)/\pi$. We have $f(\lambda) \simeq (As(\alpha)/2)|\lambda|^{-\alpha+1}$ as $|\lambda| \to 0$. Hence, if $\alpha > 1$, the process is a long-memory process. The following proposition about $f(\lambda)$ holds (Fox & Taqqu, 1986). We take A such that $\int_{-\pi}^{\pi}\log f(\lambda)\mathrm{d}\lambda = 0$ for the sake of convenience.

Proposition 2.7. *Let $\partial_\alpha = \partial/\partial\alpha$ and $\partial_\lambda = \partial/\partial\lambda$. Then, for each $\delta > 0$ and $k \in \mathbb{N}$, there exists a constant $C = C(\delta, k) > 0$ such that*
(1) *$f(\lambda|\alpha)$ is continuous at all (λ,α), $\lambda \neq 0$ and $f(\lambda|\alpha) \geq C^{-1}|\lambda|^{-\alpha+1+\delta}$.*
(2) *For each $0 \leq p \leq k$ and $0 \leq q \leq k$, $\partial_\lambda^p\partial_\alpha^q f(\lambda|\alpha)$ is continuous at all (λ,α), $\lambda \neq 0$ and*

$$|\partial_\lambda^p\partial_\alpha^q f(\lambda|\alpha)| \;\leq\; C|\lambda|^{-\alpha+1-p-\delta}.$$

As a result of (1) and (2), a function $g(\alpha) = \int_{-\pi}^{\pi}\log f(\lambda|\alpha)\mathrm{d}\lambda$ can be differentiated arbitrarily times under the integral sign. □

Whittle's quasi likelihood is defined by

$$\exp\left[-\frac{n}{4\pi}\int_{-\pi}^{\pi}\left\{\log\sigma^2+\frac{I_n(\lambda)}{\sigma^2 f(\lambda|\alpha)}\right\}\mathrm{d}\lambda\right], \quad (2.26)$$

where $I_n(\lambda) = (2\pi n)^{-1}|\sum_{j=1}^{n}\mathrm{e}^{-\sqrt{-1}j\lambda}Z_j|^2$. The quasi-maximum likelihood estimator is

$$\hat{\alpha}_n \;=\; \operatorname{argmin}\hat{\sigma}_n^2(\alpha), \quad (2.27)$$
$$\hat{\sigma}_n^2(\alpha) \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{I_n(\lambda)}{f(\lambda|\alpha)}\mathrm{d}\lambda. \quad (2.28)$$

Then $\hat{\alpha}_n$ is strongly consistent, asymptotically normal (Fox & Taqqu, 1986) and asymptotically efficient (Dahlhaus, 1989). The asymptotic variance is equal to the inverse J^{-1} of the Fisher information

$$J \;=\; \frac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\frac{\partial\log f(\lambda|\alpha)}{\partial\alpha}\right\}^2\mathrm{d}\lambda.$$

The graph of J^{-1} is shown in Fig. 2.3.

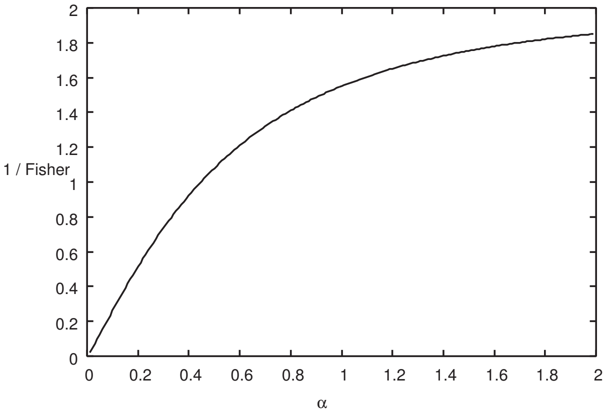


Figure 2.3: The inverse J^{-1} of Fisher information for each α.

Chapter 3

Asymptotic decision theory

In this chapter, the facts on asymptotic decision theory are summarized. In particular, the local asymptotic minimax theorem for prediction problem is proved.

3.1 Contiguity

An important concept in asymptotic decision theory is contiguity of two sequences of probability measures. We first prepare some basic terms and facts.

Definition 3.1. Let $\{P_n\}$ and $\{Q_n\}$ be two sequences of probability measures on a sequence of measurable spaces $(\Omega_n, \mathcal{B}_n)$. Let $\{A_n\}$ be a sequence of measurable sets. It is said that $\{Q_n\}$ is *contiguous to* $\{P_n\}$ if $Q_n(A_n) \to 0$ whenever $P_n(A_n) \to 0$. If P_n is contiguous to Q_n and vice versa, then P_n and Q_n are called *contiguous.* □

The following two lemmas are useful. The proofs are given, for example, in Chapter 6 of van der Vaart (1998).

Lemma 3.2 (Le Cam's first lemma). *The following conditions are equivalent.*

(i) Q_n *is contiguous to* P_n.
(ii) If $\mathrm{d}P_n/\mathrm{d}Q_n \overset{Q_n}{\rightsquigarrow} U$ *along a subsequence, then* $\mathrm{P}[U > 0] = 1$.
(iii) If $\mathrm{d}Q_n/\mathrm{d}P_n \overset{P_n}{\rightsquigarrow} V$ *along a subsequence, then* $\mathrm{E}[V] = 1$.
(iv) For any statistic $T_n : \Omega_n \to \mathbb{R}^k$*: If* $T_n \to^{P_n} 0$*, then* $T_n \to^{Q_n} 0$. □

Lemma 3.3 (Le Cam's third lemma). *Let* $X_n : \Omega_n \to \mathbb{R}^k$ *be a sequence of random variables. Suppose that* Q_n *is contiguous to* P_n *and* $(X_n, \mathrm{d}Q_n/\mathrm{d}P_n) \overset{P_n}{\rightsquigarrow} (X, V)$*. Then* $L(B) = \mathrm{E}[1_B(X)V]$ *defines a probability distribution, and* $X_n \overset{Q_n}{\rightsquigarrow} L$. □

Elementary examples are given below. The first example is treated in Section 3.2.

Example 3.4. Let $p_n(x|\theta)$ be the product density of a probability density $p(x|\theta)$ on $\mathbb{R}$. Then $p_n(x|\theta + h/\sqrt{n})$ and $p(x|\theta)$ are contiguous under mild conditions. See e.g. Chapter 7 of van der Vaart (1998). □

Example 3.5. Let $U(a,b)$ be the uniform distribution on the interval $[a,b]$ in $\mathbb{R}$. Let $P_n = U(0,p_n)$ and $Q_n = U(0,q_n)$ with positive numbers p_n and q_n. Then Q_n is contiguous to P_n if and only if $\limsup_n (q_n/p_n) \le 1$. □

3.2 LAQ, LAMN and LAN

Let Θ be an open subset of $\mathbb{R}^k$. We consider the following conditions on a sequence of models $\mathcal{P}_n = \{P_{\theta,n} \mid \theta \in \Theta\}$. If the measures have density functions with respect to a common measure, we write like $\mathcal{P}_n = \{p(x|\theta) \mid \theta \in \Theta\}$.

[LA1] There exists a sequence $\{\gamma_n\}$ of positive numbers such that for any convergent sequence $h_n \to h \in \mathbb{R}^k$, two measures $P_{\theta+\gamma_n h_n,n}$ and $P_{\theta,n}$ are contiguous.

[LA2] There exists a sequence (ξ_n, J_n) such that

$$\log \frac{\mathrm{d}P_{\theta+\gamma_n h_n,n}}{\mathrm{d}P_{\theta,n}} - \left(h_n' J_n \xi_n - \frac{1}{2} h_n' J_n h_n\right) \xrightarrow{\theta} 0 \tag{3.1}$$

for any convergent sequence $h_n \to h \in \mathbb{R}^k$. Further, the matrices J_n converges to J in law, where J is almost surely positive definite and generally depends on θ.

[LA3] For any convergent sequence $h_n \to h$, the limit distribution of J_n under $P_{\theta+h_n/\sqrt{n},n}$ is independent of h.

[LA4] The matrix J is not random.

Definition 3.6 (LAQ,LAMN,LAN). If a model $\mathcal{P}_n$ satisfies [LA1] and [LA2], then it is called *a locally asymptotically quadratic (LAQ) model.* If an LAQ model satisfies also [LA3], then it is called *a locally asymptotically mixed normal (LAMN) model.* If an LAMN model satisfies also [LA4], then it is called *a locally asymptotically normal (LAN) model.* □

Example 3.7 (AR model). Let X_t be an autoregressive (AR) process defined by

$$X_t = \theta X_{t-1} + \epsilon_t, \quad \theta \in \mathbb{R}, \quad t = 1, 2, \cdots . \tag{3.2}$$

For simplicity, assume $\epsilon_t \sim N(0,1)$. Then the model for $(X_1, \ldots, X_n)$ is LAQ (but not LAMN) at $|\theta| = 1$, LAMN (but not LAN) at $|\theta| > 1$, and LAN at $|\theta| < 1$, respectively (See e.g. Chapter 9 of van der Vaart (1998)). □

We prepare the following additional condition and give a necessary and sufficient condition to LAMN.

[LA3'] The sequence (ξ_n, J_n) in [LA2] converges to a random variable (ξ, J) in law. Further, ξ given J is normal with mean 0 and variance J^{-1}.

Theorem 3.8 (Equivalent condition to LAMN). *The conditions [LA2] and [LA3'] imply [LA1]–[LA3]. The converse also holds.*

Proof. The equivalence of [LA3] and [LA3'] under [LA1] and [LA2] is proved in Le Cam & Yang (2000). It is sufficient to show that [LA2] and [LA3'] imply [LA1]. This follows from the Le Cam's first lemma. □

Remark 3.9. Several authors assume the conditions [LA1]–[LA3] only for any constant sequence $h_n = h \in \mathbb{R}$ (e.g. Ibragimov & Has'minskii (1981)), then we can construct the following unusual example:

$$P_{\theta,n} = \begin{cases} N(\theta, 1/n) & \text{if } \theta \neq 1/n, \\ N(\theta, 1) & \text{if } \theta = 1/n. \end{cases} \tag{3.3}$$

This model satisfies [LA1]–[LA4] with $\gamma_n = 1/\sqrt{n}$ for any constant sequence $h_n = h$ but not for $h_n = 1/\sqrt{n}$ at $\theta = 0$. Nevertheless, the theorems in Section 3.4 hold under this weaker conditions. If we plug-in the maximum likelihood estimator $\hat{\theta}$ to the parameter θ, it is more convenient to assume the stronger condition [LA1]-[LA4]. □

3.3 Decision theory on estimation

In this section, we give only a brief description about decision theory on estimation. The details of notations and proofs are given in Chapter 8 of van der Vaart (1998). We assume that the model is LAN. The LAMN case is similarly described by Jeganathan (1982, 1983).

3.3.1 Non-asymptotic results

Let $\xi \sim N(h, \Sigma)$, where $h \in \mathbb{R}^k$ is an unknown vector and $\Sigma = J^{-1} \in \mathbb{R}^{k \times k}$ is a known matrix.

Lemma 3.10 (Convolution). *Let $m \in \mathbb{N}$ and $A \in \mathbb{R}^{m \times k}$. The null distribution L of any randomized equivariant-in-law estimator of Ah can be decomposed as $L = N(0, A\Sigma A') * M$ for some probability measure M, where $*$ denotes the convolution. The only randomized equivariant-in-law estimator for which M is degenerate at 0 is $A\xi$.* □

The loss function ℓ defined on $\mathbb{R}^k$ is called bowl-shaped if the level set $\{x \mid \ell(x) \leq c\}$ is convex and symmetric about the origin for any $c \geq 0$. The risk r of an estimator T of Ah is defined by $r(h) = \mathrm{E}_h[\ell(T - Ah)]$.

Lemma 3.11 (Anderson's lemma). *For any bowl-shaped loss function ℓ on $\mathbb{R}^k$, every probability measure M on $\mathbb{R}^k$ and every covariance matrix Σ,*

$$\int \ell \mathrm{d}N(0, \Sigma) \quad \leq \quad \int \ell \mathrm{d}[N(0, \Sigma) * M]. \tag{3.4}$$

□

Lemma 3.12 (Minimax property). *For any bowl-shaped loss function ℓ, the maximum risk of any randomized estimator T of Ah is bounded below by $\mathrm{E}_0\ell(A\xi)$. Consequently, $A\xi$ is a minimax estimator for Ah.* □

3.3.2 Asymptotic results

The following theorems are important consequence of LAN.

Theorem 3.13 (Convolution). *Let $(P_{\theta,n} \mid \theta \in \Theta)$ be a LAN model. Let ψ be differentiable at θ. Let T_n be a regular estimator sequence, that is, for any h, $\sqrt{n}(T_n - \psi(\theta + h/\sqrt{n}))$ converges to a limit distribution L_θ under the true parameter $\theta + h/\sqrt{n}$. Then there exist a probability measure M_θ such that $L_\theta = N(0, \dot{\psi}_\theta J_\theta^{-1} \dot{\psi}'_\theta) * M_\theta$.* □

Theorem 3.14 (Local asymptotic minimax theorem). *Let $(P_{\theta,n} \mid \theta \in \Theta)$ be a LAN model. Let ψ be differentiable at θ. Let T_n be any estimator sequence. Then for any bowl-shaped loss function ℓ*

$$\sup_{I \in F(\mathbb{R}^k)} \liminf_{n \to \infty} \max_{h \in I} \mathrm{E}_{\theta + h/\sqrt{n}} \ell(\sqrt{n}(T_n - \psi(\theta + h/\sqrt{n}))) \quad \geq \quad \int \ell \mathrm{d}N(0, \dot{\psi}_\theta J_\theta^{-1} \dot{\psi}'_\theta). \tag{3.5}$$

Here $F(\mathbb{R}^k)$ denotes all finite subsets of $\mathbb{R}^k$. □

3.4 Decision theory on prediction

3.4.1 Formulation of prediction problem

We fix a LAN model $\{p_n(x|\theta) \mid \theta \in \Theta \subset \mathbb{R}^k\}$ throughout this section. The results are naturally generalized to LAMN models. The corresponding *limit model* is defined by $\{p(\xi|h) = \phi(\xi|h, J^{-1}) \mid h \in \mathbb{R}^k\}$, where h, ξ and J are defined in the conditions [LA1]–[LA4] of Section 3.2 and $\phi(x|\mu, \Sigma)$ is the density of normal distribution with the mean vector μ and the covariance matrix Σ. Although we can always take the identity matrix for J by a coordinate transformation, we leave it since J plays a role of the metric.

Let x and y be random variables according to $p_n(\cdot|\theta)$. We consider a prediction problem of the distribution of y from the observed value x. A predictive density function, or a density estimator, is written by $q_n(y|x)$. A typical predictive density is the plug-in predictive density $q_n(y|x) = p_n(y|\hat{\theta})$ with some estimator $\hat{\theta}$. Another typical one is the Bayesian predictive density $q_n(y|x) = \int p_n(y|\theta)p_n(\theta|x)\mathrm{d}\theta$, where $p_n(\theta|x)$ is the posterior density under a prior distribution. The loss function of the prediction is defined by the Kullback-Leibler divergence $\int p_n(y|\theta)[\log p_n(y|\theta)/q_n(y|x)]\mathrm{d}y$. The risk is $\int\int p_n(x|\theta)p_n(y|\theta)[\log p_n(y|\theta)/q_n(y|x)]\mathrm{d}y\mathrm{d}x$. The symbol E_θ denotes expectation with respect to both x and y under the true parameter θ. We usually abbreviate an index θ as $J = J_\theta$.

3.4.2 Non-asymptotic results

We first consider the prediction problem for limit models. The problem is prediction of η from an observation ξ, where η and ξ are independently and identically distributed according to $p(\eta|h)$ and $p(\xi|h)$ with true parameter $h \in \mathbb{R}^k$. Expectations are taken with respect to both ξ and η: $\mathrm{E}_h f(\xi,\eta) = \int\int f(\xi,\eta)p(\xi|h)p(\eta|h)\mathrm{d}\xi\mathrm{d}\eta$.

We denote $E = E_k$ as the identity matrix of size k. For Bayesian inference, we often use a prior density function $p(h|\Lambda) = \phi(h|0,\Lambda)$ with some covariance matrix Λ. Recall that ϕ is the probability density of the normal distribution. We also consider the Lebesgue measure $P(\mathrm{d}h)$ on $\mathbb{R}^k$ as an *improper prior distribution*. We call it the *uniform prior*. A positive measure $P(\mathrm{d}h)$ with the infinite total measure is called an improper prior distribution if the posterior distribution

$$P(\mathrm{d}h|\xi) \quad = \quad \frac{p(\xi|h)P(\mathrm{d}h)}{\int p(\xi|h)P(\mathrm{d}h)} \tag{3.6}$$

expresses a probability distribution. If we take $p(\xi|h)$ and $P(\mathrm{d}h)$ as above, we obtain $P(\mathrm{d}h|\xi) = \phi(h|\xi,J)\mathrm{d}h$. Advantages to use the uniform prior are described later.

The loss of a predictive distribution $q(\eta|\xi)$ is defined by the Kullback-Leibler divergence

$$l_h(q(\cdot|\xi)) \quad = \quad \int p(\eta|h)\log\frac{p(\eta|h)}{q(\eta|\xi)}\,\mathrm{d}\eta,$$

The risk is denoted by $r_h(q) = \int p(\xi|h)l(q(\cdot|\xi))\,\mathrm{d}\xi$.

The next two lemmas are elementary obtained.

Lemma 3.15. *Assume that the prior distribution for h is $N(0,\Lambda)$. Then the Bayesian predictive density $q^{\Lambda}(\eta|\xi)$ is*

$$q^{\Lambda}(\eta|\xi) \;=\; \phi(\eta|A\xi,\Sigma), \tag{3.7}$$

where $A=(J+\Lambda^{-1})^{-1}J$ and $\Sigma=J^{-1}+(J+\Lambda^{-1})^{-1}$. Twice the loss and risk are given by

$$\begin{aligned} 2l_h(q^{\Lambda}(\cdot|\xi)) &= (h-A\xi)'\Sigma^{-1}(h-A\xi)+\mathrm{tr}[\Sigma^{-1}J^{-1}]-k-\log\det[\Sigma^{-1}J^{-1}], \\ 2r_h(q^{\Lambda}) &= h'(E-A)'\Sigma^{-1}(E-A)h+\mathrm{tr}[A'\Sigma^{-1}AJ^{-1}] \\ &\quad +\mathrm{tr}[\Sigma^{-1}J^{-1}]-k-\log\det[\Sigma^{-1}J^{-1}]. \end{aligned}$$

□

Lemma 3.16. *Assume that the prior distribution of h is the uniform prior. Then the Bayesian predictive density $q^{\mathrm{B}}(\eta|\xi)$ is*

$$q^{\mathrm{B}}(\eta|\xi) \;=\; \phi(\eta|\xi,2J^{-1}). \tag{3.8}$$

Twice the loss and risk are given by

$$\begin{aligned} 2l_h(q^{\mathrm{B}}(\cdot|\xi)) &= (h-\xi)'J(h-\xi)/2-k/2+k\log 2, \\ 2r_h(q^{\mathrm{B}}) &= k\log 2. \end{aligned}$$

In particular, $l_h(q^{\Lambda}(\cdot|\xi))$ and $r_h(q^{\Lambda})$ tend to $l_h(q^{\mathrm{B}}(\cdot|\xi))$ and $r_h(q^{\mathrm{B}})$, respectively, as $\Lambda\to\infty$. Here $\Lambda\to\infty$ means that all the eigenvalues of Λ tend to ∞ simultaneously. □

We note the concept of randomization here. In general, the predictive density $q(\eta|\xi)$ does not need to be measurable with respect to (ξ,η), that is, one can consider *a randomized predictive density* $q(\eta|\xi,u)$ with a random variable u independent of ξ and η. The quantity u is considered as an ancillary statistic. The word "randomized" is usually omitted in the following because u can be always conditioned without any costs.

The following lemma is an analogy of Lemma 3.10. It states that the Bayesian predictive distribution is optimal in the class of translation-equivariant predictive distributions.

Lemma 3.17 (Translation-equivariant distribution). *Assume that $q(\eta|\xi)$ is equivariant, that is, $q(\eta+c|\xi+c)$ is independent of $c\in\mathbb{R}^k$. Then there exists a probability density function f such that*

$$q(\eta|\xi)=f(\eta-\xi). \tag{3.9}$$

Furthermore, the Bayesian predictive density $q^{\mathrm{B}}(\eta|\xi)$ under the uniform prior is the unique best equivariant predictive density.

Proof. Any predictive density is written as $q(\eta|\xi) = g(\eta - \xi, \eta + \xi)$ with a function g since (η, ξ) and $(\eta - \xi, \eta + \xi)$ are one-to-one. From the assumption, $g(\eta - \xi, \eta + \xi + 2c)$ is independent of c. Thus q is written as (3.9). The function f is indeed a probability density function because $\int f(x)\mathrm{d}x = \int q(\eta|\xi)\mathrm{d}\eta = 1$. We next prove the latter statement. Since the Bayesian predictive density function $q^{\mathrm{B}}(\eta|\xi) = \phi(\eta|\xi, 2J^{-1})$ under the uniform prior is same as the marginal density function $\phi(\eta - \xi|0, 2J^{-1})$ of $\eta - \xi$, we obtain

$$r_h(q) - r_h(q^{\mathrm{B}}) \;=\; \mathrm{E}_h\left[\log \frac{q^B(\eta|h)}{f(\eta - \xi)}\right] \;\geq\; 0 \tag{3.10}$$

for any probability density function f and any $h \in \mathbb{R}^k$. Here the equality holds if and only if $f(\eta - \xi) = q^{\mathrm{B}}(\eta|\xi)$. □

Lemma 3.18 (Minimax property of Bayesian prediction). *Let $q(\eta|\xi)$ be any predictive density. Then the following inequality holds*:

$$\sup_{h \in \mathbb{R}^k} r_h(q) \;\geq\; \sup_{h \in \mathbb{R}^k} r_h(q^{\mathrm{B}}) \;=\; r_0(q^{\mathrm{B}}). \tag{3.11}$$

Proof. For any predictive density $q(\eta|\xi)$ and any finite positive definite matrix Λ, we obtain

$$\begin{aligned} \sup_{h \in \mathbb{R}^k} r_h(q) &\geq \int r_h(q)\phi(h|0, \Lambda)\mathrm{d}h \\ &\geq \int r_h(q^{\Lambda})\phi(h|0, \Lambda)\mathrm{d}h \\ &= (1/2)\mathrm{tr}[(E-A)'\Sigma^{-1}(E-A)\Lambda] + (1/2)\mathrm{tr}[A'\Sigma^{-1}AJ^{-1}] \\ &\quad\; (1/2)\mathrm{tr}[\Sigma^{-1}J^{-1}] - (k/2) - (1/2)\log\det[\Sigma^{-1}J^{-1}], \end{aligned}$$

where $A = (J + \Lambda^{-1})^{-1}J$ and $\Sigma = J^{-1} + (J + \Lambda^{-1})^{-1}$. The second inequality above follows from the fact that the Bayesian predictive density attains the Bayes risk. Finally, putting $\Lambda = \lambda J^{-1}$ and taking $\lambda \to \infty$, we obtain

$$\sup_{h \in \mathbb{R}^k} r_h(q) \;\geq\; (k/2)\log 2. \tag{3.12}$$

The right hand side is the risk of $q^{\mathrm{B}}(\eta|\xi)$. □

3.4.3 Asymptotic results

Recall that $\{p_n(x|\theta) \mid \theta \in \Theta \subset \mathbb{R}^k\}$ is LAN. The following lemma is an analogy of Lemma 8.3 in van der Vaart (1998).

Lemma 3.19. *Let $q_n(y|x)$ be a sequence of predictive densities such that the joint distribution sequence $q_n(y|x)p_n(x|\theta)$ is contiguous to $p_n(y|\theta)p_n(x|\theta)$ and*

$$\frac{q_n(y|x)}{p_n(y|\theta)} \overset{\theta+h/\sqrt{n}}{\rightsquigarrow} L_h, \tag{3.13}$$

where L_h is some limit distribution depending on h. Then there exists a predictive distribution $q(\eta|\xi)$ such that $q(\eta|\xi)/p(\eta|0)$ has distribution L_h for every h.

Proof. By Theorem 7.10 in van der Vaart (1998), there exists a randomized statistic $T(\xi,\eta,u)$ such that

$$\frac{q_n(y|x)}{p_n(y|\theta)} \overset{\theta+h/\sqrt{n}}{\rightsquigarrow} T(\xi,\eta,u) \tag{3.14}$$

for any $h \in \mathbb{R}^k$, where (ξ,η) is distributed according to $p(\xi|h)p(\eta|h)$ and u is a random variable independent of (ξ,η). Define $q(\eta|\xi)$ by $T(\xi,\eta,u) = q(\eta|\xi)/p(\eta|0)$. We show that $q(\eta|\xi)$ is a conditional probability density function of η given ξ. The proof is similar to Le Cam's first lemma (van der Vaart, 1998, Lemma 6.4). Let

$$Q_n = q_n(y|x)p_n(x|\theta)\mathrm{d}x\mathrm{d}y, \quad P_n = p_n(y|\theta)p_n(x|\theta)\mathrm{d}x\mathrm{d}y, \quad \mu_n = \frac{P_n+Q_n}{2}.$$

Let

$$U_n = \frac{\mathrm{d}P_n}{\mathrm{d}Q_n} = \frac{p_n(y|\theta)}{q_n(y|x)}, \quad V_n = \frac{\mathrm{d}Q_n}{\mathrm{d}P_n} = \frac{q_n(y|x)}{p_n(y|\theta)}, \quad W_n = \frac{\mathrm{d}P_n}{\mathrm{d}\mu_n} = \frac{2p_n(y|\theta)}{p_n(y|\theta)+q_n(y|x)}.$$

Let ξ_n be that defined in the condition [LA2]. For any subsequence of n, there exists a further subsequence (denoting it by n for simplicity) such that

$$(U_n,\xi_n)\overset{Q_n}{\rightsquigarrow}(U,\xi), \quad (V_n,\xi_n)\overset{P_n}{\rightsquigarrow}(V,\xi), \quad (W_n,\xi_n)\overset{\mu_n}{\rightsquigarrow}(W,\xi) \tag{3.15}$$

with some random variables U, V and W. Since V is just equal to T under $h=0$, it is sufficient to prove that $\mathrm{E}[V|\xi]=1$. Let $f:[0,\infty]\to\mathbb{R}$ and $\phi:\mathbb{R}^k\to\mathbb{R}$ are bounded continuous functions. Since (3.15), we obtain

$$\mathrm{E}[f(U)\phi(\xi)] \;=\; \mathrm{E}[(2-W)f(W/(2-W))\phi(\xi)].$$

Take a sequence $\{f_i(x)\}_{i=1}^{\infty}$ such that $f_i(x) \searrow \mathbb{I}_{(x=0)}$. By using the dominated-convergence theorem, we obtain

$$\mathrm{E}[\mathbb{I}_{(U=0)}\phi(\xi)] \;=\; 2\mathrm{E}[\mathbb{I}_{(W=0)}\phi(\xi)]. \tag{3.16}$$

Similarly,

$$\mathrm{E}[f(V)\phi(\xi)] \;= \mathrm{E}[Wf((2-W)/W)\phi(\xi)].$$

By taking $f_i(x) \nearrow x$ and using the monotone convergence theorem, we obtain

$$\mathrm{E}[V\phi(\xi)] \;=\; \mathrm{E}[(2-W)\mathbb{I}_{(W>0)}\phi(\xi)] \;=\; 2\mathrm{E}[\mathbb{I}_{(W>0)}\phi(\xi)] - \mathrm{E}[W\phi(\xi)]. \tag{3.17}$$

Since $\mathrm{E}_{\mu_n}[W_n|x] = 1$, $\mathrm{E}_{\mu_n}[W_n\phi(\xi_n)] = \mathrm{E}_{\mu_n}[\phi(\xi_n)]$. By taking $n \to \infty$ and by using boundedness of W_n, we obtain

$$\mathrm{E}[W\phi(\xi)] \;=\; \mathrm{E}[\phi(\xi)]. \tag{3.18}$$

By (3.16), (3.17) and (3.18), we obtain

$$\mathrm{E}[\mathbb{I}_{(U=0)}\phi(\xi)] + \mathrm{E}[V\phi(\xi)] \;=\; \mathrm{E}[\phi(\xi)]. \tag{3.19}$$

By contiguity of Q_n to P_n, the first term in the left hand side is zero (Le Cam's first lemma). Thus $\mathrm{E}[V\phi(\xi)] = \mathrm{E}[\phi(\xi)]$ for any bounded continuous function ϕ. This implies $\mathrm{E}[V|\xi] = 1$. □

Definition 3.20 (Regularity). The predictive density function $q_n(y|x)$ is said to be regular if there exists a probability density function f such that

$$\frac{q_n(y|x)}{p_n(y|\theta)} \overset{\theta+h/\sqrt{n}}{\rightsquigarrow} \frac{f(\eta-\xi)}{p(\eta|0)} \tag{3.20}$$

for any h. □

Theorem 3.21. *Assume that $\{p_n(\cdot|\theta) \mid \theta \in \Theta\}$ is LAN. Let $q_n(y|x)$ be a regular predictive density. Then, for any $h_n \to h$,*

$$\liminf_{n\to\infty} \mathrm{E}_{\theta+h/\sqrt{n}}\left[\log\frac{p_n(y|\theta+h/\sqrt{n})}{q_n(y|x)}\right] \;\geq\; \mathrm{E}_h\left[\log\frac{p(\eta|h)}{q^{\mathrm{B}}(\eta|\xi)}\right]. \tag{3.21}$$

Proof. Note that $\int p\log(p/q) = \int p\,[\log(p/q) - 1 + (q/p)]$ for any densities p and q. By using the Portmanteau lemma for a non-negative continuous function (van der Vaart, 1998), we obtain

$$\liminf_{n\to\infty} \mathrm{E}_{\theta+h/\sqrt{n}}\left[\log\frac{p_n(y|\theta+h/\sqrt{n})}{q_n(y|x)}\right] \;\geq\; \mathrm{E}_h\left[\log\frac{p(\eta|h)}{q(\eta|\xi)}\right]. \tag{3.22}$$

The right hand side is minimized at $q = q^{\mathrm{B}}$ by Lemma 3.17. □

Theorem 3.22 (Local asymptotic minimax theorem). *Assume that $\{p_n(\cdot|\theta) \mid \theta \in \Theta\}$ is LAN. Then*

$$\sup_{I\in F(\mathbb{R}^k)} \liminf_{n\to\infty} \max_{h\in I} \mathrm{E}_{\theta+h/\sqrt{n}}\left[\log\frac{p_n(y|\theta+h/\sqrt{n})}{q_n(y|x)}\right] \;\geq\; \mathrm{E}_0\left[\log\frac{p(\eta|0)}{q^{\mathrm{B}}(\eta|\xi)}\right], \tag{3.23}$$

where $F(\mathbb{R}^k)$ denotes all finite subsets of $\mathbb{R}^k$.

Proof. The proof is similar to the local asymptotic minimax theorem for estimation (van der Vaart, 1998). We recall that E_θ (resp. E_h) denotes expectation not only with respect to x (resp. ξ) but also with respect to y (resp. η). It is sufficient to prove that, for any subsequence of $\{n\}$, there exists a further subsequence such that (3.23) holds along the sequence. We can assume that the distributions of $\log(p_n(y|\theta)/q_n(y|x))$ are tight under θ along such a sequence. Otherwise, the left hand side of (3.23) is infinity. Prohorov's lemma implies that, for any subsequence, there exist a further subsequence (denoting it by n for simplicity) such that

$$\left(\frac{p_n(y|\theta)}{q_n(y|x)}, \log\frac{p_n(y|\theta+h/\sqrt{n})}{p_n(y|\theta)}\right) \tag{3.24}$$

converges in distribution to a limit under θ. By Le Cam's third lemma, $p_n(y|\theta)/q_n(y|x)$ converges in distribution also under every $\theta + h/\sqrt{n}$. Therefore Lemma 3.19 implies that there exist a (randomized) predictive density $q(\eta|\xi)$ such that

$$\frac{p_n(y|\theta)}{q_n(y|x)} \overset{\theta+h/\sqrt{n}}{\rightsquigarrow} \frac{p(\eta|0)}{q(\eta|\xi)}$$

for any h. From Lemma 3.18, it is sufficient to prove that

$$\sup_{I\in F(\mathbb{R}^k)} \liminf_{n\to\infty} \sup_{h\in I} \mathrm{E}_{\theta+h/\sqrt{n}}\left[\log\frac{p_n(y|\theta+h/\sqrt{n})}{q_n(y|x)}\right] \geq \sup_{h\in\mathbb{R}^k} \mathrm{E}_h\left[\log\frac{p(\eta|h)}{q(\eta|\xi)}\right].$$

Let $h_0 \in \mathbb{Q}^k$ and $F(\mathbb{Q}^k)$ be all finite subsets of $\mathbb{Q}^k$. Take a sequence $I_j \in F(\mathbb{Q}^k)$ such that $I_1 \subset I_2 \subset \cdots$ and $\bigcup_j I_j = F(\mathbb{Q}^k)$. The following evaluation holds:

$$\begin{aligned}
&\sup_{I\in F(\mathbb{R}^k)} \liminf_{n\to\infty} \sup_{h\in I} \mathrm{E}_{\theta+h/\sqrt{n}}\left[\log\frac{p_n(y|\theta+h/\sqrt{n})}{q_n(y|x)}\right]\\
\geq\ & \sup_{I\in F(\mathbb{Q}^k)} \liminf_{n\to\infty} \sup_{h\in I} \mathrm{E}_{\theta+h/\sqrt{n}}\left[\log\frac{p_n(y|\theta+h/\sqrt{n})}{q_n(y|x)}\right]\\
=\ & \lim_{j\to\infty} \liminf_{n\to\infty} \sup_{h\in I_j} \mathrm{E}_{\theta+h/\sqrt{n}}\left[\log\frac{p_n(y|\theta+h/\sqrt{n})}{q_n(y|x)}\right]\\
=\ & \lim_{j\to\infty} \sup_{h\in I_j} \mathrm{E}_{\theta+h/\sqrt{n_j}}\left[\log\frac{p_{n_j}(y|\theta+h/\sqrt{n_j})}{q_{n_j}(y|x)}\right] &(3.25)\\
\geq\ & \liminf_{j\to\infty} \mathrm{E}_{\theta+h_0/\sqrt{n_j}}\left[\log\frac{p_{n_j}(y|\theta+h_0/\sqrt{n_j})}{q_{n_j}(y|x)}\right]\\
\geq\ & \mathrm{E}_{h_0}\left[\log\frac{p(\eta|h_0)}{q(\eta|\xi)}\right], &(3.26)
\end{aligned}$$

where the equality in (3.25) uses some subsequence $\{n_j\}$ by the diagonal argument and the last inequality (3.26) comes from the Portmanteau lemma for non-negative continuous functions (see the proof of Theorem 3.21). Since $h_0 \in \mathbb{Q}^k$ is arbitrary, we obtain

$$\sup_{I \in F(\mathbb{R}^k)} \liminf_{n\to\infty} \sup_{h\in I} \mathrm{E}_{\theta+h/\sqrt{n}}\left[\log \frac{p_n(y|\theta+h/\sqrt{n})}{q_n(y|x)}\right] \geq \sup_{h\in\mathbb{Q}^k} \mathrm{E}_h\left[\log\frac{p(\eta|h)}{q(\eta|\xi)}\right]. \quad (3.27)$$

The right hand side is equal to the supremum over $\mathbb{R}^k$ by Fatou's lemma. □

The following theorem states a sufficient condition that the equality in Theorem 3.22 holds. It is expected that the Bayesian predictive density $q_n^\pi(y|x)$ with a smooth prior π is asymptotically optimal. However, it is not case in general. More reasonable conditions may be constructed but we do not pursue this problem. A related topic is treated in Chapter III of Ibragimov & Has'minskii (1981).

Theorem 3.23. *Let $q_n(y|x)$ be a sequence of predictive densities. Assume that the sequence $\log p_n(y|\theta)/q_n(y|x)$ converges to $\log p(\eta|0)/q(\eta|\xi)$ in distribution $P_{\theta,n}$. Assume also that the set of random variables $\log p_n(y|\theta+h/\sqrt{n})/q_n(y|x)$ under $P_{\theta+h/\sqrt{n},n}$ are uniformly integrable in all n for each h. Then*

$$\sup_{I \in F(\mathbb{R}^k)} \liminf_{n\to\infty} \max_{h\in I} \mathrm{E}_{\theta+h/\sqrt{n}}\left[\log \frac{p_n(y|\theta+h/\sqrt{n})}{q_n(y|x)}\right] = \sup_{h\in\mathbb{R}^k} \mathrm{E}_h\left[\log\frac{p(\eta|h)}{q(\eta|\xi)}\right]. \quad (3.28)$$

In particular, if $q(\eta|\xi) = q^{\mathrm{B}}(\eta|\xi)$, the sequence $q_n(y|x)$ is optimal in the sense of the local asymptotic minimax property.

Proof. Put $a_h = \mathrm{E}_h[\log p(\eta|h)/q(\eta|\xi)]$ and $a_{n,h} = \mathrm{E}_{\theta+h/\sqrt{n}}[\log p_n(y|\theta+h/\sqrt{n})/q_n(y|x)]$. By the uniform integrability, $\lim_n a_{n,h} = a_h$ for any h follows. For $I \in F(\mathbb{R}^k)$, we obtain $\lim_n \max_{h\in I} a_{n,h} = \max_{h\in I} \lim_n a_{n,h} = \max_{h\in I} a_h$. Taking supremum over $I \in F(\mathbb{R}^k)$, we obtain the result. □

Chapter 4

Information criteria

In this chapter, the theory of information criteria are summarized. The asymptotic decision theory for model selection is touched on.

4.1 General definition

An important problem in data analysis is to select one from several proposed models based on data. Information criteria provide a tool for selecting a suitable model. We refer Burnham & Anderson (2002) for a review on the model selection theory.

Let $\mathcal{X}$ be a measurable space. Let $\mathcal{P} = \{p(\cdot|\theta) \mid \theta \in \Theta\}$ be a statistical model on $\mathcal{X}$, that is, a set of probability densities with respect to a fixed measure μ on $\mathcal{X}$. Suppose that Θ is a k-dimensional manifold. In most cases, Θ is an open subset of $\mathbb{R}^k$. We consider subfamilies indexed by a finite set A. Specifically, Let $\mathcal{P}_\alpha = \{p(\cdot|\theta) \mid \theta \in \Theta_\alpha\}$ for $\alpha \in A$, where Θ_α is a k_α-dimensional submanifold of Θ. The general definition of information criteria is as follows.

Definition 4.1 (Information criterion). An information criterion is a function $s : A \times \mathcal{X} \mapsto \mathbb{R}$. For given data $x \in \mathcal{X}$, the selected model based on s is $\hat{\alpha} = \hat{\alpha}(s)$ that minimizes $s(\alpha, x)$. □

Almost information criteria are defined as an asymptotically unbiased estimator for the risk of the predictive density. Let $q_\alpha(y|x)$ be a predictive density when the submodel Θ_α is assumed. The risk we use is the expected Kullback-Leibler divergence:

$$r_\theta(q_\alpha) \quad = \quad \iint p(y|\theta)p(x|\theta) \log \frac{p(y|\theta)}{q_\alpha(y|x)} \, \mathrm{d}y \, \mathrm{d}x. \tag{4.1}$$

This is equivalent to $-2 \iint p(y|\theta)p(x|\theta) \log q_\alpha(y|x) \, \mathrm{d}y \, \mathrm{d}x$. Then an information criterion has a form $-2 \log q_\alpha(x|x) + 2b_\alpha(x)$ with a bias-corrected term $b_\alpha(x)$.

4.2 Several criteria

We list several information criteria. AIC and PIC defined below are related to our criterion for LAMN models discussed in Chapter 7. The criteria except AIC and PIC are not used in the other chapters.

Definition 4.2 (AIC, Akaike (1974)). Akaike's Information Criterion (AIC) is defined by

$$\mathrm{AIC}(\alpha) \;=\; -2\log p(x|\hat{\theta}_\alpha) + 2k_\alpha, \tag{4.2}$$

where $\hat{\theta}_\alpha$ is the maximum likelihood estimator in Θ_α. □

Definition 4.3 (PIC, Kitagawa (1997)). Let $\pi_\alpha(\theta)$ be a prior distribution that ranges over Θ_α for each $\alpha \in A$. Then PIC is defined by

$$\mathrm{PIC}(\alpha) \;=\; -2\log q^{\pi_\alpha}(x|x) + k_\alpha, \tag{4.3}$$

where $q^{\pi_\alpha}(y|x)$ is the Bayesian predictive density under the prior $\pi_\alpha(\theta)$ and k_α is the dimension of Θ_α. □

Remark 4.4. Kitagawa's original paper (Kitagawa, 1997) deals with the criterion for more general prior under the linear model. Specifically, it assumes that a Gaussian Bayes model of $x|\theta \sim N(A\theta, R)$ and $\theta \sim N(\theta_0, Q)$. Then two types of PIC are defined by

$$\mathrm{PIC}_1(\alpha) \;=\; -2\log q^{\pi}(x|x) + 2\mathrm{tr}[R^{-1}(W+R)(2W+R)^{-1}W], \tag{4.4}$$

$$\mathrm{PIC}_2(\alpha) \;=\; -2\log q^{\pi}(x|x) + 2\mathrm{tr}[(2W+R)^{-1}W], \tag{4.5}$$

where $W = AQA'$. The difference of the bias term comes from the difference of the true distribution assumed. The first assumes that the marginal distribution $p(x) = \int p(x|\theta)\pi(\theta)\mathrm{d}\theta$ is true and the second assumes that the conditional distribution $p(x|\theta)$ with some θ is true. Our definition above corresponds to PIC_2 under $Q \to \infty$. In addition, PIC is also found in (Akaike, 1980, eq. (3.8)). □

In the following, we restrict the sample space to $\mathcal{X} = E^n$ where $E = \mathbb{R}^d$ with some fixed d. Assume that $x = \{x_1, \cdots, x_n\} \in E^n$ is an i.i.d. sequence. The model is the set $\{f(x_1|\theta) \mid \theta \in \Theta\}$ of probability densities on E. Let $\hat{G}$ be the the empirical distribution $n^{-1}\sum_{i=1}^n \delta_{x_i}(\cdot)$, where δ_z is the Dirac measure concentrated at z. Let $\mathcal{G}$ be the space of all distributions on E and F_θ be the distribution corresponding to θ. Let $T_\alpha : \mathcal{G} \to \Theta_\alpha$ be a Fisher consistent estimator of θ, that is, $T_\alpha(F_\theta) = \theta$ if $\theta \in \Theta_\alpha$. We put $\hat{\theta}_\alpha(x) = T_\alpha(\hat{G})$. The influence function $T_\alpha^{(1)}(z; G)$ for $z \in E$ and $G \in \mathcal{G}$ is defined by

$$T_\alpha^{(1)}(z; G) \;=\; \lim_{\epsilon \to 0} \frac{T_\alpha((1-\epsilon)G + \epsilon\delta_z)}{\epsilon}. \tag{4.6}$$

Definition 4.5 (GIC, Konishi & Kitagawa (1996)). The Generalized information criterion (GIC) is defined by

$$\mathrm{GIC}(\alpha) \;=\; -2\sum_{i=1}^{n}\log f(x_i|T_\alpha(\hat{G}))+2b_\alpha(\hat{G}), \tag{4.7}$$

where

$$b_\alpha(G) \;=\; \mathrm{tr}\left[\int T_\alpha^{(1)}(z;G)\frac{\partial}{\partial\theta'_\alpha}\log f(z|T_\alpha(G))\mathrm{d}G(z)\right]. \tag{4.8}$$

□

Definition 4.6 (TIC, Takeuchi (1976)). When T_α is the maximum likelihood estimator $\hat{\theta}_\alpha$, GIC is called Takeuchi's Information Criterion (TIC). □

Definition 4.7 (EIC, Ishiguro et al. (1997)). EIC is defined by

$$\mathrm{EIC}(\alpha) \;=\; -2\sum_{i=1}^{n}\log f(x_i|\hat{\theta}_\alpha(x))+2b_\alpha(\hat{G}), \tag{4.9}$$

where

$$b_\alpha(G) \;=\; B^{-1}\sum_{b=1}^{B}\sum_{i=1}^{n}\left[\log f(x^*_{b,i}|\hat{\theta}_\alpha(x^*_b))-\log f(x^*_{b,i}|\hat{\theta}_\alpha(x))\right] \tag{4.10}$$

and $x^*_b=\{x^*_{b,1},\cdots,x^*_{b,n}\}$ for $b=1,\cdots,B$ are bootstrap samples (Efron, 1979) that are sampled subject to G. □

Definition 4.8 (Cross validation, e.g. Stone (1977)). Cross validation is defined by

$$\mathrm{CV}(\alpha) \;=\; -2\sum_{i=1}^{n}\log f(x_i|\hat{\theta}_\alpha(x_{-i})), \tag{4.11}$$

where $x_{-i}=\{x_j\}_{j\neq i}$ and $\hat{\theta}_\alpha(x_{-i})$ is the maximum likelihood estimator based on data x_{-i}. □

Definition 4.9 (BIC, Schwarz (1978)). BIC is defined by

$$\mathrm{BIC}(\alpha) \;=\; -2\sum_{i=1}^{n}\log f(x_i|\hat{\theta}_\alpha)+k_\alpha\log n. \tag{4.12}$$

□

Remark 4.10. BIC has larger risk than AIC in the local uniform sense as shown in Section 4.3. This contrasts with BIC's pointwise consistent property. This phenomenon is similar to that of Hodges' superefficient estimator (van der Vaart, 1998, p.109). □

We summarize the properties about asymptotic unbiasedness into the following theorem. We simply say that the submodel $\mathcal{P}_\alpha$ is true instead of saying that $\mathcal{P}_\alpha$ includes the true density. The regularity conditions and proofs are omitted.

Theorem 4.11. *AIC is an asymptotically unbiased estimator for the risk of the plug-in predictive density $p_n(y|\hat{\theta}_\alpha(x))$ if $\mathcal{P}_\alpha$ is true. PIC is an asymptotically unbiased estimator for the risk of the Bayesian predictive density $\int_{\Theta_\alpha} p_n(y|\theta)\pi_\alpha(\theta|x)\mathrm{d}\theta$ if $\mathcal{P}_\alpha$ is true. GIC, TIC and EIC are asymptotically unbiased estimators for the risk of the plug-in predictive density without supposing that $\mathcal{P}_\alpha$ is true.* □

Remark 4.12. As considered in Chapter 7, the asymptotic unbiasedness of AIC and PIC holds also under the local alternative hypothesis, that is, the hypothesis that the true density is $p_n(\cdot|\theta + h/\sqrt{n})$ for $\theta \in \Theta_\alpha$ and $h \in \mathbb{R}^k$. □

4.3 Local asymptotic maximum risk

We fix a LAN model $\{p_n(x|\theta) \mid \theta \in \Theta\}$ and submodels $\{\Theta_\alpha \mid \alpha \in A\}$. Let $\{p(\xi|h) \mid h \in \mathbb{R}^k\}$ be the limit model. Assume that the limit $H_\alpha \subset \mathbb{R}^k$ of the submodel Θ_α exists (see Chapter 7 for the exact meaning). Consider the model-selection problem for the non-limit model and the limit model.

We prepare some notations for the non-limit models. We use the expected Kullback-Leilber divergence

$$r_\theta(q_n) = \int p_n(x|\theta) \int p_n(y|\theta) \log \frac{p_n(y|\theta)}{q_n(y|x)} \mathrm{d}x\mathrm{d}y, \qquad (4.13)$$

as the risk of prediction. The risk of an information criterion s is defined by $r_\theta(q_{s,n})$, where $q_{s,n}$ is the plug-in predictive density corresponding to the selected model by s:

$$q_{s,n}(y|x) = \sum_{\alpha \in A} p_n(y|\hat{\theta}_\alpha)\mathbb{I}_{(\hat{\alpha}(s)=\alpha)}.$$

The symbols $r_h(q)$ and q_s for the limit models are similarly defined.

Shibata (1986) studied the minimax risk and minimax regret of generalized FPE. Although Stone (1982) studied minimax property of AIC from the viewpoint of LAN setting, the risk he adopted is sum of the Kullback-Leibler divergence and a penalty term.

The following simple example elucidates a weak point of BIC. The result is essentially stated in Shibata (1986).

Example 4.13. Assume that $X_1, \cdots, X_n \sim N(\theta, 1)$ i.i.d. with $\theta \in \Theta = \mathbb{R}$. Consider two models $\Theta_{\mathrm{I}} = \{0\}$ and $\Theta_{\mathrm{II}} = \mathbb{R}$. Thus $A = \{\mathrm{I}, \mathrm{II}\}$. Then, the asymptotic maximum risk of AIC and BIC is, respectively,

$$\limsup_{n\to\infty} \sup_{\theta\in\mathbb{R}} r_\theta(q_{\mathrm{AIC},n}) = \sup_{h\in\mathbb{R}} r_h(q_{\mathrm{AIC}}) < \infty, \tag{4.14}$$

$$\limsup_{n\to\infty} \sup_{\theta\in\mathbb{R}} r_\theta(q_{\mathrm{BIC},n}) = \infty. \tag{4.15}$$

The risk of each criterion is shown in Fig. 4.1 for $n = 10$ and $n = 100$.

Proof. The equality in (4.14) is easily checked. Let us consider a generalized criterion $s(\alpha) = -2\log p_n(x|\hat{\theta}_\alpha) + c_n k_\alpha$, where c_n is a sequence of positive numbers. AIC and BIC correspond to $c_n = 2$ and $c_n = \log n$, respectively. We show that the asymptotic maximum risk is finite if and only if $\{c_n\}$ is bounded. The maximum likelihood estimator under each model is $\hat{\theta}_{\mathrm{I}} = 0$ and $\hat{\theta}_{\mathrm{II}} = \bar{x} = \sum x_i/n$, respectively. The information criteria are $s(\mathrm{I}) = \sum x_i^2$ and $s(\mathrm{II}) = \sum x_i^2 - n\bar{x}^2 + c_n$ (common constants are neglected). The selected predictive distribution satisfies

$$-2\log q_{s,n}(y|x) = \{\textstyle\sum y_i^2\}\mathbb{I}_{(n\bar{x}^2\le c_n)} + \{\textstyle\sum (y_i - \bar{x})^2\}\mathbb{I}_{(n\bar{x}^2 > c_n)}.$$

Twice the Kullback-Leibler divergence is

$$2\int p_n(y|\theta)\log\frac{p_n(y|\theta)}{q_{s,n}(y|x)}\mathrm{d}y = n\theta^2\mathbb{I}_{(n\bar{x}^2\le c_n)} + n(\bar{x}-\theta)^2\mathbb{I}_{(n\bar{x}^2>c_n)}. \tag{4.16}$$

Twice the risk is

$$\begin{aligned} 2r_\theta(q_{s,n}) &= n\theta^2\int_{n\bar{x}^2\le c_n} p_n(x|\theta)\mathrm{d}x + \int_{n\bar{x}^2>c_n} n(\bar{x}-\theta)^2 p_n(x|\theta)\mathrm{d}x \\ &= n\theta^2\int_{(\sqrt{n}\theta+z)^2\le c_n}\phi(z)\mathrm{d}z + \int_{(\sqrt{n}\theta+z)^2>c_n} z^2\phi(z)\mathrm{d}z, \end{aligned} \tag{4.17}$$

where $\phi(z)$ is the probability density function of $N(0,1)$. The second term of (4.17) is [0,1]-valued and therefore bounded. The supremum of the first term over $\theta \in \mathbb{R}$ is

$$\sup_{\theta\in\mathbb{R}} n\theta^2\int_{(\sqrt{n}\theta+z)^2\le c_n}\phi(z)\mathrm{d}z = \sup_{h\in\mathbb{R}} h^2\int_{(h+z)^2\le c_n}\phi(z)\mathrm{d}z =: \sup_{h\in\mathbb{R}}\rho_n(h). \tag{4.18}$$

This is finite if and only if $\{c_n\}$ is bounded. In fact, if $c_n \to \infty$, then $\rho_n(-\sqrt{c_n}) = c_n\int_0^{2\sqrt{c_n}}\phi(z)\mathrm{d}z \to \infty$. On the other hand, if $c_n \le c$, then

$$\lim_{h\to\infty}\rho_n(h) \le \lim_{h\to\infty} h^2\frac{2\sqrt{c}\,\mathrm{e}^{-(h-\sqrt{c})^2/2}}{\sqrt{2\pi}} = 0$$

and $\rho_n(h) \le h^2$ imply the claim. □

The above example suggests a more general result. AIC is naturally defined for the limit model by $\mathrm{AIC}(\alpha) = -2\log p(\xi|\pi_\alpha\xi) + 2k_\alpha$, where π_α is the projection to H_α with respect to the metric J. BIC for the limit model is $\mathrm{BIC}(\alpha) = -2\log p(\xi|0)$ if $H_\alpha = \{0\}$ and $\mathrm{BIC}(\alpha) = \infty$ if $H_\alpha \neq \{0\}$. Then, under some mild conditions, AIC and BIC for the non-limit model converges to those for the limit model.

Theorem 4.14 (Local asymptotic minimax theorem for information criterion). *Let $\{p_n(\cdot|\theta) \mid \theta \in \Theta\}$ be a LAN model and s be an information criterion. Take ξ_n as* [LA2] *in Chapter 3. Assume that $\hat{\theta}_\alpha - (\theta + \pi_\alpha\xi_n/\sqrt{n})$ converges to 0 for each $\alpha \in A$. Assume also that the selected model $\hat{\alpha}_n(s)$ converges in distribution to $\hat{\alpha}(s)$. Then*

$$\sup_{I\in F(\mathbb{R}^k)} \liminf_{n\to\infty} \sup_{h\in I} r_{\theta+h/\sqrt{n}}(q_{s,n}) \quad\geq\quad \sup_{h\in\mathbb{R}^k} r_h(q_s),$$

where $F(\mathbb{R}^k)$ denotes all finite subsets of $\mathbb{R}^k$.

Proof. From the assumption, $p(y|\hat{\theta}_\alpha)/p(y|\theta)$ converges in distribution to $p(\eta|\pi_\alpha\xi)/p(\eta|h)$. Thus, for any subsequence, there exists a further subsequence such that $q_{s,n}(y|x)/p(y|\theta)$ converges to $q_s(\eta|\xi)/p(\eta|h)$. Then the proof is same as Theorem 3.22 if we replace q_n and q with $q_{s,n}$ and q_s, respectively. □

We define the regret RG of an information criterion s by

$$RG_\theta(q_{s,n}) \quad=\quad r_\theta(q_{s,n}) - r_\theta(q_{\mathrm{OPT},n}),$$

where the distribution $q_{\mathrm{OPT},n}$ is the predictive distribution corresponding to the submodel attaining the minimum loss, that is,

$$q_{\mathrm{OPT},n}(y|x) \quad=\quad \sum_{\alpha\in A} \mathbb{I}_{\{\alpha = \operatorname*{argmin}_{\alpha'} l_\theta(q_{\alpha',n}(\cdot|x))\}} q_{\alpha,n}(y|x). \qquad (4.19)$$

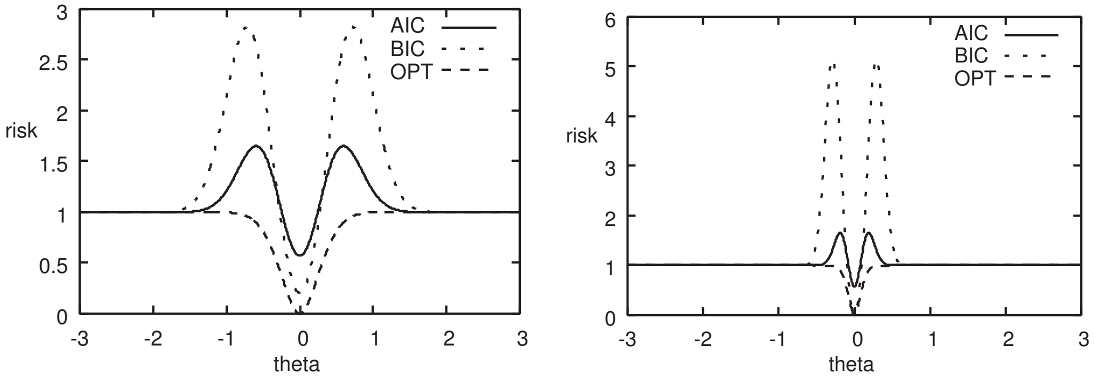
The $q_{\mathrm{OPT},n}$ depends on the true θ and provides the lower bound for risk of model selection. The above definition of the regret is slightly different from that defined by Shibata (1986). He considered the difference $r_\theta(q_{s,n}) - r_\theta(q_{\alpha_0,n})$, where α_0 is the minimal submodel including the true density. His definition is well-defined only under a hypothesis that the submodels are nested.

The local asymptotic minimax theorem for the regret holds. The proof is similar to Theorem 4.14.

Theorem 4.15 (Local asymptotic minimax theorem for regret). *Under the same conditions as Theorem 4.14, we have*

$$\sup_{I\in F(\mathbb{R}^k)} \liminf_{n\to\infty} \sup_{h\in I} RG_{\theta+h/\sqrt{n}}(q_{s,n}) \quad\geq\quad \sup_{h\in\mathbb{R}^k} \left\{ r_h(q_s) - \limsup_{n\to\infty} r_{\theta+h/\sqrt{n}}(q_{\mathrm{OPT},n}) \right\}.$$

If the set of random variables $\log p_n(y|\theta+h/\sqrt{n})/q_{\mathrm{OPT},n}(y|x)$ *under* $P_{\theta+h/\sqrt{n}}$ *are uniformly integrable in all* n *for each* h*, the right hand side is* $\sup_h RG_h(q_s)$*.* □



(a) $n = 10$

(b) $n = 100$

Figure 4.1: The risk of AIC, BIC and OPT in Example 4.13. OPT is the lower bound obtained by a model-selection procedure that selects the model attaining minimum loss by using true θ.

Chapter 5

Estimation of fractal index

In this chapter, the asymptotic properties of the quasi maximum likelihood estimator (QMLE) for the fractal index proposed by Stein (1995) are studied under the fixed domain asymptotics. The asymptotic variance is smaller than the estimators considered in Chapter 2.

5.1 Introduction

Let us consider a stationary Gaussian-process $\{X_t \in \mathbb{R} \mid t \in [0,1]\}$ with the constant mean $\mu \in \mathbb{R}$ and variogram

$$\gamma(t) \;=\; \frac{1}{2}\mathrm{E}[(X_{s+t} - X_s)^2]. \tag{5.1}$$

We assume that $\gamma(t)$ varies regularly at 0 with exponent $\alpha \in (0,2)$, that is,

$$\lim_{t\to 0}\frac{\gamma(tu)}{\gamma(t)} \;=\; u^{\alpha} \tag{5.2}$$

for each $u > 0$ (Feller, 1971, p.276). The parameter α is called the fractal index (Chapter 2). Details of further assumptions on γ are given in the next section. Our goal is to find a good estimator of α from the discrete observations $\{X_{i/n} \mid i = 0, 1, \cdots, n\}$. The other parameters contained in $\gamma(\cdot)$ are considered to be (infinite-dimensional) nuisance parameters. As stated in Chapter 2, a number of estimators for α have been studied in recent years.

We study the asymptotic property of the quasi maximum likelihood estimator (QMLE) proposed by Stein (1995) under the fixed domain asymptotics (see Chapter 2). QMLE is defined by (5.5) below. We first describe the background of the estimator. The increments of $X_{i/n}$ are denoted by $x_i = X_{i/n} - X_{(i-1)/n}$. Let $x = \{x_i \mid i = 1, \cdots, n\}$. By the assumption

(5.2), the autocovariance of x satisfies

$$\begin{aligned}\mathrm{E}[x_{i+h}x_i] &= \gamma((h+1)/n)+\gamma((h-1)/n)-2\gamma(h/n) \\ &\sim \gamma(1/n)\{|h+1|^\alpha+|h-1|^\alpha-2|h|^\alpha\}\end{aligned}$$

as $n \to \infty$ for each fixed h. As remarked by Kent & Wood (1997), the process x is approximated by the *fractional Gaussian noise* (FGN). FGN is defined by a discrete-time stationary Gaussian process $y = \{y_i \mid i = 1, \cdots, n\}$ with the covariance function $\mathrm{E}[y_{i+h}y_i] = (A/2)\{|h+1|^\alpha+|h-1|^\alpha-2|h|^\alpha\}$ with some $A > 0$. Statistical inference for FGN was reviewed in Section 2.5. Since the normalized process $\gamma(1/n)^{-1/2}x_i$ is approximated by FGN, it is natural to use an estimator for FGN to estimate the parameter of the process X_t. Although the normalization constant $\gamma(1/n)^{-1/2}$ depends on α, it causes no problem by considering $\gamma(1/n)$ as a new nuisance parameter since the coordinates of the nuisance parameter can be transformed by any function including the parameter of interest (Amari, 1985, Chapter 8).

We define QMLE analogous to that for FGN. The spectral density of FGN is denoted by

$$f(\lambda|\alpha) = 2As(\alpha)\sin^2(\lambda/2)\sum_{j=-\infty}^{\infty}|\lambda+2\pi j|^{-\alpha-1}, \tag{5.3}$$

where $s(\alpha) = \Gamma(\alpha+1)\sin(\alpha\pi/2)/\pi$. The constant A is determined by, just for convenience,

$$\int_{-\pi}^{\pi}\log f(\lambda|\alpha)\mathrm{d}\lambda \;=\; 0. \tag{5.4}$$

Let $I_n(\lambda; x) = (2\pi n)^{-1}|\sum_{j=1}^n x_j \mathrm{e}^{-\mathrm{i}j\lambda}|^2$ be the periodogram of the increments x_i. Then QMLE is defined by

$$\hat{\alpha}_n \;=\; \underset{\alpha'}{\mathrm{argmin}}\int_{-\pi}^{\pi}\frac{I_n(\lambda; x)}{f(\lambda|\alpha')}\mathrm{d}\lambda. \tag{5.5}$$

In practice, one uses $\sum_{k=1}^{n-1} I_n(\lambda_k; x)/f(\lambda_k)$ instead of the integral in (5.5), where $\lambda_k = 2\pi k/n$. Stein showed that $\hat{\alpha}_n$ behaves well by numerical experiments. On the other hand, if x is exactly FGN, the consistency and asymptotic normality hold due to Fox & Taqqu (1986). We prove the consistency and asymptotic normality under mild conditions.

We assume that the true parameter α belongs to $\mathcal{A} := (1 + a, 2 - a)$ with some *known* $a > 0$. Thus the solution of the optimization problem (5.5) is searched over $\bar{\mathcal{A}} = [1 + a, 2 - a]$. Although this is an unrealistic assumption, we assume it from a technical reason.

This chapter is organized as follows. In Section 5.2 we describe the theorems about the consistency and the asymptotic normality of QMLE. The regularity conditions are also stated there. Numerical experiments are given in Section 5.3. Proofs are given in Section 5.4. Lastly we give some discussions in Section 5.5.

5.2 Main results

5.2.1 Consistency and asymptotic normality

We first give a regularity condition [FS] that ensures the consistency and asymptotic normality of $\hat{\alpha}_n$ (the letters "FS" denote the "fractal" and "spectrum"). Other conditions on the variogram sufficient to prove the consistency and asymptotic normality are given in Subsection 5.2.2. As stated in the introduction, we assume that the parameter space is $\bar{\mathcal{A}} = [1+a, 2-a]$ with some known $a > 0$ and the true α belongs to its interior.

Assume that there exists the spectral density $\phi(x)$ of $\{X_t \mid t \in [0,1]\}$, that is, the autocovariance function of X_t is given by $\sigma(t) = \int_{-\infty}^{\infty} \phi(x) \mathrm{e}^{\mathrm{i}xt} \mathrm{d}x$. The regularity condition is stated as follows. The coefficient $s(\alpha) = \Gamma(\alpha+1)\sin(\alpha\pi/2)/\pi$ in (5.6) is just for the sake of convenience.

[FS] The function ϕ is bounded over $\mathbb{R}$, and there exist $\nu > 0$ and $\beta > 0$ such that

$$\phi(x) \;\; = \;\; \nu s(\alpha)|x|^{-\alpha-1} + \mathrm{O}(|x|^{-\alpha-\beta-1}) \tag{5.6}$$

as $|x| \to \infty$. It is assumed that $\beta \le 2-\alpha$ without loss of generality.

We prepare some notations. Assume the condition [FS]. We consider the normalized increments $\tilde{x}_i = (n^\alpha A/2\nu)^{1/2} x_i$, where A is the same as one in (5.3). The spectral density $f_n(\lambda)$ for $\tilde{x}_i$ is expressed as

$$f_n(\lambda) \;\; = \;\; \frac{2A}{\nu}\sin^2(\lambda/2)\sum_{j=-\infty}^{\infty} n^{1+\alpha}\phi(n(\lambda+2\pi j)) \tag{5.7}$$

(see Stein (1995)). We show that $|f_n(\lambda) - f(\lambda|\alpha)| = \mathrm{O}(n^{-\beta})$ for each $\lambda \neq 0$ in Subsection 5.4.1. The periodogram of $\tilde{x}_i$ is denoted by $\tilde{I}_n(\lambda) = I_n(\lambda;\tilde{x}) = (n^\alpha A/2\nu) I_n(\lambda; x)$. QMLE $\hat{\alpha}_n$ is invariant under replacing I_n in (5.5) with $\tilde{I}_n$. The expectation of $\tilde{I}_n(\lambda)$ is given by

$$S_n[f_n](\lambda) \;\; = \;\; \frac{1}{2\pi n}\int_{-\pi}^{\pi} f_n(\lambda')\frac{\sin^2(n(\lambda'-\lambda)/2)}{\sin^2((\lambda'-\lambda)/2)}\mathrm{d}\lambda'. \tag{5.8}$$

Let α_n be the parameter corresponding to the spectral density closest to $S_n[f_n]$, that is,

$$\alpha_n \;=\; \operatorname*{argmin}_{\alpha'} \int_{-\pi}^{\pi} \frac{S_n[f_n](\lambda)}{f(\lambda|\alpha')} \mathrm{d}\lambda. \tag{5.9}$$

Let ∂_α be the derivative with respect to α. The Fisher information for FGN is

$$J \;=\; \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{(\partial_\alpha f(\lambda|\alpha))^2}{f(\lambda|\alpha)^2} \mathrm{d}\lambda. \tag{5.10}$$

Then the consistency and asymptotic normality is stated as follows.

Theorem 5.1 (consistency). *Assume* [FS]. *Then* $\hat{\alpha}_n \xrightarrow{\mathrm{P}} \alpha$ *as* $n \to \infty$. □

Theorem 5.2 (asymptotic normality). *Assume* [FS]. *Then*

$$\sqrt{n}(\hat{\alpha}_n - \alpha_n) \;\rightsquigarrow\; N(0, J^{-1}) \tag{5.11}$$

and

$$\alpha_n - \alpha \;=\; \frac{1}{4\pi J} \int_{-\pi}^{\pi} \frac{(\partial_\alpha f(\lambda|\alpha)) f_n(\lambda)}{f(\lambda|\alpha)^2} \mathrm{d}\lambda + \mathrm{o}(n^{-\beta}) \;=\; \mathrm{O}(n^{-\beta}). \tag{5.12}$$

□

Remark 5.3. If the number of observation is restricted to $n = 2^m$, we can also prove the strong consistency of $\hat{\alpha}_n$ under $m \to \infty$ by using the Borel-Cantelli lemma. □

Remark 5.4. The asymptotic variance J^{-1} is optimal in the following sense: if x_i is exactly FGN, it is optimal due to Dahlhaus (1989). The order $\mathrm{O}(n^{-\beta})$ of α_n is the same as the other estimators discussed in Chapter 2. It is larger than the order of the stochastic term $\hat{\alpha}_n - \alpha_n$ if $\beta < 1/2$. □

Remark 5.5. QMLE $\hat{\alpha}_n$ and the quantity α_n are interpreted as the orthogonal projection of $\tilde{I}_n$ and $S_n[f_n]$, respectively, to the FGN model in the space of all spectral densities (Fig. 5.1). □

5.2.2 Other conditions

Let us consider some regularity conditions not on the spectral density but on the variogram. Since $\gamma(t)$ is an even function, the conditions are stated only for $t \geq 0$. Let $\gamma^{(k)}(t)$ denote the k-th derivative of $\gamma(t)$ and $\alpha_{(k)} = \alpha(\alpha - 1) \cdots (\alpha - k + 1)$.

We first give a set of conditions [F1]–[F4] that ensures the consistency.

[F1] There exists $\delta > 0$ such that $\gamma(t)$ is positive for $t \in (0, \delta)$.

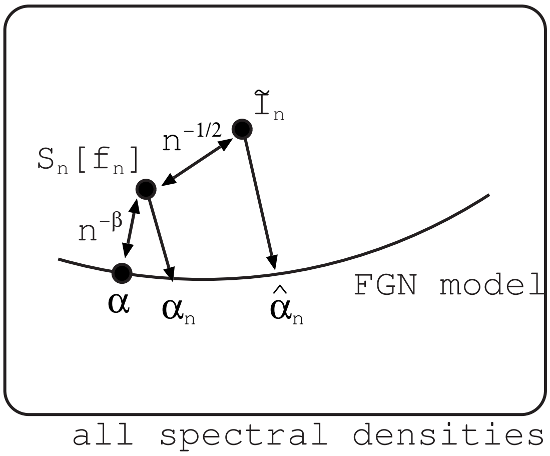


Figure 5.1: Geometrical interpretation of QMLE.

[F2] The function $\gamma(t)$ has regular variation with exponent α:

$$\lim_{t\downarrow 0}\frac{\gamma(tu)}{\gamma(t)} \;=\; u^{\alpha} \tag{5.13}$$

for any $u > 0$.

[F3] The function $\gamma(t)$ is continuously twice differentiable over $t \in (0, 1]$.

[F4] For $k = 1$ and 2, it holds that

$$\lim_{t\downarrow 0}\frac{t^k\gamma^{(k)}(t)}{\gamma(t)} \;=\; \alpha_{(k)}. \tag{5.14}$$

Theorem 5.6. *Assume* [F1]–[F4]. *Then* $\hat{\alpha}_n \xrightarrow{\mathrm{P}} \alpha$ *as* $n \to \infty$. □

We next give a set of conditions [F5]–[F10] that implies [FS]. It is related to the Tauberian theorem given by Pitman (1968). Here we assume that the domain of the variogram function $\gamma(t)$ can be extended to $\mathbb{R}$, that is, there exists a non-negative definite function $\tilde{\gamma}(t)$ on $t \in \mathbb{R}$ such that $\tilde{\gamma}(t) = \gamma(t)$ for $t \in [-1, 1]$. We denote the extended function $\tilde{\gamma}(t)$ by $\gamma(t)$ for simplicity.

[F5] The function $\gamma(t)$ is continuously three-times differentiable over $(0, \infty)$.

[F6] There exist $\nu > 0$ and $\beta \in (0, 2 - \alpha]$ such that

$$\gamma^{(k)}(t) \;=\; \nu\alpha_{(k)}t^{\alpha-k} + \mathrm{O}(t^{\alpha+\beta-k}) \tag{5.15}$$

as $t \downarrow 0$ for $k = 0, 1$ and 2.

[F7] There exists $\gamma(\infty) = \lim_{t\to\infty} \gamma(t)$. Furthermore,

$$\int_0^\infty |\gamma(\infty) - \gamma(t)| \mathrm{d}t \quad < \quad \infty. \tag{5.16}$$

[F8] For $k = 1$ and 2,

$$\lim_{t\to\infty} \gamma^{(k)}(t) \quad = \quad 0. \tag{5.17}$$

[F9] For each $\delta > 0$,

$$\int_\delta^\infty |\gamma^{(3)}(t)| \mathrm{d}t \quad < \quad \infty. \tag{5.18}$$

[F10] There exists $\delta > 0$ such that the function $\gamma^{(2)}(t) - \nu\alpha_{(2)}t^{\alpha-2}$ is monotone over $t \in (0, \delta)$.

Theorem 5.7. *The conditions* [F5]–[F10] *imply* [FS]. □

The following corollary is immediately obtained.

Corollary 5.8. *The conditions* [F5]–[F10] *imply the consistency and asymptotic normality of QMLE.* □

5.3 Numerical experiments

We first see how the spectral density f_n converges to f. Let ϕ be

$$\phi(x) \quad = \quad x_0^{-\alpha-1} \lceil x/x_0 \rceil^{-\alpha-1}, \quad x_0 = 20, \quad \alpha = 1.5.$$

The graphs of the spectral densities f_n ($n = 100$ and 1000) and f are shown in Fig. 5.2.

We next consider the following spectral densities.

$$\begin{aligned} \phi_1(x) &= (1 + x^2/\alpha)^{-(\alpha+1)/2}, \\ \phi_2(x) &= (\lfloor |x| \rfloor \vee 1)^{-\alpha-1}, \\ \phi_3(x) &= (2 + \sin|x|)(1 + x^2/\alpha)^{-(\alpha+1)/2}. \end{aligned}$$

The densities ϕ_1 and ϕ_2 satisfy [FS] but ϕ_3 does not satisfy it. We remark that ϕ_1 is called the Matérn class and the second one is not continuous. We also consider the following variogram functions.

$$\begin{aligned} \gamma_1(t) &= 1 - \exp(-|t|^\alpha), \\ \gamma_2(t) &= 1 - \exp(-|t|^\alpha) r(|t|), \\ \gamma_3(t) &= \frac{|t|^\alpha - |t|^2}{\log(1/|t|)}, \end{aligned}$$

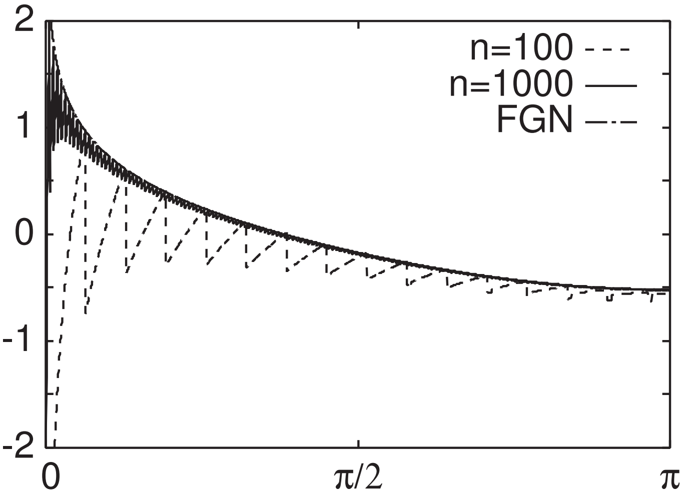


Figure 5.2: Convergence of f_n to f. The graphs of $\log f_n(\lambda)$ ($n = 100$ and 1000) and $\log f(\lambda|\alpha)$ are shown. The fractal index is $\alpha = 1.5$.

where

$$r(t) \;=\; \left(\frac{(1-t)\sin(2\pi t)}{2\pi t} + \frac{1-\cos(2\pi t)}{2\pi^2 t}\right)\mathbb{I}_{(t\leq 1)}.$$

The functions γ_1 and γ_2 satisfy [F5]–[F10]. The function γ_3 does not satisfy [F5]–[F10] but satisfy [F1]–[F4]. We remark that γ_1 is called the stable class and γ_2 has the compact support (Gneiting, 2002).

For the above six examples, we compare QMLE $\hat{\alpha}_{\text{QMLE}}$ with an increment-based estimator $\hat{\alpha}^{(1)}_{\text{OLS}}$ in Kent & Wood (1997) because the latter one is considered as a good estimator in early researches. The results are shown in Table 5.1, Table 5.2 and Table 5.3. From Table 5.3, the order of bias and variance seems to be consistent with our theoretical result: $\mathrm{O}(n^{-\beta})$ and $\mathrm{O}(n^{-1})$, respectively. From the three tables, $\hat{\alpha}_{\text{QMLE}}$ always has smaller variance than $\hat{\alpha}^{(1)}_{\text{OLS}}$. It is consistent with our theoretical result. On the other hand, the bias of $\hat{\alpha}_{\text{QMLE}}$ is sometimes larger than $\hat{\alpha}^{(1)}_{\text{OLS}}$.

We also compare the prediction error of the two estimators if we use the kriging method (see e.g. Chilès & Delfiner (1999)). The method of the experiment is as follows. We generate the sample $\{X_{i/2n}\}_{i=0}^{2n}$ based on the true variogram $\gamma(t)$. Assume that $\{X_{i/n}\}_{i=0}^{n}$ is the observed data and $\{X_{(2i-1)/2n}\}_{i=1}^{n}$ is the unobserved data. Then the estimator $\hat{\alpha}$ is calculated from $\{X_{i/n}\}_{i=0}^{n}$ and the predicted value $\hat{X}_{(2i-1)/2n}$ is calculated based on the kriging method. The prediction error is defined by $n^{-1}\sum_{i=1}^{n}(\hat{X}_{(2i-1)/2n} - X_{(2i-1)/2n})^2$. A result is shown in Table 5.4. From the table, the prediction error of QMLE is slightly less than that of OLS.

Table 5.1: Comparison of QMLE v.s. a increment-based estimator. The true variogram is (a) $\gamma(t) = 2 - \exp(-|t|^{\alpha}) - \exp(-|t|^{1.9})$ and (b) $\gamma(t) = 2 - \exp(-|t|^{\alpha}) - \exp(-|t|^{1.99})$. The nuisance parameter β is (a) $\beta = 1.9 - \alpha$ and (b) $\beta = 1.99 - \alpha$, respectively. Number of simulation is 5000 in each case. The numerical value of the inverse of Fisher information J^{-1} is indicated in the last column.

(a) $\gamma(t) = 2 - \exp(-|t|^{\alpha}) - \exp(-|t|^{1.9})$.

α	β	n	$\hat{\alpha}_{\text{QMLE}}$		$\hat{\alpha}^{(1)}_{\text{OLS}}$		J^{-1}
			$n^{\beta}\times$bias	$n\times$var.	$n^{\beta}\times$bias	$n\times$var.	
1.2	0.7	1024	0.817	1.75	0.028	4.01	1.64
1.2	0.7	2048	0.816	1.67	0.088	3.98	1.64
1.2	0.7	4096	0.865	1.71	0.112	4.05	1.64
1.5	0.4	1024	0.308	1.96	0.105	4.25	1.74
1.5	0.4	2048	0.285	1.88	0.123	4.26	1.74
1.5	0.4	4096	0.279	1.81	0.138	4.26	1.74

(b) $\gamma(t) = 2 - \exp(-|t|^{\alpha}) - \exp(-|t|^{1.99})$.

α	β	n	$\hat{\alpha}_{\text{QMLE}}$		$\hat{\alpha}^{(1)}_{\text{OLS}}$		J^{-1}
			$n^{\beta}\times$bias	$n\times$var.	$n^{\beta}\times$bias	$n\times$var.	
1.2	0.79	1024	0.862	1.77	-0.301	4.00	1.64
1.2	0.79	2048	1.023	1.73	-0.092	4.07	1.64
1.2	0.79	4096	0.876	1.71	-0.122	4.07	1.64
1.8	0.19	1024	0.137	2.98	0.005	4.42	1.81
1.8	0.19	2048	0.105	2.76	0.010	4.26	1.81
1.8	0.19	4096	0.074	2.63	0.007	4.28	1.81
1.9	0.09	1024	0.093	2.20	0.007	4.29	1.82
1.9	0.09	2048	0.079	3.21	0.005	4.36	1.82
1.9	0.09	4096	0.061	3.87	0.009	4.24	1.82

Table 5.2: Comparison of QMLE v.s. an increment-based estimator. Number of the sampling points is $n = 2048$. The spectral density is (a) $\phi(x) = (1 + x^2/\alpha)^{-(\alpha+1)/2}$, (b) $\phi(x) = (\lfloor |x| \rfloor \vee 1)^{-\alpha-1}$ and (c) $\phi(x) = (2 + \sin|x|)(1 + x^2/\alpha)^{-(\alpha+1)/2}$. Number of simulation is 1000 in each case. The numerical value of the inverse of Fisher information J^{-1} is indicated in the last column.

(a) $\phi(x) = (1 + x^2/\alpha)^{-(\alpha+1)/2}$.

α	$\hat{\alpha}_{\text{QMLE}}$		$\hat{\alpha}^{(1)}_{\text{OLS}}$		J^{-1}
	bias	$n\times$var.	bias	$n\times$var.	
0.1	0.1002	0.55	0.1042	1.13	0.26
0.4	0.0300	1.07	0.0303	2.20	0.92
0.7	0.0078	1.39	0.0083	3.03	1.31
1.0	-0.0004	1.52	-0.0004	3.67	1.54
1.3	-0.0029	1.60	-0.0013	4.16	1.68
1.6	-0.0032	1.80	-0.0034	4.22	1.77
1.9	-0.0013	2.01	-0.0015	4.41	1.82

(b) $\phi(x) = (\lfloor |x| \rfloor \vee 1)^{-\alpha-1}$.

α	$\hat{\alpha}_{\text{QMLE}}$		$\hat{\alpha}^{(1)}_{\text{OLS}}$		J^{-1}
	bias	$n\times$var.	bias	$n\times$var.	
0.1	0.0987	0.55	0.1039	1.16	0.26
0.4	0.0315	1.02	0.0326	2.14	0.92
0.7	0.0080	1.37	0.0055	3.05	1.31
1.0	0.0019	1.55	0.0023	3.83	1.54
1.3	-0.0021	1.87	-0.0018	3.93	1.68
1.6	-0.0022	1.82	-0.0028	4.20	1.77
1.9	-0.0013	1.93	-0.0043	4.46	1.82

(c) $\phi(x) = (2 + \sin|x|)(1 + x^2/\alpha)^{-(\alpha+1)/2}$.

α	$\hat{\alpha}_{\text{QMLE}}$		$\hat{\alpha}^{(1)}_{\text{OLS}}$		J^{-1}
	bias	$n\times$var.	bias	$n\times$var.	
0.1	0.0996	0.53	0.1043	1.123	0.26
0.4	0.0312	0.99	0.0317	2.148	0.92
0.7	0.0076	1.31	0.0067	3.046	1.31
1.0	0.0009	1.58	0.0030	3.625	1.54
1.3	-0.0011	1.83	-0.0008	4.061	1.68
1.6	-0.0038	1.82	-0.0026	4.165	1.77
1.9	-0.0053	1.73	-0.0046	4.165	1.82

Table 5.3: Comparison of QMLE v.s. an increment-based estimator. Number of the sampling points is $n = 2048$. The variogram is (a) $\gamma(t) = 1 - \exp(-|t|^\alpha)$, (b) $\gamma(t) = 1 - r(|t|)\exp(-|t|^\alpha)$ and (c) $\gamma(t) = (|t|^\alpha - |t|^2)/\log(1/|t|)$. Number of simulation is 1000 in each case. The numerical value of the inverse of Fisher information J^{-1} is indicated in the last column.

(a) $\gamma(t) = 1 - \exp(-|t|^\alpha)$.

α	$\hat{\alpha}_{\text{QMLE}}$		$\hat{\alpha}^{(1)}_{\text{OLS}}$		J^{-1}
	bias	$n\times$var.	bias	$n\times$var.	
0.1	-0.0193	0.34	-0.0235	0.58	0.26
0.4	-0.0167	0.95	-0.0134	1.95	0.92
0.7	-0.0052	1.32	-0.0029	2.78	1.31
1.0	0.0002	1.66	0.0007	3.60	1.54
1.3	0.0012	1.84	-0.0001	4.35	1.68
1.6	0.0046	1.74	-0.0002	4.21	1.77
1.9	0.0235	2.79	-0.0010	4.10	1.82

(b) $\gamma(t) = 1 - r(|t|)\exp(-|t|^\alpha)$.

α	$\hat{\alpha}_{\text{QMLE}}$		$\hat{\alpha}^{(1)}_{\text{OLS}}$		J^{-1}
	bias	$n\times$var.	bias	$n\times$var.	
0.1	-0.0168	0.32	-0.0241	0.59	0.26
0.4	-0.0126	0.88	-0.0145	1.92	0.92
0.7	0.0017	1.36	-0.0032	3.17	1.31
1.0	0.0104	1.72	-0.0006	3.75	1.54
1.3	0.0241	2.09	-0.0022	4.31	1.68
1.6	0.0554	2.82	0.0040	3.91	1.77
1.9	0.0980	0.06	0.0564	4.43	1.82

(c) $\gamma(t) = (|t|^\alpha - |t|^2)/\log(1/|t|)$.

α	$\hat{\alpha}_{\text{QMLE}}$		$\hat{\alpha}^{(1)}_{\text{OLS}}$		J^{-1}
	bias	$n\times$var.	bias	$n\times$var.	
0.1	0.1716	0.68	0.1480	1.40	0.26
0.4	0.1593	1.32	0.1430	2.56	0.91
0.7	0.1525	1.50	0.1414	3.28	1.31
1.0	0.1387	1.65	0.1320	3.78	1.54
1.3	0.1220	1.83	0.1161	4.38	1.68
1.6	0.0915	1.94	0.0876	4.44	1.77
1.9	0.0327	1.77	0.0266	4.13	1.82

Table 5.4: The prediction error of each estimators. The true variogram is assumed to be $\gamma(t) = 1 - \exp(-|t|^\alpha)$. The number of sampling points is $n = 32$ and the number of simulation is 1000. The last column corresponds to the result of kriging when the true parameter is used.

α	QMLE	OLS	true
1.0	0.0162	0.0168	0.0155
1.3	3.54e-3	3.61e-3	3.43e-3
1.6	6.09e-4	6.12e-4	6.04e-4
1.9	4.49e-5	4.49e-5	4.47e-5

5.4 Proofs

5.4.1 Preliminary lemmas

We give the behavior of $f_n(\lambda)$ and $S_n[f_n](\lambda)$. Recall that S_n is defined by $S_n[g](\lambda) = \int_{-\pi}^{\pi} g(\lambda')K_n(\lambda' - \lambda)\mathrm{d}\lambda'$ with the kernel function $K_n(t) = (2\pi n)^{-1}\sin^2(\frac{nt}{2})/\sin^2(\frac{t}{2})$. The relations that $K_n(t) \le n^{-1}(n^2 \wedge |t|^{-2})$ and $\int_{-\pi}^{\pi} K_n(t)\mathrm{d}t = 1$ are useful. We abbreviate $f(\lambda|\alpha)$ to $f(\lambda)$.

Lemma 5.9. *Assume* [FS]. *There exists $C > 0$ such that, for any $n \ge 1$,*

$$|f_n(\lambda) - f(\lambda)| \le Cn^{-\beta}|\lambda|^{-\alpha-\beta+1} \quad \text{for } |\lambda| \in [n^{-1}, \pi], \tag{5.19}$$

$$f_n(\lambda) \le C(n^{\alpha-1} \wedge |\lambda|^{-\alpha+1}) \quad \text{for } |\lambda| \in (0, \pi], \tag{5.20}$$

$$|S_n[f_n](\lambda) - f_n(\lambda)| \le Cn^{-\beta}|\lambda|^{-\alpha-\beta+1} \quad \text{for } |\lambda| \in [n^{-1}, \pi], \tag{5.21}$$

$$S_n[f_n](\lambda) \le C(n^{\alpha-1} \wedge |\lambda|^{-\alpha+1}) \quad \text{for } |\lambda| \in (0, \pi]. \tag{5.22}$$

Proof. In the proof, the constant C independent of λ and n is changed step-by-step.
Proof of (5.19). The assumption [FS] implies that, for any $|x| \ge 1$,

$$\left|\phi(x) - \nu s(\alpha)|x|^{-\alpha-1}\right| \le C|x|^{-\alpha-\beta-1}.$$

Thus, for any $|\lambda| \in [n^{-1}, \pi]$,

$$\begin{aligned} |f_n(\lambda) - f(\lambda)| &= \left|2As(\alpha)\sin^2(\lambda/2)\sum_{j=-\infty}^{\infty}\left(\frac{n^{\alpha+1}\phi(n(\lambda+2\pi j))}{\nu s(\alpha)} - |\lambda + 2\pi j|^{-\alpha-1}\right)\right| \\ &\le C|\lambda|^2\sum_{j=-\infty}^{\infty} n^{-\beta}|\lambda + 2\pi j|^{-\alpha-\beta-1} \\ &\le Cn^{-\beta}|\lambda|^{-\alpha-\beta+1}, \end{aligned}$$

where a formula $\sum_{j\neq 0}|\lambda+2\pi j|^{-\alpha-\beta+1}\le C$ is used.
Proof of (5.20). The condition [FS] implies that, for $\lambda\in(0,n^{-1}]$,

$$f_n(\lambda) \;=\; \frac{2A}{\nu}\sin^2(\lambda/2)\sum_{j=-\infty}^{\infty} n^{\alpha+1}\phi(n(\lambda+2\pi j)) \;\le\; C\sin^2(\lambda/2)n^{\alpha+1} \;\le\; Cn^{\alpha-1}.$$

On the other hand, (5.19) and Proposition 2.7 imply $f_n(\lambda)\le C\lambda^{-\alpha+1}$ for $\lambda\in[n^{-1},\pi]$.
Proof of (5.21). Let $\lambda\in[n^{-1},\pi]$. Let $d_n=|f_n(\lambda')-f_n(\lambda)|$ and $k_n=K_n(\lambda'-\lambda)$. It is sufficient to evaluate $\int_0^\pi d_nk_n\mathrm{d}\lambda'$. Let us partition $[0,\pi]$ into $\cup_{i=1}^5 A_i$, where

$$A_1=[0,(2n)^{-1}],\quad A_2=[(2n)^{-1},\lambda/2],\quad A_3=[\lambda/2,\lambda-(2n)^{-1}]\cup[\lambda+(2n)^{-1},3\lambda/2],$$
$$A_4=[\lambda-(2n)^{-1},\lambda+(2n)^{-1}],\quad A_5=[3\lambda/2,\pi].$$

If $\lambda'\in A_1$, then $d_n\le Cn^{\alpha-1}$ and $k_n\le Cn^{-1}|\lambda|^{-2}$. We have

$$\int_{A_1} d_nk_n\mathrm{d}\lambda' \;\le\; C\int_0^{(2n)^{-1}} n^{\alpha-2}\lambda^{-2}\mathrm{d}\lambda' \;\le\; Cn^{\alpha-3}\lambda^{-2} \;\le\; Cn^{-\beta}\lambda^{-\alpha-\beta+1}.$$

If $\lambda'\in A_2$, then $d_n\le Cn^{-\beta}|\lambda'|^{-\alpha-\beta+1}$ and $k_n\le Cn^{-1}|\lambda|^{-2}$. Let $r=1$ if $\beta<2-\alpha$ and $r=\log(n\lambda)$ if $\beta=2-\alpha$. Then we have

$$\int_{A_2} d_nk_n\mathrm{d}\lambda' \;\le\; C\int_{(2n)^{-1}}^{\lambda/2} n^{-\beta-1}|\lambda'|^{-\alpha-\beta+1}\lambda^{-2}\mathrm{d}\lambda' \;\le\; Cn^{\alpha-3}r\lambda^{-2} \;\le\; Cn^{-\beta}\lambda^{-\alpha-\beta+1}.$$

If $\lambda'\in A_3$, then $d_n\le Cn^{-\beta}\lambda^{-\alpha-\beta+1}$ and $k_n\le Cn^{-1}|\lambda'-\lambda|^{-2}$. We have

$$\int_{A_3} d_nk_n\mathrm{d}\lambda' \;\le\; C\int_{\lambda/2}^{\lambda-(2n)^{-1}} n^{-\beta-1}\lambda^{-\alpha-\beta+1}|\lambda'-\lambda|^{-2}\mathrm{d}\lambda' \;\le\; Cn^{-\beta}\lambda^{-\alpha-\beta+1}.$$

If $\lambda'\in A_4$, then

$$d_n \;\le\; |f_n(\lambda')-f(\lambda')|+|f(\lambda')-f(\lambda)|+|f(\lambda)-f_n(\lambda)| \;\le\; Cn^{-\beta}\lambda^{-\alpha-\beta+1},$$

where (5.19) and Proposition 2.7 are used. Since $k_n\le n$, we have

$$\int_{A_4} d_nk_n\mathrm{d}\lambda' \;\le\; Cn^{-\beta}\lambda^{-\alpha-\beta+1}.$$

Lastly, if $\lambda'\in A_5$, then $d_n\le Cn^{-\beta}|\lambda'|^{-\alpha-\beta+1}$ and $k_n\le Cn^{-1}|\lambda'|^{-2}$. We have

$$\int_{A_5} d_nk_n\mathrm{d}\lambda' \;\le\; C\int_{3\lambda/2}^{\pi} n^{-\beta-1}|\lambda'|^{-\alpha-\beta-1}\mathrm{d}\lambda' \;\le\; Cn^{-\beta-1}\lambda^{-\alpha-\beta} \;\le\; Cn^{-\beta}\lambda^{-\alpha-\beta-1}.$$

Proof of (5.22). The inequality (5.20) implies $S_n[f_n](\lambda)\le Cn^{\alpha-1}\int_{-\pi}^{\pi}K_n(t)\mathrm{d}t=Cn^{\alpha-1}$. On the other hand, if $\lambda\in[n^{-1},\pi]$, (5.20) and (5.21) imply $S_n[f_n](\lambda)\le C\lambda^{-\alpha+1}$. □

Remark 5.10. Although the relation (5.21) is also derived by Proposition 1 of Stein (1995), our proof is different because he assumed differentiability of ϕ. Instead of differentiability, we assume detailed tail behavior of ϕ as [FS]. □

As a corollary of Lemma 5.9, we obtain the following lemma.

Lemma 5.11. *Assume* [FS]. *For each* $0 < t \leq \pi$*,*

$$\begin{aligned} \sup_{t \leq |\lambda| \leq \pi} |f_n(\lambda) - f(\lambda)| &= \mathrm{O}(n^{-\beta}), \\ \sup_{t \leq |\lambda| \leq \pi} |S_n[f_n](\lambda) - f_n(\lambda)| &= \mathrm{O}(n^{-\beta}). \end{aligned}$$

□

The following lemma is used to prove Theorem 5.2. The proof is similar to Theorem (3.15) in Zygmund (2002).

Lemma 5.12. *Let* $\alpha \in (1, 2)$. *Let* g *be an even function on* $[-\pi, \pi]$. *Assume that* g *is continuously differentiable for* $\lambda \neq 0$ *and* $|\mathrm{d}^i g/\mathrm{d}\lambda^i| \leq |\lambda|^{\alpha-1-i}$ *for* $i = 0$ *and* 1. *Then there exists* $C > 0$ *such that*

$$|S_n[g](\lambda) - g(\lambda)| \;\leq\; C|\lambda|^{\alpha-2}n^{-1}\log n \tag{5.23}$$

Proof. The constant C is changed step-by-step. Let $\lambda > 0$. We evaluate $\int_0^\pi a_n \mathrm{d}\lambda'$, where $a_n = |g(\lambda') - g(\lambda)|K_n(\lambda' - \lambda)$. If $\lambda \in [n^{-1}, \pi]$, then

$$\begin{aligned} \int_0^{\lambda/2} a_n \mathrm{d}\lambda' &\leq C\int_0^{\lambda/2} \lambda^{\alpha-1}n^{-1}|\lambda' - \lambda|^{-2}\mathrm{d}\lambda' \;\leq\; C\lambda^{\alpha-2}n^{-1}, \\ \left(\int_{\lambda/2}^{\lambda-n^{-1}} + \int_{\lambda+n^{-1}}^{\pi}\right) a_n \mathrm{d}\lambda' &\leq C\int_{\lambda+n^{-1}}^{\pi} \lambda^{\alpha-2}n^{-1}|\lambda' - \lambda|^{-1}\mathrm{d}\lambda' \;\leq\; C\lambda^{\alpha-2}n^{-1}\log n, \\ \int_{\lambda-n^{-1}}^{\lambda+n^{-1}} a_n \mathrm{d}\lambda' &\leq C\int_{\lambda-n^{-1}}^{\lambda+n^{-1}} \lambda^{\alpha-2}|\lambda' - \lambda|n\mathrm{d}\lambda' \;\leq\; C\lambda^{\alpha-2}n^{-1}. \end{aligned}$$

If $\lambda \in (0, n^{-1}]$, then

$$\begin{aligned} \int_0^{\lambda+n^{-1}} a_n \mathrm{d}\lambda' &\leq C\int_0^{2n^{-1}} n^{-\alpha+1}n\mathrm{d}\lambda' \;\leq\; Cn^{-\alpha+1} \;\leq\; C\lambda^{\alpha-2}n^{-1}, \\ \int_{\lambda+n^{-1}}^{\pi} a_n \mathrm{d}\lambda' &\leq C\int_{\lambda+n^{-1}}^{\pi} \lambda^{\alpha-2}n^{-1}|\lambda' - \lambda|^{-1}\mathrm{d}\lambda' \;\leq\; C\lambda^{\alpha-2}n^{-1}\log n. \end{aligned}$$

Thus the result follows. □

5.4.2 Proof of Theorem 5.1

The outline of the proof of Theorem 5.1 is in line with Fox & Taqqu (1986). Note that $\tilde{x}_i$, $\tilde{I}_n$, f_n and S_n are defined in Subsection 5.2.1.

Let

$$\bar{c}_n(h) \;=\; \frac{1}{n}\sum_{i=1}^{n-|h|}\tilde{x}_{i+|h|}\tilde{x}_i \;=\; \int_{-\pi}^{\pi}\tilde{I}_n(\lambda)\mathrm{e}^{\mathrm{i}h\lambda}\mathrm{d}\lambda \tag{5.24}$$

and

$$c(h) \;=\; \frac{A}{2}\{|h+1|^\alpha+|h-1|^\alpha-2|h|^\alpha\} \;=\; \int_{-\pi}^{\pi}f(\lambda)\mathrm{e}^{\mathrm{i}h\lambda}\mathrm{d}\lambda. \tag{5.25}$$

Lemma 5.13. *Assume* [FS]*. For any fixed h, $\bar{c}_n(h)\xrightarrow{\mathrm{P}}c(h)$ as $n\to\infty$.*

Proof. It is sufficient to prove that $\mathrm{E}[\bar{c}_n(h)]\to c(h)$ and $\mathrm{Var}[\bar{c}_n(h)]\to 0$. Since

$$\mathrm{E}[\bar{c}_n(u)] = \int_{-\pi}^{\pi}S_n[f_n](\lambda)\mathrm{e}^{\mathrm{i}u\lambda}\mathrm{d}\lambda \quad\text{and}\quad c(u)=\int_{-\pi}^{\pi}f(\lambda)\mathrm{e}^{\mathrm{i}u\lambda}\mathrm{d}\lambda,$$

Lemma 5.9 implies that

$$\begin{aligned}\sup_{|u|\le n-1}|\mathrm{E}[\bar{c}_n(u)]-c(u)| \;&\le\; \int_{-\pi}^{\pi}|S_n[f_n](\lambda)-f(\lambda)|\mathrm{d}\lambda\\ &=\; \mathrm{O}\left[\int_0^{n^{-1}}(n^{\alpha-1}+\lambda^{-\alpha+1})\mathrm{d}\lambda+\int_{n^{-1}}^{\pi}n^{-\beta}\lambda^{-\alpha-\beta+1}\mathrm{d}\lambda\right]\\ &=\; \mathrm{O}(n^{-\beta'})\end{aligned}$$

for any $0<\beta'<\beta$. Let $\sigma_n(u)=\mathrm{E}[\tilde{x}_{i+|u|}\tilde{x}_i]$. Then

$$\sup_{|u|\le n-1}\left|\frac{n-|u|}{n}\sigma_n(u)-c(u)\right| \;=\; \mathrm{O}(n^{-\beta'})$$

Fix $h\ge 0$ and let $n'=n-h$. Since x_i's are Gaussian and $c(u)=\mathrm{O}(u^{\alpha-2})$, we obtain

$$\begin{aligned}\mathrm{Var}[\bar{c}_n(h)] \;&=\; \frac{1}{n^2}\sum_{u=-n'+1}^{n'-1}(n'-|u|)\left[\sigma_n(u)^2+\sigma_n(u+h)\sigma_n(u-h)\right]\\ &\le\; \frac{3}{n^2}\sum_{u=-n+1}^{n-1}(n-|u|)\sigma_n(u)^2\\ &=\; 3\sum_{u=-n+1}^{n-1}\frac{c(u)^2}{n-|u|}+\mathrm{O}\left[n^{-\beta'}\sum_{u=-n+1}^{n-1}\frac{1}{n-|u|}\right]\\ &=\; \mathrm{O}\left[\sum_{u=1}^{n-1}\frac{u^{2\alpha-4}}{n-u}\right]+\mathrm{O}\left[n^{-\beta'}\log n\right].\end{aligned}$$

Fix a number $p > (4-2\alpha)^{-1} \vee 1$. Then Hölder's inequality implies that

$$\sum_{u=1}^{n-1} \frac{u^{2\alpha-4}}{n-u} \;=\; \mathrm{O}\left[\sum_{u=1}^{n-1} u^{p(2\alpha-4)}\right]^{1/p} \;=\; \mathrm{O}(n^{2\alpha-4+(1/p)}) \;=\; \mathrm{o}(1).$$

Thus $\mathrm{Var}[\bar{c}_n(h)] = \mathrm{o}(1)$ □

Lemma 5.14. *Assume* [FS]. *Let φ be any continuous function on $[-\pi,\pi] \times \bar{\mathcal{A}}$. Then*

$$\int_{-\pi}^{\pi} \tilde{I}_n(\lambda)\varphi(\lambda,\alpha')\mathrm{d}\lambda \;\overset{\mathrm{P}}{\to}\; \int_{-\pi}^{\pi} f(\lambda)\varphi(\lambda,\alpha')\mathrm{d}\lambda \tag{5.26}$$

uniformly in $\alpha' \in \bar{\mathcal{A}}$.

Proof. The outline of this proof is due to Fox & Taqqu (1986). Define $\bar{c}_n(h)$ and $c(h)$ as (5.24) and (5.25). Lemma 5.13 implies

$$\int_{-\pi}^{\pi} \tilde{I}_n(\lambda)\mathrm{e}^{\mathrm{i}h\lambda}\mathrm{d}\lambda \;=\; \bar{c}_n(h) \overset{\mathrm{P}}{\to} c(h) \;=\; \int_{-\pi}^{\pi} f(\lambda)\mathrm{e}^{\mathrm{i}h\lambda}\mathrm{d}\lambda. \tag{5.27}$$

Let φ_m be the m-th Cesàro sum of φ:

$$\varphi_m(\lambda,\alpha') \;=\; \sum_{h=-m}^{m} (1-|h|/m)p_h(\alpha')\mathrm{e}^{\mathrm{i}h\lambda}. \tag{5.28}$$

The h-th Fourier coefficient $p_h(\alpha') = (2\pi)^{-1}\int_{-\pi}^{\pi}\varphi(\lambda,\alpha)\mathrm{e}^{-\mathrm{i}h\lambda}\mathrm{d}\lambda$ is continuous with respect to α'. We usually abbreviate the arguments α, α' and λ of any function below. For any $\epsilon > 0$, there exists a positive integer m such that $\sup_{\lambda,\alpha'}|\varphi - \varphi_m| < \epsilon/2c(0)$. Thus

$$\sup_{\alpha'}\left|\int_{-\pi}^{\pi}(\tilde{I}_n - f)(\varphi-\varphi_m)\mathrm{d}\lambda\right| \;\le\; \frac{\epsilon}{2c(0)}\int_{-\pi}^{\pi}(\tilde{I}_n + f)\mathrm{d}\lambda \;=\; \frac{\epsilon(\bar{c}_n(0)+c(0))}{2c(0)} \overset{\mathrm{P}}{\to} \epsilon.$$

On the other hand, $|\int_{-\pi}^{\pi}(\tilde{I}_n - f)\varphi_m\mathrm{d}\lambda| \overset{\mathrm{P}}{\to} 0$ uniformly in α' by (5.27) and (5.28). Thus the lemma follows. □

Proof of Theorem 5.1. The function $\{f(\lambda|\alpha')\}^{-1}$ is continuous on $[-\pi,\pi]\times\bar{\mathcal{A}}$ (see Section 2.5). Thus Lemma 5.14 implies

$$\int_{-\pi}^{\pi}\frac{\tilde{I}_n(\lambda)}{f(\lambda|\alpha')}\mathrm{d}\lambda \;\overset{\mathrm{P}}{\to}\; \int_{-\pi}^{\pi}\frac{f(\lambda|\alpha)}{f(\lambda|\alpha')}\mathrm{d}\lambda \tag{5.29}$$

uniformly in $\alpha' \in \bar{\mathcal{A}}$. The right hand side of this expression has the unique minimum at $\alpha' = \alpha$. The theorem follows from Theorem 5.7 in van der Vaart (1998). □

5.4.3 Proof of Theorem 5.2

The outline of the proof of Theorem 5.2 is in line with Fox & Taqqu (1986). Note that $\tilde{x}_i$, $\tilde{I}_n$, f_n and S_n are defined in Subsection 5.2.1.

Let g and g_n be continuous functions defined on $[-\pi,\pi]\setminus\{0\}$ with the following properties: $|g(\lambda)| = \mathrm{O}(|\lambda|^{\alpha-1-\delta})$ and $|g_n(\lambda)| = \mathrm{O}(|\lambda|^{\alpha-1-\delta})$ as $\lambda\to 0$ for any $\delta>0$; for any $0<t\le\pi$, there exists $C=C(t)>0$ such that $\sup_{t\le|\lambda|\le\pi}|g_n(\lambda)-g(\lambda)|\le Cn^{-\beta}$. We show the asymptotic normality of $\int_{-\pi}^{\pi}\tilde{I}_n(\lambda)g_n(\lambda)\mathrm{d}\lambda$ by using the idea of Fox & Taqqu (1987). We shall take $g_n(\lambda)=\partial_\alpha f^{-1}(\lambda|\alpha_n)$ later. We prepare some additional notations. Most of the notations are the same as those in Fox & Taqqu (1987). The argument α of a function is usually abbreviated as $f(\lambda)=f(\lambda|\alpha)$. Fix a positive integer p. Let $U_t=[-t,t]^{2p}$ for each $t\in(0,\pi]$. For $y=(y_1,\cdots,z_{2p})\in U_\pi$, we put

$$\begin{aligned} P_n(y) &= \sum_{j_1=0}^{n-1}\cdots\sum_{j_{2p}=0}^{n-1}\mathrm{e}^{\mathrm{i}(j_1-j_2)y_1}\mathrm{e}^{\mathrm{i}(j_2-j_3)y_2}\ldots\mathrm{e}^{\mathrm{i}(j_{2p}-j_1)y_{2p}},\\ Q_n(y) &= f_n(y_1)g_n(y_2)f_n(y_3)\cdots g_n(y_{2p}),\\ Q(y) &= f(y_1)g(y_2)f(y_3)\cdots g(y_{2p}). \end{aligned}$$

Let

$$\begin{aligned} r_{jk} &= \int_{-\pi}^{\pi} f_n(\lambda)\mathrm{e}^{\mathrm{i}(j-k)\lambda}\mathrm{d}\lambda,\\ a_{jk} &= \int_{-\pi}^{\pi} g_n(\lambda)\mathrm{e}^{\mathrm{i}(j-k)\lambda}\mathrm{d}\lambda. \end{aligned}$$

Let R_n and A_n be the matrix whose (j,k) component is r_{jk} and a_{jk}, respectively. Then $\mathrm{tr}(R_nA_n)^p=\int_{U_\pi}P_nQ_n\mathrm{d}y$. Let μ be the measure on U_π which is concentrated on $D=\{y\in U_\pi|y_1=\ldots=y_{2p}\}$ and satisfies $\mu\{y|a\le y_1=\ldots=y_{2p}\le b\}=b-a$ for all $-\pi\le a\le b\le\pi$. Introduce the sets $W_k=\{y\in\mathbb{R}^{2p}\mid |y_k|\le|y_{k+1}|/2\}$ for $k=1,\cdots,2p$, where we interpret $y_{2p+1}=y_1$, and $W=W_1\cup W_2\cup\cdots\cup W_{2p}$. For each $t\in(0,\pi]$, we define three sets E_t, F_t and G by $E_t=U_\pi\setminus(W\cup U_t)$, $F_t=U_t\setminus W$ and $G=U_\pi\cap W$.

Lemma 5.15. *Assume* [FS]. *For any* $0<t\le 1$,

$$\sup_{y\in E_t}|Q_n(y)-Q(y)| = \mathrm{O}(n^{-\beta}).$$

Proof. Let $y\in E_t$. Then $y_j>t/2^{2p-1}$ for $j=1,\cdots,2p$ (Fox & Taqqu, 1987, p.237). Let $\Lambda=\{1,2,\cdots,2p\}$. Put $F_n^{(i)}=f_n$ if i is odd and g_n if i is even. Similarly, put $F^{(i)}=f$ if i is odd and g if i is even. It holds that

$$Q_n-Q = \prod_{i\in\Lambda}F_n^{(i)}(y_i)-\prod_{i\in\Lambda}F^{(i)}(y_i) = \sum_{\emptyset\subsetneq S\subset\Lambda}\prod_{i\in S}(F_n^{(i)}(y_i)-F^{(i)}(y_i))\prod_{j\notin S}F^{(i)}(y_j).$$

The result follows from Lemma 5.11 and the assumption on g_n. □

Lemma 5.16.

$$n^{-1}\int_{U_\pi} |P_n(y)|\mathrm{d}y, \quad = \quad \mathrm{O}((\log n)^{2p-1}). \tag{5.30}$$

Proof. We have

$$P_n(y) \quad = \quad \frac{\sin(n(y_1-y_{2p})/2)}{\sin((y_1-y_{2p})/2)}\frac{\sin(n(y_2-y_1)/2)}{\sin((y_2-y_1)/2)}\cdots\frac{\sin(n(y_{2p}-y_{2p-1})/2)}{\sin((y_{2p}-y_{2p-1})/2)}.$$

By putting $u_1 = y_2 - y_1, \cdots, u_{2p-1} = y_{2p} - y_{2p-1}$,

$$\begin{aligned}|P_n(y)| \quad &= \quad \left|\frac{\sin(n(u_1+\cdots+u_{2p-1})/2)}{\sin((u_1+\cdots+u_{2p-1})/2)}\frac{\sin(nu_1/2)}{\sin(u_1/2)}\cdots\frac{\sin(nu_{2p-1}/2)}{\sin(u_{2p-1}/2)}\right|\\ &\leq \quad h_n(u_1+\cdots+u_{2p-1})h_n(u_1)\cdots h_n(u_{2p-1}),\end{aligned}$$

where $h_n(x)$ is a 2π-periodic function with $h_n(x) = 4(n\wedge|x|^{-1})$ for $x\in[-\pi,\pi]$. Then we obtain

$$\begin{aligned}n^{-1}\int_{U_\pi}|P_n(y)|\mathrm{d}y \quad &\leq \quad 2\pi n^{-1}\int_{[-2\pi,2\pi]^{2p-1}} h_n(u_1+\cdots+u_{2p-1})h_n(u_1)\cdots h_n(u_{2p-1})\mathrm{d}u\\ &\leq \quad 2\pi\left[\int_{-2\pi}^{2\pi} h_n(u)\mathrm{d}u\right]^{2p-1}\\ &= \quad (2\pi)16^{2p-1}(1+\log(n/\pi))^{2p-1},\end{aligned}$$

where a formula $\int_{-\pi}^{\pi} h_n(x)\mathrm{d}x = 8(1+\log(n/\pi))$ is used. □

The following lemma and proposition are generalization of Theorem 1 and Theorem 2 in Fox & Taqqu (1987), respectively. Their results correspond to the case that $f_n = f$ and $g_n = g$ for all n.

Lemma 5.17. *Assume* [FS]. *Then*

$$\lim_{n\to\infty} n^{-1}\mathrm{tr}(R_nA_n)^p \quad = \quad (2\pi)^{2p-1}\int_{-\pi}^{\pi}[f(\lambda)g(\lambda)]^p\mathrm{d}\lambda. \tag{5.31}$$

Proof. The left hand side of (5.31) is equal to $\lim_n n^{-1}\int_{U_\pi} P_nQ_n\mathrm{d}y$. It is sufficient to show that

$$\lim_{n\to\infty} n^{-1}\int_{E_t} P_nQ_n\mathrm{d}y \quad = \quad (2\pi)^{2p-1}\int_{t\leq|\lambda|\leq\pi}[f(\lambda)g(\lambda)]^p\mathrm{d}\lambda, \quad 0<t\leq 1, \tag{5.32}$$

$$\lim_{t\to 0}\limsup_{n\to\infty} n^{-1}\int_{F_t} P_nQ_n\mathrm{d}y \quad = \quad 0, \tag{5.33}$$

and

$$\lim_{n\to\infty} n^{-1}\int_G P_nQ_n\mathrm{d}y \quad = \quad 0. \tag{5.34}$$

Lemma 5.15 and 5.16 imply

$$\left| n^{-1} \int_{E_t} P_n(Q_n - Q)\mathrm{d}y \right| \;=\; \mathrm{O}(n^{-\beta}(\log n)^{2p-1}) \;=\; \mathrm{o}(1).$$

Therefore (5.32) is shown if we prove that

$$\lim_{n\to\infty} n^{-1} \int_{E_t} P_n Q \mathrm{d}y \;=\; (2\pi)^{2p-1} \int_{t\le|\lambda|\le\pi} [f(\lambda)g(\lambda)]^p \mathrm{d}\lambda.$$

This equality is exactly shown in the proof of Theorem 1 in Fox & Taqqu (1987). Next, for any fixed $\delta > 0$, Lemma 5.9 implies

$$|Q_n(y)| \;\le\; R_\delta(y) \;:=\; C|y_1|^{-\alpha+1-\delta}|y_2|^{\alpha-1-\delta}|y_3|^{-\alpha+1-\delta}\cdots|y_{2p}|^{\alpha-1-\delta}$$

with some $C > 0$. Therefore (5.33) and (5.34) are shown if we prove the following two relations for some $\delta > 0$:

$$\lim_{t\to 0}\limsup_{n\to\infty} n^{-1} \int_{F_t} |P_n| R_\delta \mathrm{d}y \;=\; 0$$

and

$$\lim_{n\to\infty} n^{-1} \int_{G} |P_n| R_\delta \mathrm{d}y \;=\; 0.$$

The above two formulas are proved in the proof of Theorem 1 in Fox & Taqqu (1987). Therefore we obtain (5.32), (5.33) and (5.34). □

Proposition 5.18. *Assume* [FS]. *Let* $Z_n = \int_{-\pi}^{\pi} g_n(\lambda)\tilde{I}_n(\lambda)\mathrm{d}\lambda$. *Then* $n^{1/2}(Z_n - \mathrm{E}[Z_n])$ *converges in distribution to* $N(0, V)$*, where*

$$V \;=\; 4\pi \int_{-\pi}^{\pi} [f(\lambda)g(\lambda)]^2 \mathrm{d}\lambda.$$

Proof. The p-th order cumulant of $n^{1/2}Z_n$ is $c_{p,n} = 2^{-1}(p-1)!\pi^{-p}n^{-p/2}\mathrm{tr}(R_n A_n)^p$. Lemma 5.17 implies $\lim_n c_{p,n} = 0$ for $p \ge 3$ and $\lim_n c_{2,n} = 4\pi \int (fg)^2 \mathrm{d}\lambda$ for $p = 2$. □

Now we give the proof of Theorem 5.2.

Proof of Theorem 5.2. Put $f_{\alpha'} = f(\lambda|\alpha')$ and $\partial = \partial_{\alpha'} = \partial/\partial\alpha'$. Theorem 5.1 implies $\hat{\alpha}_n \xrightarrow{\mathrm{P}} \alpha$. We have $\alpha_n \to \alpha$. In fact, by Lemma 5.9, the objective function $\int_{-\pi}^{\pi} S_n[f_n]/f_{\alpha'}\mathrm{d}\lambda$ converges to $\int_{-\pi}^{\pi} f_\alpha/f_{\alpha'}\mathrm{d}\lambda$ uniformly in $\alpha' \in \bar{\mathcal{A}}$ and the limit has the unique minimum at $\alpha' = \alpha$. By Taylor's formula, there exists $\hat{\alpha}_n^*$ between $\hat{\alpha}_n$ and α_n such that

$$0 \;=\; \int_{-\pi}^{\pi} (\partial f_{\hat{\alpha}_n}^{-1})\tilde{I}_n \mathrm{d}\lambda \;=\; \int_{-\pi}^{\pi} (\partial f_{\alpha_n}^{-1})\tilde{I}_n \mathrm{d}\lambda + (\hat{\alpha}_n - \alpha_n)\int_{-\pi}^{\pi} (\partial^2 f_{\hat{\alpha}_n^*}^{-1})\tilde{I}_n \mathrm{d}\lambda.$$

By using Lemma 5.14, we obtain

$$\int_{-\pi}^{\pi}(\partial^2 f_{\hat{\alpha}_n^*}^{-1})\tilde{I}_n \mathrm{d}\lambda \ \xrightarrow{\mathrm{P}} \ \int_{-\pi}^{\pi}(\partial^2 f_{\alpha}^{-1})f_\alpha \mathrm{d}\lambda \ = \ \int_{-\pi}^{\pi}(\partial f_\alpha/f_\alpha)^2 \mathrm{d}\lambda \ = \ 4\pi J.$$

By Proposition 5.18, we obtain

$$\sqrt{n}\left(\int_{-\pi}^{\pi}(\partial f_{\alpha_n}^{-1})\tilde{I}_n \mathrm{d}\lambda - \mathrm{E}\left[\int_{-\pi}^{\pi}(\partial f_{\alpha_n}^{-1})\tilde{I}_n \mathrm{d}\lambda\right]\right) \ \rightsquigarrow \ N(0, 16\pi^2 J).$$

By the definition of α_n, we obtain

$$\mathrm{E}\left[\int_{-\pi}^{\pi}(\partial f_{\alpha_n}^{-1})\tilde{I}_n \mathrm{d}\lambda\right] = \int_{-\pi}^{\pi}(\partial f_{\alpha_n}^{-1})S_n[f_n]\mathrm{d}\lambda = 0.$$

Thus $\sqrt{n}(\hat{\alpha}_n - \alpha_n) \rightsquigarrow N(0, J^{-1})$ is proved. We next evaluate α_n. By Taylor's formula, there exists α_n^* between α_n and α such that

$$0 \ = \ \int_{-\pi}^{\pi}(\partial f_{\alpha_n}^{-1})S_n[f_n]\mathrm{d}\lambda \ = \ \int_{-\pi}^{\pi}(\partial f_{\alpha}^{-1})S_n[f_n]\mathrm{d}\lambda + (\alpha_n - \alpha)\int_{-\pi}^{\pi}(\partial^2 f_{\alpha_n^*}^{-1})S_n[f_n]\mathrm{d}\lambda.$$

Since $\int_{-\pi}^{\pi}(\partial^2 f_{\alpha_n^*}^{-1})S_n[f_n]\mathrm{d}\lambda \xrightarrow{\mathrm{P}} 4\pi J$ and

$$\int_{-\pi}^{\pi}(\partial f_{\alpha}^{-1})S_n[f_n]\mathrm{d}\lambda \ = \ \int_{-\pi}^{\pi}(\partial f_{\alpha}^{-1})(S_n[f_n] - f_\alpha)\mathrm{d}\lambda \ = \ \mathrm{O}(n^{-\beta}),$$

we obtain

$$\alpha_n - \alpha \ = \ \frac{1}{4\pi J}\int_{-\pi}^{\pi}\frac{\partial f_\alpha S_n[f_n]}{f_\alpha^2}\mathrm{d}\lambda + \mathrm{o}(n^{-\beta}) \ = \ \mathrm{O}(n^{-\beta}).$$

Lastly, we have

$$\int_{-\pi}^{\pi}\partial f_\alpha^{-1}(S_n[f_n] - f_n)\mathrm{d}\lambda \ = \ \mathrm{o}(n^{-\beta})$$

because the left hand side is equal to $\int_{-\pi}^{\pi}\{S_n[\partial f_\alpha^{-1}] - \partial f_\alpha^{-1}\}f_n \mathrm{d}\lambda$ and it is $\mathrm{O}(n^{-1}(\log n)^2)$ due to Lemma 5.9 and Lemma 5.12. Thus the theorem is proved. □

5.4.4 Proof of Theorem 5.6

Let

$$\sigma_n(h) \ = \ \frac{A}{2\gamma(\frac{1}{n})}\mathrm{E}[x_{i+h}x_i] \ = \ \frac{A}{2\gamma(\frac{1}{n})}\{\gamma(\tfrac{h+1}{n}) + \gamma(\tfrac{h-1}{n}) - 2\gamma(\tfrac{h}{n})\}$$

and

$$c(h) \ = \ \frac{A}{2}\{|h+1|^\alpha + |h-1|^\alpha - 2|h|^\alpha\}.$$

Lemma 5.19. *Assume* [F1]–[F4]. *Then* $\sup_{|h|\le n-1}|\sigma_n(h) - c(h)| = \mathrm{o}(1)$ *as* $n \to \infty$.

We prepare two lemmas before proving Lemma 5.19. They are shown in Bingham et al. (1987).

Lemma 5.20 (Uniform convergence theorem). *Assume* [F1] *and* [F2]. *Let* $0 < \underline{u} < \overline{u}$. *Then*

$$\lim_{t\downarrow 0} \sup_{u\in[\underline{u},\overline{u}]} \left| \frac{\gamma(tu)}{\gamma(t)} - u^{\alpha} \right| = 0.$$

□

Lemma 5.21 (Potter's theorem). *Assume* [F1] *and* [F2]. *For any* $\beta > 0$, *there exists* $\delta > 0$ *such that* $\gamma(t)/\gamma(u) \le (t/u)^{\alpha+\beta}$ *for any* $0 < u \le t \le \delta$. □

Proof of Lemma 5.19. Let $g_n(x) = \gamma(\frac{x}{n})/\gamma(\frac{1}{n}) - |x|^{\alpha}$. Then

$$\sigma_n(h) - c(h) = \frac{A}{2}\left\{g_n(h+1) + g_n(h-1) - 2g_n(h)\right\}.$$

The condition [F2] implies that $g_n(h) = \mathrm{o}(1)$ for fixed h. Thus it is sufficient to prove that $\sup_{h\in[2,n-1]} |g_n(h+1) + g_n(h-1) - 2g_n(h)| = \mathrm{o}(1)$. By the mean-value theorem,

$$g_n(h+1) + g_n(h-1) - 2g_n(h) = \int_0^1\int_0^1 g_n^{(2)}(h+u-v)\mathrm{d}u\mathrm{d}v.$$

The expression of $g_n^{(2)}(x)$ is

$$g_n^{(2)}(x) = \frac{\gamma^{(2)}(\frac{x}{n})}{n^2\gamma(\frac{1}{n})} - \alpha_{(2)}x^{\alpha-2}.$$

Fix $\epsilon > 0$. It is sufficient to prove that $\limsup_{n\to\infty} \sup_{x\in[1,n]} |g_n^{(2)}(x)| \le \epsilon$. We partition the interval $[1, n]$ into $[1, x_0]$, $[x_0, n\delta]$ and $[n\delta, n]$, where x_0 and δ are fixed numbers defined as follows. There exists a number x_0 such that $x_0 \ge 1$ and $3\alpha_{(2)}x_0^{\alpha/2-1} \le \epsilon$. By the condition [F4] and Lemma 5.21, there exists $\delta > 0$ such that $\sup_{t\in(0,\delta]} |t^2\gamma^{(2)}(t)/\gamma(t)| \le 2\alpha_{(2)}$ and $\gamma(t)/\gamma(u) \le (t/u)^{\alpha/2+1}$ for any $0 < u \le t \le \delta$. Now we evaluate $\sup|g_n^{(2)}(x)|$ for the three intervals $[1, x_0]$, $[x_0, n\delta]$ and $[n\delta, n]$. From the condition [F4] and Lemma 5.20,

$$\sup_{x\in[1,x_0]} |g_n^{(2)}(x)| = \sup_{x\in[1,x_0]} \left| \frac{(\frac{x}{n})^2\gamma^{(2)}(\frac{x}{n})}{\gamma(\frac{x}{n})}\frac{\gamma(\frac{x}{n})}{\gamma(\frac{1}{n})}x^{-2} - \alpha_{(2)}x^{\alpha-2} \right| = \mathrm{o}(1).$$

By the definition of δ and x_0,

$$\begin{aligned}\sup_{x\in[x_0,n\delta]} |g_n^{(2)}(x)| &= \sup_{x\in[x_0,n\delta]} \left| \frac{(\frac{x}{n})^2\gamma^{(2)}(\frac{x}{n})}{\gamma(\frac{x}{n})}\frac{\gamma(\frac{x}{n})}{\gamma(\frac{1}{n})x^{\alpha/2+1}}x^{\alpha/2-1} - \alpha_{(2)}x^{\alpha-2} \right| \\ &\le \sup_{x\in[x_0,n\delta]} \left| \frac{(\frac{x}{n})^2\gamma^{(2)}(\frac{x}{n})}{\gamma(\frac{x}{n})} \right| \left| \frac{\gamma(\frac{x}{n})}{\gamma(\frac{1}{n})x^{\alpha/2+1}} \right| x_0^{\alpha/2-1} + \alpha_{(2)}x_0^{\alpha-2} \\ &\le 2\alpha_{(2)}x_0^{\alpha/2-1} + \alpha_{(2)}x_0^{\alpha-2} \le 3\alpha_{(2)}x_0^{\alpha/2-1} \le \epsilon.\end{aligned}$$

By the condition [F3], $G_0 := \sup_{t\in[\delta,1]} |\gamma^{(2)}(t)| < \infty$. Thus we obtain

$$\sup_{x\in[n\delta,n]} |g_n^{(2)}(x)| \ \leq \ \frac{G_0}{n^2\gamma(\frac{1}{n})} + \alpha(\alpha-1)(n\delta)^{\alpha-2} \ = \ \mathrm{o}(1).$$

Therefore we obtain $\limsup_{n\to\infty}\sup_{x\in[1,n]} |g_n^{(2)}(x)| \leq \epsilon$. □

Proof of Theorem 5.6. Define $\tilde{x}_i = (A/2\gamma(n^{-1}))^{1/2}x_i$, $\bar{c}_n(h) = n^{-1}\sum_{i=1}^{n-|h|}\tilde{x}_{i+|h|}\tilde{x}_i$ and $\tilde{I}_n(\lambda) = I_n(\lambda;\tilde{x})$. The proofs of Lemma 5.14 and Theorem 5.1 remains valid if one shows that $\bar{c}_n(h) \overset{\mathrm{P}}{\to} c(h)$ as $n\to\infty$ for any fixed h. By Lemma 5.19, we obtain

$$\begin{aligned}
\mathrm{E}[\bar{c}_n(h)] &= (1-|h|/n)\sigma_n(h) \ \to \ c(h),\\
\mathrm{Var}[\bar{c}_n(h)] &= n^{-2}\sum_{u=-n+h+1}^{n-h-1}(n-h-|u|)(\sigma_n(u)^2+\sigma_n(u+h)\sigma_n(u-h))\\
&= \mathrm{O}\left[n^{-1}\sum_{u=1}^{n}c(u)^2\right] + \mathrm{o}(1) \ = \ \mathrm{o}(1),
\end{aligned}$$

where $c(u) = \mathrm{O}(u^{\alpha-2})$ is used for the last equality. Thus $\bar{c}_n(h) \overset{\mathrm{P}}{\to} c(h)$. □

5.4.5 Proof of Theorem 5.7

Proof of Theorem 5.7. Assume [F5]–[F10]. The boundedness of $\phi(x)$, and also continuity, follows from the condition [F7] because the autocovariance function is given by $\sigma(t) = \gamma(\infty)-\gamma(t)$. We prove (5.6). Let $\rho^{(2)}(t) = \gamma^{(2)}(t) - \alpha_{(2)}\nu|t|^{\alpha-2}$. By the condition [F10], we have $\delta > 0$ such that $\rho^{(2)}(t)$ is monotone over $t\in(0,\delta)$. Let $x>0$. By [F5]–[F8] and integration by parts, we obtain

$$\begin{aligned}
\pi\phi(x) &= \int_0^\infty(\gamma(\infty)-\gamma(t))\cos(tx)\mathrm{d}t\\
&= x^{-1}\int_0^\infty\gamma^{(1)}(t)\sin(tx)\mathrm{d}t\\
&= x^{-2}\int_0^\infty\gamma^{(2)}(t)\cos(tx)\mathrm{d}t\\
&= x^{-2}\int_0^\infty\nu\alpha_{(2)}t^{\alpha-2}\cos(tx)\mathrm{d}t + x^{-2}\int_0^\infty\rho^{(2)}(t)\cos(tx)\mathrm{d}t\\
&= I_1 + I_2.
\end{aligned}$$

Then

$$I_1 \ = \ \nu\alpha_{(2)}x^{-\alpha-1}\int_0^\infty u^{\alpha-2}\cos(u)\mathrm{d}u \ = \ \pi\nu s(\alpha)x^{-\alpha-1}.$$

We show $I_2 = \mathrm{O}(x^{-\alpha-\beta-1})$. The integral I_2 is decomposed into three parts as follows.

$$\begin{aligned} I_2 &= x^{-2}\int_0^{1/x} \rho^{(2)}(t)\cos(tx)\mathrm{d}t + x^{-2}\int_{1/x}^{\delta} \rho^{(2)}(t)\cos(tx)\mathrm{d}t + x^{-2}\int_{\delta}^{\infty} \rho^{(2)}(t)\cos(tx)\mathrm{d}t \\ &= J_1 + J_2 + J_3. \end{aligned}$$

The condition [F6] implies $\rho^{(2)}(t) = \mathrm{O}(t^{\alpha+\beta-2})$ as $t \to 0$, and therefore

$$J_1 = x^{-2}\int_0^{1/x} \rho^{(2)}(t)\cos(tx)\mathrm{d}tx = \mathrm{O}\left(x^{-2}\int_0^{1/x} t^{\alpha+\beta-2}\mathrm{d}t\right) = \mathrm{O}(x^{-\alpha-\beta-1}).$$

By the definition of δ and the second mean-value theorem, there exists $\delta' \in (1/x, \delta)$ such that

$$\begin{aligned} J_2 &= x^{-2}\int_{1/x}^{\delta} \rho^{(2)}(t)\cos(tx)\mathrm{d}tx \\ &= x^{-2}\left[\rho^{(2)}(1/x)\int_{1/x}^{\delta'} \cos(tx)\mathrm{d}tx - \rho^{(2)}(\delta)\int_{\delta'}^{\delta} \cos(tx)\mathrm{d}tx\right] \\ &= x^{-3}\left[\rho^{(2)}(1/x)(\sin(\delta' x) - \sin(1)) - \rho^{(2)}(\delta)(\sin(\delta x) - \sin(\delta' x))\right] \\ &= \mathrm{O}(x^{-\alpha-\beta-1}) + \mathrm{O}(x^{-3}). \end{aligned}$$

Finally, integration by parts and the condition [F9] imply

$$\begin{aligned} J_3 &= x^{-2}\int_{\delta}^{\infty} \rho^{(2)}(t)\cos(tx)\mathrm{d}tx \\ &= \left[x^{-3}\rho^{(2)}(t)\sin(tx)\right]_{\delta}^{\infty} - x^{-3}\int_{\delta}^{\infty} \rho^{(3)}(t)\sin(tx)\mathrm{d}t \\ &= \mathrm{O}(x^{-3}). \end{aligned}$$

Therefore $I_2 = J_1 + J_2 + J_3 = \mathrm{O}(x^{-\alpha-\beta-1}) + \mathrm{O}(x^{-3}) = \mathrm{O}(x^{-\alpha-\beta-1})$ as $x \to \infty$. □

5.5 Discussions

We proved that QMLE $\hat{\alpha}_n$ has the consistency and asymptotic normality. To determine the efficiency of estimators is not easy because we consider the semiparametric model. The asymptotic variance of QMLE is optimal because it is same as the inverse of Fisher information matrix of FGN and the semiparametric model includes FGN. The bias is of $\mathrm{O}(n^{-\beta})$ that is same as the other estimators discussed in Chapter 2.

We assumed that X_t is stationary throughout this chapter. However, the essential condition is a weaker condition that X_t is an intrinsic random field (Chapter 2). In fact, Theorem 5.6 holds also under the weaker assumption. To prove the other theorems under the weaker condition is a future work. To study the multi-fractal index (i.e. spatially varying fractal index) and non-lattice sampling schemes is also a challenging problem.

Chapter 6

Transformed Gaussian model

The contents of this chapter are reported in Sei (2004).

6.1 Introduction

In this chapter, we prove the *local asymptotic mixed normality* (LAMN; see Chapter 3) of a class of transformed Gaussian models for random fields with time-parameter space $[0,1]^d$ with discrete observations, where d is a positive integer.

The transformed Gaussian model for random fields is an important model for processes with non-Gaussian marginal distributions. Several geostatistical methods including trans-Gaussian kriging (see e.g. Cressie (1993)) assume it. The transformed model was initiated by Box & Cox (1964), who applied it to factorial experiments. A Bayesian prediction procedure for the transformed Gaussian models was treated by De Oliveira et al. (1997).

We suppose that the process is observed at discrete lattice points in a unit cube and investigate *fixed domain asymptotics*, which means that the observed points increase densely in a fixed domain as described in Section 2.2. Chan & Wood (2004) studied increment-based estimators of the fractal dimension of transformed Gaussian models under the framework of fixed domain asymptotics. They showed that the normalized difference between the estimator and the true parameter converges to a mixed normal random variable. Their setting is close to ours. However, their interest is semiparametric inference and they did not give any likelihood-based results. Synthesis of their result and our result is a future work.

The original Gaussian random field of our transformed Gaussian models is assumed to be the product of a deterministic process and a process with independent increments. A stationary random field called the Ornstein-Uhlenbeck sheet is included in our setting as explained in Section 6.4. We also assume that the original Gaussian process is known and that only the transformation function is unknown. Estimation of the transformation function is important when the marginal distribution is of interest.

The LAMN property implies the convergence of the likelihood ratio to that of the corresponding mixed normal model (van der Vaart, 1998, Theorem 9.8). Therefore it allows us to reduce statistical problems to those of the mixed normal model asymptotically. We gave the related facts in Chapter 3.

Our result is a generalization of 1-dimensional case by Dohnal (1987) and Genon-Catalot & Jacod (1993, 1994), who proved the LAMN property of discretely observed diffusion models with unknown diffusion coefficients.

The chapter is organized as follows. In Section 6.2 we describe the transformed Gaussian models. In Section 6.3 we state our main theorem of LAMN (Theorem 6.1). Several notations and regularity conditions are explained there. In Section 6.4 the quantities appeared in Theorem 6.1 are calculated for several examples. Section 6.5 is devoted to the proof of Theorem 6.1. Finally we give some discussions in Section 6.6.

6.2 Transformed Gaussian model

Let us consider a d-dimensional time parameter Gaussian random field $Y = (Y_t \mid t \in [0,1]^d)$ defined by

$$Y_t = \gamma_t \int_{(-\infty,t]} \beta_s \nu(\mathrm{d}s) \tag{6.1}$$

for $t = (t_1, \cdots, t_d) \in [0,1]^d$, where $(-\infty, t] = \prod_{i=1}^d (-\infty, t_i]$, β is a nonnegative function, γ is a positive function, and a random measure ν is a Gaussian white noise on $\mathbb{R}^d$. The regularity conditions for β and γ are described in Subsection 6.3.2. The white noise ν on $\mathbb{R}^d$ is defined as a Gaussian process on the Borel-field $\mathcal{B}(\mathbb{R}^d)$ of $\mathbb{R}^d$ with $\mathbb{E}[\nu(A)] = 0$ and $\mathbb{E}[\nu(A)\nu(B)] = \mathrm{Leb}(A \cap B)$ for all $A, B \in \mathcal{B}(\mathbb{R}^d)$, where Leb is the Lebesgue measure on $\mathbb{R}^d$. The covariance matrix of Y is given by

$$\mathrm{E}[Y_t Y_s] = \gamma_t \gamma_s \int_{(-\infty, t \curlywedge s]} \beta_u^2 \mathrm{d}u,$$

where $t \curlywedge s = (t_1 \wedge s_1, \cdots, t_d \wedge s_d)$ and $t_i \wedge s_i = \min(t_i, s_i)$. The process $(\gamma_t^{-1} Y_t \mid t \in [0,1]^d)$ has independent increments as stated in Subsection 6.5.4. For example, the Brownian sheet and the Ornstein-Uhlenbeck sheet are examples of this class of processes as described in Section 6.4.

The *transformed Gaussian model* for a random field $X = (X_t \mid t \in [0,1]^d)$ is defined by

$$g(X_t; t, \theta) = Y_t, \quad t \in [0,1]^d,$$

where $\theta \in \Theta = [\underline{\theta}, \overline{\theta}] \subset \mathbb{R}$ is an unknown parameter, $Y = (Y_t \mid t \in [0,1]^d)$ is defined by (6.1) with known functions β and γ, and $g : \mathbb{R} \times [0,1]^d \times \Theta \to \mathbb{R}$ is a function satisfying the regularity conditions described in Subsection 6.3.2. Since g depends on t, we can always assume that $\gamma_t = 1$ without loss of generality. However, γ_t is left for convenience (see examples in Section 6.4).

We assume that the process X is discretely observed at $(n+1)^d$ equipatitioned lattice points $\bar{D}_n^d = \{0, n^{-1}, 2n^{-1}, \cdots, 1\}^d$ in $[0,1]^d$. Thus we treat a model

$$g(X_t; t, \theta) \quad = \quad Y_t, \quad t \in \bar{D}_n^d. \tag{6.2}$$

The model is useful for modeling processes with non-Gaussian marginal distributions. Another advantage of the model is that the likelihood function is explicitly expressed.

6.3 Main result

6.3.1 Notations

Let $(\Omega, \mathcal{B}, \mathrm{P})$ be a probability space on which the white noise ν is defined. If the parameter θ is specified, the probability measure induced by X is denoted by P_θ. All random fields treated in the chapter are real-valued and d-parameter processes unless otherwise stated. Furthermore, we always assume almost sure continuity of the processes.

For a positive integer k, let $[k] = \{1, 2, \cdots, k\}$ and $\overline{[k]} = \{0, 1, 2, \cdots, k\}$. For a finite set S, $\sharp S$ denotes the cardinality of S. Recall that a variable $\mathbb{I}_{(A)}$ takes 1 if a proposition A is true, 0 otherwise.

An order $\preceq$ on $\mathbb{R}^d$ is introduced: for $s = (s_1, \cdots, s_d)$ and $t = (t_1, \cdots, t_d) \in \mathbb{R}^d$, we write $s \preceq t$ if $s_j \leq t_j$ holds for any $j \in [d]$. The infimum of s and t with respect to the order $\preceq$ is $s \curlywedge t$. Rectangles $\prod_{i=1}^d [s_i, t_i]$ and $\prod_{i=1}^d (s_i, t_i]$ generated by s and t ($s \preceq t$) are denoted by $[s, t]$ and $(s, t]$, respectively.

In the present chapter, discrete observations of a d-parameter stochastic process $X = (X_s \mid s \in [0,1]^d)$ are considered. Let n be a positive integer and $\delta = 1/n$. The set of observed points is lattice points $\bar{D}_n^d = \{0, \delta, 2\delta, \cdots, n\delta\}^d$ in the cube $[0,1]^d$. Let $D_n^d = \{\delta, 2\delta, \cdots, n\delta\}^d$.

For $t \in \mathbb{R}^d$ and $a \subset [d]$, the a-marginal $(t_j)_{j \in a}$ of t is denoted by t_a. For $t, u \in \mathbb{R}^d$ and $a \subset [d]$, $t + u_a$ means $(t_j + u_j \mathbb{I}_{(j \in a)})_{j=1}^d$. For any $\lambda \in \mathbb{R}$, the vector $(\lambda, \cdots, \lambda) \in \mathbb{R}^d$ is abbreviated by λ if there is no confusion. For example, $t - \delta + \delta_a$ denotes a vector whose j-th component is t_j if $j \in a$, $t_j - \delta$ otherwise.

The symbol ∂_x denotes the partial derivative with respect to an argument x, that is, $\partial_x = \partial / \partial x$.

6.3.2 Regularity conditions

We consider a transformed Gaussian model (6.2) for discretely observed values $(X_t \mid t \in \bar{D}_n^d)$. The original Gaussian process $(Y_t \mid t \in [0,1]^d)$ is defined by (6.1). The space of the unknown parameter is $\Theta = [\underline{\theta}, \overline{\theta}] \subset \mathbb{R}$. We assume the true parameter is an interior point of Θ for simplicity.

The functions β and γ in (6.1) are assumed to satisfy the following conditions:

[Y1] The function β is positive, continuous and square integrable on $(-\infty, 1]^d$.

[Y2] The function γ is positive and differentiable $d+1$ times on $[0,1]^d$.

The transformation function $g : \mathbb{R} \times [0,1]^d \times \Theta \to \mathbb{R}$ is assumed to satisfy the following regularity conditions:

[g1] For each $(t,\theta) \in [0,1]^d \times \Theta$, the function $g(\cdot; t, \theta) : \mathbb{R} \to \mathbb{R}$ is one-to-one.

[g2] For each $(i,j,k) \in \overline{[d+2]} \times \overline{[d+2]}^d \times \overline{[4]}$, g has continuous derivatives $\partial_x^i \partial_t^j \partial_\theta^k g(x;t,\theta)$ on $\mathbb{R} \times [0,1]^d \times \Theta$, where the multi-index notation $\partial_t^j = \partial_{t_1}^{j_1} \cdots \partial_{t_d}^{j_d}$ for t and j is used.

[g3] For all $(x,t,\theta) \in \mathbb{R} \times [0,1]^d \times \Theta$, $|\partial_x g(x;t,\theta)| > 0$ and $|\partial_x \partial_\theta g(x;t,\theta)| > 0$.

From the condition [g1], there exists the inverse function $g^{-1}(\cdot; t, \theta)$ of $g(\cdot; t, \theta)$ for each t and θ. The former condition in [g3] is needed for regularity of the variation of X and the latter condition is needed for non-degeneracy of the random Fisher information defined later.

The condition [g4] below is useful for the proof of Theorem 6.1 but it does not need to be assumed since a truncation method is available as discussed in Subsection 6.5.2. This fact is also used in Chan & Wood (2004).

[g4] The derivatives $\partial_x^i \partial_t^j \partial_\theta^k g(x;t,\theta)$ for all $(i,j,k) \in \overline{[d+2]} \times \overline{[d+2]}^d \times \overline{[4]} \setminus \{(0, \cdots, 0)\}$ and $1/\partial_x g$ are bounded over $\mathbb{R} \times [0,1]^d \times \Theta$.

6.3.3 Local asymptotic mixed normality

We prepare some additional notations. Let $L_n(\theta)$ be the likelihood function for the model (6.2) of X. Let $\mathcal{A}_p$ be the set of all partitions of $[d]$ into p subsets, that is,

$$\mathcal{A}_p = \{\{a_1, \cdots, a_p\} \mid \emptyset \subsetneq a_i \subset [d] \ (\forall i); \ a_i \cap a_j = \emptyset \ (\forall i \neq \forall j); \ \cup_{j=1}^p a_j = [d]\}. \quad (6.3)$$

For each $t \in [0,1]^d$ and $\theta \in \Theta$, we put

$$F_\theta(y;t,\theta) = (\partial_\theta g \circ g^{-1})(y;t,\theta) = (\partial_\theta g)(g^{-1}(y;t,\theta);t,\theta).$$

For each $a \subset [d]$ and $t \in [0,1]^d$, we put

$$q_a(t) \;=\; \gamma_t^2 \int_{(-\infty, t_{[d]\setminus a}]} \beta_u^2\big|_{u_a = t_a} \, \mathrm{d}u_{[d]\setminus a}. \tag{6.4}$$

In particular, $q_{[d]}(t) = \gamma_t^2 \beta_t^2$.

The next theorem is our main result. The proof is given in Section 6.5.

Theorem 6.1. *Assume the conditions* [Y1], [Y2] *and* [g1]–[g3]. *Then the model* (6.2) *satisfies that, for any $\theta \in \Theta$, there exist random variables ξ_n, J_n and J such that*

$$\begin{aligned} &\log L_n(\theta + \delta^{d/2} h) - \log L_n(\theta) - (h J_n \xi_n - \frac{h^2}{2} J_n) \xrightarrow{\mathrm{P}} 0, && (6.5)\\ &J_n \xrightarrow{\mathrm{P}} J, && (6.6)\\ &(\xi_n, J_n) \;\rightsquigarrow\; (\xi, J), \quad \textit{where} \quad \xi | J \sim N(0, J^{-1}), && (6.7) \end{aligned}$$

uniformly in $h \in I$ for any bounded interval $I \subset \mathbb{R}$ under P_θ. *In particular, the model is LAMN. The random Fisher information J is given by*

$$J \;=\; \int_{[0,1]^d} \left[\left(\sum_{p=1}^{d} (\partial_y^p F_\theta(Y_t; t, \theta))^2 \sum_{\{a_1, \cdots, a_p\} \in \mathcal{A}_p} \frac{q_{a_1}(t) \cdots q_{a_p}(t)}{q_{[d]}(t)} \right) + (\partial_y F_\theta(Y_t; t, \theta))^2 \right] \mathrm{d}t. \tag{6.8}$$

□

Remark 6.2. Although we assume that Θ is a subset of $\mathbb{R}$ for simplicity, a generalization of Theorem 6.1 to the case of $\Theta \subset \mathbb{R}^k$ $(k \geq 2)$ can be proved. □

Remark 6.3. Dohnal (1987) proved the LAMN property for univariate 1-parameter diffusion processes. If the stochastic differential equation is integrable, Dohnal's model results in our model with $d = 1$. Genon-Catalot & Jacod (1993, 1994) generalized Dohnal's result to multivariate 1-parameter Markov processes with random-sampling schemes. □

6.4 Examples

6.4.1 Examples of Y

We consider two examples of Gaussian processes Y and calculate the "weight"

$$\frac{q_{a_1}(t) q_{a_2}(t) \cdots q_{a_p}(t)}{q_{[d]}(t)}$$

in the expression (6.8) of the random Fisher information. Note that this quantity is always 1 when $d = 1$.

Example 6.4 (Ornstein-Uhlenbeck sheet). Let $\lambda_j > 0$ for all $j \in [d]$. If $\beta_t = \gamma_t^{-1} = \exp(\sum_{j=1}^d \lambda_j t_j)$, then Y_t is called the Ornstein-Uhlenbeck sheet (e.g. Arato et al. (2001), Ying (1993)), whose covariance matrix is $\mathrm{E}[Y_tY_s] = \prod_{j=1}^d \exp(-\lambda_j|t_j - s_j|)$. Since $q_a(t) = \prod_{j\notin a}(2\lambda_j)^{-1}$, the weight is

$$\frac{q_{a_1}(t)q_{a_2}(t)\cdots q_{a_p}(t)}{q_{[d]}(t)} = \left(\frac{1}{2^d\lambda_1\lambda_2\cdots\lambda_d}\right)^{p-1}.$$

This is independent of each partition $\{a_1, \cdots, a_p\}$ and time t. □

Example 6.5 (Brownian sheet). If $\beta_t = \mathbb{I}_{(t\succeq 0)}$ and $\gamma_t = 1$ in (6.1), then Y_t is the Brownian sheet (e.g. Khoshnevisan (2002)), whose covariance matrix is $\mathrm{E}[Y_tY_s] = \prod_{j=1}^d t_j \wedge s_j$. Since $q_a(t) = \prod_{j\notin a} t_j$, the weight is

$$\frac{q_{a_1}(t)q_{a_2}(t)\cdots q_{a_p}(t)}{q_{[d]}(t)} = (t_1t_2\cdots t_d)^{p-1}.$$

This does not depend on each partition $\{a_1, \cdots, a_p\}$ but depends on time t. In the example, the proof of Theorem 6.1 should be slightly modified since β_t does not satisfy [Y1]. However, the modification is straightforward and omitted. □

6.4.2 Examples of g

We give two examples of the function g and elucidate some features of the random Fisher information J. A quantity $J_n = -\delta^d\partial_\theta^2 \log L_n^{[d]}(\theta)$ as an approximation of J is numerically evaluated, where $L_n^{[d]}$ is the conditional likelihood of $(X_t \mid t \in D_n^d)$ given $(X_t \mid t \in \bar{D}_n^d \setminus D_n^d)$ (see Subsection 6.5.3). Since both the examples of g are independent of the time parameter t, the argument t of functions is omitted.

Example 6.6 (scale family). Put

$$g(x;\theta) = \frac{h(x)}{\sqrt{\theta}}$$

with a sufficiently smooth one-to-one known function h. Then $h(X_t)$ is a scale-transformed random field of Y_t. The function $F_\theta(y;\theta)$ used in Theorem 6.1 is $F_\theta(y;\theta) = -y/(2\theta)$. The random Fisher information $J = 1/(2\theta^2)$ is deterministic and independent of the covariance function of (Y_t). In this case, *local asymptotic normality* (LAN) holds. Figure 6.1 shows the distribution of J_n for each $n \in \{30, 100, 300, 1000\}$ by the Monte Carlo method, where the original Gaussian process is the 2-parameter Ornstein-Uhlenbeck sheet with $\lambda_1 = \lambda_2 = 1$ and the transformation function is $g(x;\theta) = x/\sqrt{\theta}$. The distributions tend to a degenerate distribution as $n \to \infty$. □

Example 6.7 (location family). Let g be a location family given by

$$g(x;\theta) \quad = \quad h(x-\theta)$$

with a sufficiently smooth one-to-one function h. The function $F_\theta(y;\theta)$ used in Theorem 6.1 is $F_\theta(y;\theta) = -h' \circ h^{-1}(y)$, where h' is the first derivative of h. The random Fisher information J has a common distribution to all $\theta \in \mathbb{R}$. Figure 6.2 shows the distribution of J_n for each $n \in \{30, 100, 300, 1000\}$ by the Monte Carlo method, where the original Gaussian process is the 2-parameter Ornstein-Uhlenbeck sheet with $\lambda_1 = \lambda_2 = 1$ and the transformation function is $g(x;\theta) = \log((\mathrm{e}^{x-\theta} + \mathrm{e}^{2(x-\theta)})/2)$. The distributions tend to a nondegenerate distribution as $n \to \infty$. □

Remark 6.8. The Box-Cox transformation

$$g(x;\theta) \quad = \quad \begin{cases} (x^\theta - 1)/\theta & \text{if } \theta \neq 0, \\ \log x & \text{if } \theta = 0 \end{cases}$$

is widely used in spatial statistics. However, it is not consistent with our result since the image of g is a proper subset of $\mathbb{R}$ when $\theta \neq 0$. Modifications that allow the Box-Cox transformation are not discussed here. □

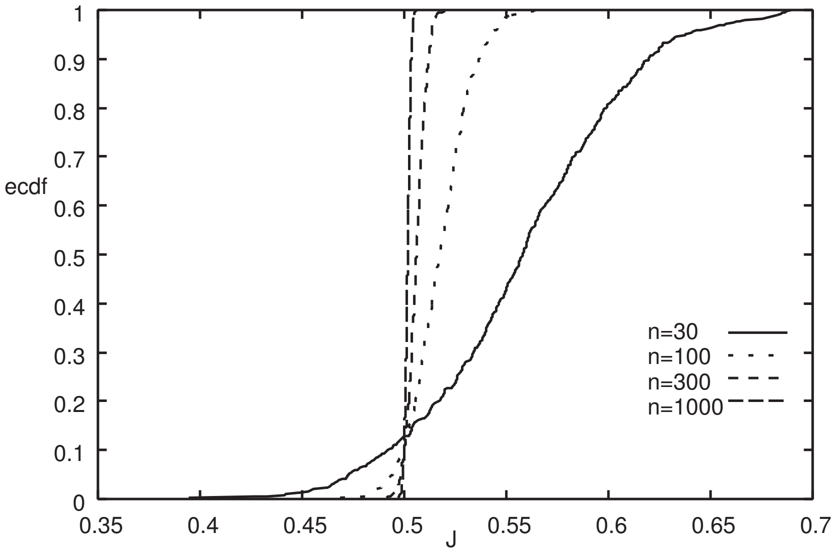


Figure 6.1: For each $n \in \{30, 100, 300, 1000\}$, an empirical cumulative distribution function (ecdf) of J_n by the Monte Carlo method is shown. The original Gaussian process is the 2-parameter Ornstein-Uhlenbeck sheet with $\lambda_1 = \lambda_2 = 1$ and the transformation function is $g(x;\theta) = x/\sqrt{\theta}$. The model is LAN. The true parameter is $\theta = 1$. The number of sampling is 500 for each n.

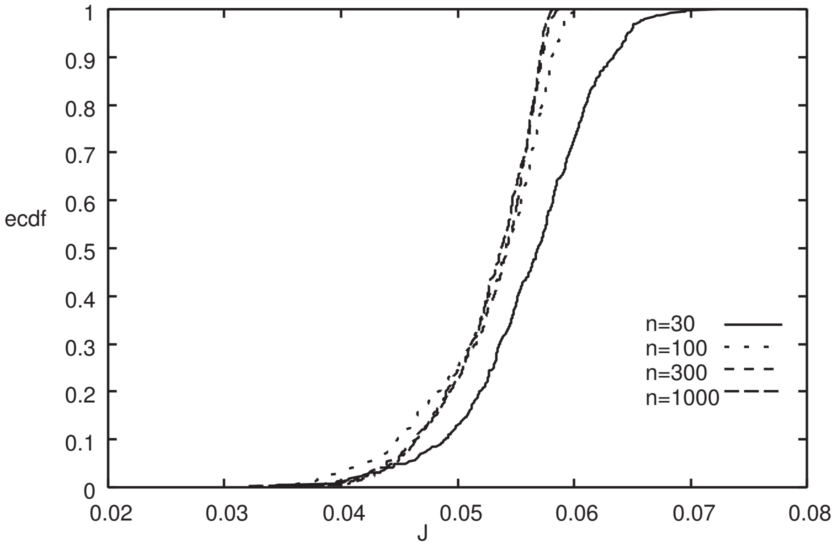


Figure 6.2: For each $n \in \{30, 100, 300, 1000\}$, an empirical cumulative distribution function (ecdf) of J_n by the Monte Carlo method is shown. The original Gaussian process is the 2-parameter Ornstein-Uhlenbeck sheet with $\lambda_1 = \lambda_2 = 1$ and the transformation function is $g(x; \theta) = \log((\mathrm{e}^{x-\theta} + \mathrm{e}^{2(x-\theta)})/2)$. The model is LAMN but not LAN. The true parameter is $\theta = 1$. The number of sampling is 500 for each n.

6.5 Proofs

Some lemmas are prepared before giving the proof of Theorem 6.1. Subsections 6.5.1-6.5.3 are devoted to simplification of the proof and Subsections 6.5.4-6.5.7 give technical lemmas.

We use a notation about rate of convergence as follows. Let $(z_n)_{n=1}^{\infty}$ be a sequence of random variables. The expression $z_n = r_2(u_n)$ ($z_n = r(u_n)$) for a positive sequence $(u_n)_{n=1}^{\infty}$ means that $u_n^{-1} z_n$ converges to 0 in L^2 (resp. in L^p for any $p \geq 1$). The expression $z_n = R_2(u_n)$ ($z_n = R(u_n)$) means that $u_n^{-1} z_n$ is L^2-bounded (resp. L^p-bounded for any $p \geq 1$). When a random sequence z_n depends on $t \in [0,1]^d$ and $\theta \in \Theta$, a statement that $z_n = r_2(u_n)$ uniformly in t and θ is simply denoted by $z_n = r_2(u_n)$. The abbreviation of stating uniformness is also used for r, R_2, R and the orders o and O.

6.5.1 Reduction to the case of $\gamma_t = 1$

We reduce the proof to the case of $\gamma_t = 1$. Assume that Theorem 6.1 was proved when $\gamma_t = 1$. Let γ_t be any positive-valued continuous function. Put $Y'_t = \gamma_t^{-1} Y_t$, $g'(x; t, \theta) = \gamma_t^{-1} g(x; t, \theta)$, $q'_a(t) = \gamma_t^{-2} q_a(t)$ and $F'_\theta(y'; t, \theta) = (\partial_\theta g' \circ (g')^{-1})(y'; t, \theta)$. The process X

defined by (6.2) is invariant under these transformations. Since g' satisfies [g1]-[g3] and $Y'_t = \int_{(-\infty,t]} \beta_s \nu(\mathrm{d}s)$ satisfies [Y1] and [Y2], the theorem for g' holds by the assumption. The random Fisher information is

$$\begin{aligned}
&\int_{[0,1]^d} \left[\sum_{p=1}^{d} ((\partial_{y'}^p F'_\theta)(Y'_t;t,\theta))^2 \sum_{\{a_1,\cdots,a_p\}\in\mathcal{A}_p} \frac{q'_{a_1}(t)\cdots q'_{a_p}(t)}{q'_{[d]}(t)} + ((\partial_{y'} F'_\theta)(Y'_t;t,\theta))^2 \right] \mathrm{d}t \\
= &\int_{[0,1]^d} \left[\sum_{p=1}^{d} \gamma_t^{2(p-1)} ((\partial_{y}^p F_\theta)(Y_t;t,\theta))^2 \sum_{\{a_1,\cdots,a_p\}\in\mathcal{A}_p} \frac{q_{a_1}(t)\cdots q_{a_p}(t)}{\gamma_t^{2(p-1)} q_{[d]}(t)} + ((\partial_{y} F_\theta)(Y_t;t,\theta))^2 \right] \mathrm{d}t \\
= &\ J,
\end{aligned}$$

since $F'_\theta(y';t,\theta) = \gamma_t^{-1} F_\theta(\gamma_t y';t,\theta)$. This means that the theorem for g also holds. Thus we assume $\gamma_t = 1$ without loss of generality.

6.5.2 Truncation

We explain a truncation method that reduces the proof of the theorem to one with the additional regularity condition [g4]. The reduction is easily done because the domain of the observed points is bounded.

Fix a function g that satisfies [g1]-[g3]. Let K be any positive number. Then there exists a function g_K that satisfies the conditions [g1]-[g4] and $g_K(x;t,\theta) = g(x;t,\theta)$ for all $(x,t,\theta) \in [-K,K] \times [0,1]^d \times \Theta$. We say that to take g_K is a truncation of g because, roughly speaking, the derivatives of g_K are truncated functions of the derivatives of g, respectively. We define a process $X^K = (X_t^K \mid t \in [0,1]^d)$ by $X_t^K = g_K^{-1}(Y_t;t,\theta)$. It holds that $X_t^K = X_t$ for any $K > \sup_{t\in[0,1]^d} |X_t|$, and such a number K exists almost surely since X has a continuous path almost surely.

Let Z_n be a sequence of measurable functionals from $\mathrm{C}([0,1]^d;\mathbb{R})$ to a metric space (S,ρ). Since

$$\lim_{K\to\infty} \limsup_{n\to\infty} \mathrm{P}[\rho(Z_n(X^K), Z_n(X)) \geq \epsilon] \ \leq \ \lim_{K\to\infty} \mathrm{P}[X^K \neq X] \ = \ 0$$

for any $\epsilon > 0$, the assumption of the following lemma holds if we put $Z_n^K = Z_n(X^K)$ and $Z_n = Z_n(X)$.

Lemma 6.9. *Let Z_n^K and Z_n be S-valued random variables. Suppose*

$$\lim_{K\to\infty} \limsup_{n\to\infty} \mathrm{P}[\rho(Z_n^K, Z_n) \geq \epsilon] \ = \ 0$$

for any $\epsilon > 0$. Then
(i) *If $Z_n^K \xrightarrow{\mathrm{P}} Z^K$ for any K and $Z^K \xrightarrow{\mathrm{P}} Z$, then $Z_n \xrightarrow{\mathrm{P}} Z$.*
(ii) *If $Z_n^K \rightsquigarrow Z^K$ for any K and $Z^K \rightsquigarrow Z$, then $Z_n \rightsquigarrow Z$.*

Proof. (i) Let $n \to \infty$ and then $K \to \infty$ in the following inequality: for any $\epsilon > 0$,

$$\mathrm{P}[\rho(Z_n, Z) \geq \epsilon] \quad \leq \quad \mathrm{P}[\rho(Z_n, Z_n^K) \geq \frac{\epsilon}{3}] + \mathrm{P}[\rho(Z_n^K, Z^K) \geq \frac{\epsilon}{3}] + \mathrm{P}[\rho(Z^K, Z) \geq \frac{\epsilon}{3}].$$

(ii) See Theorem 4.2 in Billingsley (1999). □

From the above lemma, we can assume the additional condition [g4] in order to prove Theorem 6.1. Specifically, we shall take Z_n as the left hand side of (6.5), (6.6) and (6.7), respectively.

6.5.3 Conditional likelihood function

Let $D_n^a = \{(t_j \mathbb{I}_{(j \in a)})_{j \in [d]} \mid t \in D_n^d\}$ for $a \subset [d]$. It holds that $\bar{D}_n^d = \cup_{a \subset [d]} D_n^a$ and $D_n^d = D_n^{[d]}$. The likelihood function $L_n(\theta)$ is decomposed as $L_n(\theta) = \prod_{a \subset [d]} L_n^a(\theta)$, where $L_n^a(\theta)$ is the conditional likelihood function of $(X_t \mid t \in D_n^a)$ given $(X_t \mid t \in D_n^b,\ b \subsetneq a)$. Only $L_n^{[d]}(\theta)$ affects the LAMN property. In fact, if we show

$$\log L_n^{[d]}(\theta + \delta^{d/2} h) - \log L_n^{[d]}(\theta) - (h J_n \xi_n - \frac{h^2}{2} J_n) \overset{\mathrm{P}}{\to} 0, \tag{6.9}$$

then we can also show, by the same way,

$$\log L_n^a(\theta + \delta^{\sharp a/2} h) - \log L_n^a(\theta) - (h J_n^a \xi_n^a - \frac{h^2}{2} J_n^a) \overset{\mathrm{P}}{\to} 0$$

with some tight sequences ξ_n^a and J_n^a for all nonempty $a \subsetneq [d]$. Since this convergence is uniform in h, we replace h by $\delta^{\sharp([d] \setminus a)/2} h$ to obtain

$$\log L_n^a(\theta + \delta^{d/2} h) - \log L_n^a(\theta) \quad \overset{\mathrm{P}}{\to} \quad 0$$

for all nonempty $a \subsetneq [d]$. It holds also for $a = \emptyset$ by direct calculations. Thus it suffices to show (6.9) instead of (6.5).

6.5.4 A difference operator

We define a difference operator $\Box_a$ for each subset $a \subset [d]$. Let ϕ be any real-valued function on $\bar{D}_n^d$. For each $t \in D_n^d$, we put

$$\Box_a \phi_t \quad = \quad \sum_{b \subset a} (-1)^{\sharp(a \setminus b)} \phi_{t - \delta + \delta_b}.$$

The inclusion-exclusion formula holds:

$$\sum_{a \subset b} \Box_a \phi_t \quad = \quad \phi_{t - \delta + \delta_b}.$$

The operator $\Box_a$ is useful to describe Taylor's expansion as shown in the next section. In particular, $\Box_{\{j\}}\phi_t$ for $j \in [d]$ is the partial difference of ϕ_t along j-th axis and $\Box_{[d]}\phi_t$ is the increment of ϕ_t (Khoshnevisan, 2002, p.40). For example, if $d = 3$ and $a = \{1, 2\}$, then

$$\Box_{\{1,2\}}\phi_{(t_1,t_2,t_3)} = \phi_{(t_1,t_2,t_3-\delta)} - \phi_{(t_1,t_2-\delta,t_3-\delta)} - \phi_{(t_1-\delta,t_2,t_3-\delta)} + \phi_{(t_1-\delta,t_2-\delta,t_3-\delta)}.$$

The domain whose volume is $\Box_a\phi_t$ is shown in Figure 6.3 when $\phi_t = \mathrm{Leb}([0, t])$.

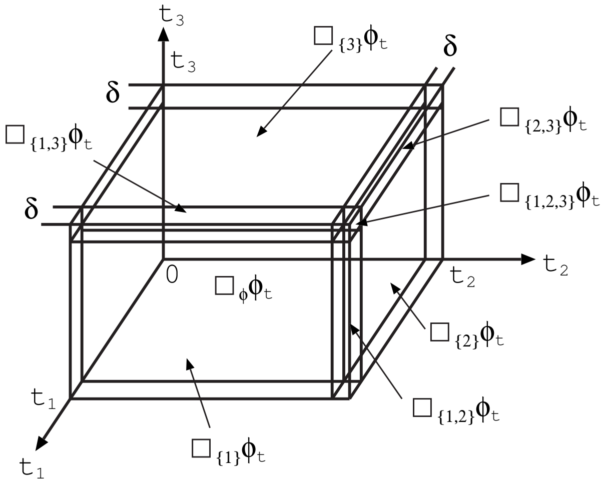


Figure 6.3: The domain whose volume is $\Box_a\phi_t$ is shown, where $\phi_t = \mathrm{Leb}([0, t])$.

Let $I_{a,t}$ be a rectangular set

$$I_{a,t} = \{u \mid u_j \in (t_j - \delta, t_j] \text{ for } j \in a,\ u_j \in (-\infty, t_j - \delta] \text{ for } j \notin a\}.$$

Put

$$\tilde{q}_a(t) = \delta^{-\sharp a} \int_{I_{a,t}} \beta_s^2 \mathrm{d}s.$$

The quantity $\tilde{q}_a(t)$ is approximated by $q_a(t)$ uniformly in t, where $q_a(t)$ is defined by (6.4) with $\gamma_t = 1$.

For the Gaussian process Y, the next lemma holds.

Lemma 6.10. *Let Y be a Gaussian process given by* (6.1) *with $\gamma_t = 1$. Fix a point t in D_n^d. Then $\Box_a Y_t$ ($a \subset [d]$) are independently distributed and $\Box_a Y_t \sim N(0, \delta^{\sharp a}\tilde{q}_a(t))$. In particular, $\Box_a Y_t = R(\delta^{\sharp a/2})$.*

Proof. A formula $\Box_a Y_t = \int_{I_{a,t}} \beta_u \nu(\mathrm{d}u)$ holds. The independence follows from the fact that $I_{a,t}$ for $a \subset [d]$ are disjoint subsets. The L^p-boundedness of $(\delta^{-\sharp a/2}\Box_a Y_t)$ holds since $\tilde{q}_a(t)$ is approximated by $q_a(t)$ uniformly in $t \in [0, 1]^d$. □

6.5.5 Taylor's expansion

Lemma 6.11. *Let Y be a Gaussian process* (6.1). *Let f be a function from $(y,t) \in \mathbb{R} \times [0,1]^d$ to $f(y,t) \in \mathbb{R}$ with continuous derivatives $\partial_y^p \partial_t^q f(y,t)$ for $p \in \overline{[d+1]}$ and $q \in \overline{[d+1]}^d$. Assume that $\partial_y^p \partial_t^q f(y,t)$ for $p + \sum_j q_j = d+1$ are bounded over $\mathbb{R} \times [0,1]^d$. Then, for any $k \geq 1$,*

$$\Box_{[d]} f(Y_t,t) \;=\; \sum_{d=1}^{p} (\partial_y^p f)(Y_{t-\delta}, t-\delta) \sum_{\{a_1,\cdots,a_p\}\in\mathcal{A}_p} \prod_{j=1}^{p} (\Box_{a_j} Y_t) + r(\delta^{d/2}),$$

where $\mathcal{A}_p$ is defined by (6.3). *In particular, $\Box_{[d]} f(Y_t,t) = R(\delta^{d/2})$.*

Proof. We prove only the case that f is independent of $t \in [0,1]^d$. The dependent case is similarly proved. Put $\Delta_{\{a_1,\cdots,a_p\}} = \prod_{j=1}^p (\Box_{a_j} Y_t)$. From the inclusion-exclusion formula, Taylor's expansion, Lemma 6.10 and boundedness of the derivatives of f,

$$\begin{aligned}
f(Y_{t-\delta+\delta_a}) &= f(Y_{t-\delta} + \sum_{\emptyset\subsetneq b\subset a} \Box_b Y_t) \\
&= f(Y_{t-\delta}) + \sum_{p=1}^{d} \frac{f^{(p)}(Y_{t-\delta})}{p!} \left(\sum_{\emptyset\subsetneq b\subset a} (\Box_b Y_t) \right)^p + r(\delta^{d/2}) \\
&= f(Y_{t-\delta}) + \sum_{p=1}^{d} \frac{f^{(p)}(Y_{t-\delta})}{p!} \sum_{\substack{a_1,\cdots,a_p\subset a,\\ \sharp a_1+\cdots+\sharp a_p\leq d}} C_{\{a_1,\cdots,a_p\}} \Delta_{\{a_1,\cdots,a_p\}} + r(\delta^{d/2}),
\end{aligned}$$

where $C_{\{a_1,\cdots,a_p\}}$ is number of the term $\Delta_{\{a_1,\cdots,a_p\}}$. In particular, $C_{\{a_1,\cdots,a_p\}} = p!$ if $\{a_1,\cdots,a_p\} \in \mathcal{A}_p$. Then, $\Box_{[d]} f$ is calculated as

$$\begin{aligned}
\Box_{[d]} f(Y_t) &= \sum_{a\subset[d]} (-1)^{\sharp([d]\setminus a)} f(Y_{t-\delta}) + S(f) + r(\delta^{d/2}) \\
&= S(f) + r(\delta^{d/2}),
\end{aligned}$$

where

$$\begin{aligned}
S(f) &= \sum_{a\subset[d]}(-1)^{\sharp([d]\setminus a)}\sum_{p=1}^{d}\frac{f^{(p)}(Y_{t-\delta})}{p!}\sum_{\substack{a_1,\cdots,a_p\subset a,\\ \sharp a_1+\cdots+\sharp a_p\le d}} C_{\{a_1,\cdots,a_p\}}\Delta_{\{a_1,\cdots,a_p\}} \\
&= \sum_{p=1}^{d}\frac{f^{(p)}(Y_{t-\delta})}{p!}\sum_{\substack{a_1,\cdots,a_p\subset[d],\\ \sharp a_1+\cdots+\sharp a_p\le d}} C_{\{a_1,\cdots,a_p\}}\Delta_{\{a_1,\cdots,a_p\}}\sum_{b\subset[d]\setminus(a_1\cup\cdots\cup a_p)}(-1)^{\sharp b} \\
&= \sum_{p=1}^{d}\frac{f^{(p)}(Y_{t-\delta})}{p!}\sum_{\substack{a_1,\cdots,a_p\subset[d],\\ \sharp a_1+\cdots+\sharp a_p\le d}} C_{\{a_1,\cdots,a_p\}}\Delta_{\{a_1,\cdots,a_p\}}\mathbb{I}_{([d]\setminus(a_1\cup\cdots\cup a_p)=\emptyset)} \\
&= \sum_{p=1}^{d} f^{(p)}(Y_{t-\delta})\sum_{\{a_1,\cdots,a_p\}\in\mathcal{A}_p}\Delta_{\{a_1,\cdots,a_p\}}.
\end{aligned}$$

Thus the lemma follows. □

6.5.6 Commuting filtration

For each $t\in D_n^d$, we define a σ-field

$$\mathcal{F}_t \;=\; \mathcal{F}_t^n \;=\; \sigma(Y_s : s\in\bar{D}_n^d,\ s\preceq t).$$

Lemma 6.12. *The set of σ-fields $(\mathcal{F}_t \mid t\in\bar{D}_n^d)$ forms a d-parameter commuting filtration, that is, it holds that*

$$\begin{aligned}
&(1) \qquad s\preceq t \implies \mathcal{F}_s\subset\mathcal{F}_t, \\
&(2) \qquad \forall s,t\in\bar{D}_n^d;\quad \mathrm{E}[\mathrm{E}[\cdot|\mathcal{F}_t]|\mathcal{F}_s]=\mathrm{E}[\mathrm{E}[\cdot|\mathcal{F}_s]|\mathcal{F}_t]=\mathrm{E}[\cdot|\mathcal{F}_{s\curlywedge t}].
\end{aligned}$$

Proof. See Theorem 2.4.1 of Chapter 7 in Khoshnevisan (2002). □

6.5.7 A lemma on convergence in L^2

Lemma 6.13. *Let $(\Omega,\mathcal{G},P,(\mathcal{G}_i)_{1\le i\le n})$ be a 1-parameter filtered probability space and $(\chi_i)_{i=1}^n$ be a $(\mathcal{G}_i)$-adapted process. If*

$$\sum_{i=1}^{n}\mathrm{E}[\chi_i|\mathcal{G}_{i-1}] \;\overset{L^2}{\to}\; Z \tag{6.10}$$

and

$$\chi_i \;=\; r_2(\delta^{1/2}), \tag{6.11}$$

then

$$\sum_{i=1}^{n} \chi_i \ \ \overset{L^2}{\to} \ \ Z.$$

Proof. Put $\xi_i = \chi_i - \mathrm{E}[\chi_i|\mathcal{G}_{i-1}]$. Then $(\xi_i)_{i=1}^n$ is a martingale difference array. Since $\mathrm{E}[(\xi_i)^2] \le \mathrm{E}[(\chi_i)^2] = \mathrm{o}(\delta)$,

$$\mathrm{E}[(\sum_{i=1}^{n} \xi_i)^2] \ = \ \sum_{i=1}^{n} \mathrm{E}[(\xi_i)^2] \ = \ \sum_{i=1}^{n} \mathrm{o}(\delta) \ = \ \mathrm{o}(1).$$

Thus $\sum_{i=1}^n \xi_i \overset{L^2}{\to} 0$. This implies $\sum_{i=1}^n \chi_i \overset{L^2}{\to} Z$. □

The next lemma is a multiparameter version of Lemma 6.13.

Lemma 6.14. *Let $(\Omega, \mathcal{F}, P, (\mathcal{F}_t)_{t \in \bar{D}_n^d})$ be a multiparameter filtered probability space and (χ_t) be a $(\mathcal{F}_t)$-adapted process. Suppose that the filtration $(\mathcal{F}_t)$ is commuting. If*

$$\sum_{t \in D_n^d} \mathrm{E}[\chi_t|\mathcal{F}_{t-\delta}] \ \ \overset{L^2}{\to} \ \ Z \tag{6.12}$$

and

$$\chi_t \ \ = \ \ r_2(\delta^{d-1/2}), \tag{6.13}$$

then

$$\sum_{t \in D_n^d} \chi_t \ \ \overset{L^2}{\to} \ \ Z.$$

Proof. For each $j \in [d]$, put

$$\begin{aligned} \xi_t^{(j)} \ &= \ \mathrm{E}[\chi_t|\mathcal{F}_{t-\delta_{[j-1]}}] - \mathrm{E}[\chi_t|\mathcal{F}_{t-\delta_{[j]}}], \\ \eta_{t_j}^{(j)} \ &= \ \sum_{s \in D_n^d : s_j = t_j} \xi_s^{(j)}, \\ \mathcal{F}_{t_j}^{(j)} \ &= \ \bigvee_{s \in \bar{D}_n^d : s_j \preceq t_j} \mathcal{F}_s \end{aligned}$$

(see Figure 6.4). The filtration $\mathcal{F}_{t_j}^{(j)}$ is called a marginal filtration. The random variables $\eta_{t_j}^{(j)}$ are not symmetrically defined with respect to j. The next decomposition of χ_t holds:

$$\chi_t \ = \ \mathrm{E}[\chi_t|\mathcal{F}_{t-\delta}] + \sum_{j=1}^{d} \xi_t^{(j)}.$$

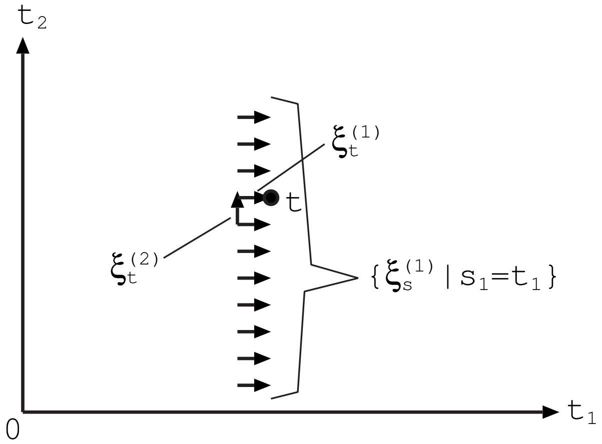


Figure 6.4: The definition of $\xi_t^{(j)}$. The arrows denote the martingale differences $\xi_t^{(j)}$. The sum of $\{\xi_s^{(1)} \mid s_1 = t_1\}$ is $\eta_{t_1}^{(1)}$.

Thus

$$\sum_t \chi_t \;=\; \sum_t \mathrm{E}[\chi_t|\mathcal{F}_{t-\delta}] + \sum_{j=1}^{d} \sum_{t_j \in D_n} \eta_{t_j}^{(j)}. \tag{6.14}$$

The first term on the right side of (6.14) converges in probability to Z by the assumption (6.12). Thus it suffices to show that

$$\sum_{t_j \in D_n} \eta_{t_j}^{(j)} \;\xrightarrow{L^2}\; 0$$

for each $j \in [d]$. To do this, we use Lemma 6.13. By the commuting property,

$$\begin{aligned}
\mathrm{E}[\xi_t^{(j)}|\mathcal{F}_{t_j-\delta}^{(j)}] &= \mathrm{E}\left[\mathrm{E}[\chi_t|\mathcal{F}_{t-\delta_{[j-1]}}] - \mathrm{E}[\chi_t|\mathcal{F}_{t-\delta_{[j]}}]\,\middle|\, \mathcal{F}_{t_j-\delta}^{(j)}\right] \\
&= \mathrm{E}\left[\chi_t|\mathcal{F}_{t-\delta_{[j]}}\right] - \mathrm{E}\left[\chi_t|\mathcal{F}_{t-\delta_{[j]}}\right] \\
&= 0.
\end{aligned}$$

This implies that

$$\mathrm{E}[\eta_{t_j}^{(j)}|\mathcal{F}_{t_j-\delta}^{(j)}] \;=\; \sum_{s:s_j=t_j} \mathrm{E}[\xi_s^{(j)}|\mathcal{F}_{s_j-\delta}^{(j)}] \;=\; 0.$$

Therefore the first condition (6.10) of Lemma 6.13 is satisfied with $Z = 0$. Next, by the assumption (6.13),

$$\mathrm{E}[(\eta_{t_j}^{(j)})^2] \;=\; \sum_{s:s_j=t_j} \sum_{u:u_j=t_j} \mathrm{E}[\xi_s^{(j)}\xi_u^{(j)}] \;=\; n^{2d-2}\mathrm{o}(\delta^{2d-1}) \;=\; \mathrm{o}(\delta).$$

Thus the second condition (6.11) of Lemma 6.13 is satisfied. □

6.5.8 Proof of Theorem 6.1

We abbreviate $\Box_{[d]}$ to $\Box$. For any function $\phi : (x,t,\theta) \mapsto \phi(x,t,\theta)$, $\phi(X_t,t,\theta)$ is abbreviated to ϕ whenever there is no confusion. We assume $\gamma_t = 1$ because of the reason mentioned in Subsection 6.5.1.

First we derive expression of the conditional likelihood function introduced in Subsection 6.5.3. The conditional density function of the Gaussian process $(Y_t \mid t \in D_n^d)$ given $(Y_t \mid t \in \bar{D}_n^d \setminus D_n^d)$ is

$$\prod_{t\in D_n^d} \frac{1}{\sqrt{2\pi\delta^d \tilde{q}}} \exp\left[-\frac{1}{2\delta^d\tilde{q}}\left(\Box Y\right)^2\right],$$

where $\tilde{q} = \tilde{q}_{[d]}(t)$. The definition of $\tilde{q}_a(t)$ is in Subsection 6.5.4. The conditional likelihood function $L_n^{[d]}(\theta)$ is explicitly given by

$$L_n^{[d]}(\theta) \;=\; \prod_{t\in D_n^d} \frac{\partial_x g}{\sqrt{2\pi\delta^d \tilde{q}}} \exp\left[-\frac{1}{2\delta^d\tilde{q}}\left(\Box g\right)^2\right].$$

The conditional log likelihood function ℓ_n is expressed as

$$\ell_n \;=\; \ell_n(\theta) \;=\; \log L_n^{[d]}(\theta) \;=\; \sum_{t\in D_n^d}\left[-\frac{(\Box g)^2}{2\delta^d\tilde{q}} + \log \partial_x g\right]. \tag{6.15}$$

By differentiating ℓ_n,

$$\begin{aligned}
\partial_\theta \ell_n &= \sum_{t\in D_n^d}\left[-\frac{(\Box g)(\Box\partial_\theta g)}{\delta^d\tilde{q}} + \frac{\partial_\theta\partial_x g}{\partial_x g}\right], & (6.16)\\
\partial_\theta^2 \ell_n &= \sum_{t\in D_n^d}\left[-\frac{(\Box\partial_\theta g)^2}{\delta^d\tilde{q}} - \frac{(\Box g)(\Box\partial_\theta^2 g)}{\delta^d\tilde{q}} + \frac{\partial_\theta^2\partial_x g}{\partial_x g} - \left(\frac{\partial_\theta\partial_x g}{\partial_x g}\right)^2\right], & (6.17)\\
\partial_\theta^3 \ell_n &= \sum_{t\in D_n^d}\left[-\frac{3(\Box\partial_\theta g)(\Box\partial_\theta^2 g)}{\delta^d\tilde{q}} - \frac{(\Box g)(\Box\partial_\theta^3 g)}{\delta^d\tilde{q}} + \partial_\theta^2\left(\frac{\partial_\theta\partial_x g}{\partial_x g}\right)\right]. & (6.18)
\end{aligned}$$

It suffices to prove (6.9), (6.6) and (6.7) with

$$\begin{aligned}
J_n &= -\delta^d\partial_\theta^2\ell_n,\\
\xi_n &= (J_n)^{-1}\delta^{\frac{d}{2}}\partial_\theta\ell_n.
\end{aligned}$$

Since the quantities ℓ_n, J_n and ξ_n are measurable functions of the path X, we assume the additional condition [g4] without loss of generality by Lemma 6.9.

<u>(6.9): $\ell_n(\theta+\delta^{d/2}h) - \ell_n(\theta) = hJ_n\xi_n - \frac{h^2}{2}J_n + \mathrm{o}_\mathrm{P}(1)$.</u>

Proof. By Taylor's formula, it suffices to prove that

$$\sup_{h:|h|\leq M} \delta^d|\partial_\theta^2\ell_n(\theta+\delta^{\frac{d}{2}}h)-\partial_\theta^2\ell_n(\theta)| \ \overset{\mathrm{P}}{\to}\ 0 \tag{6.19}$$

for all $M>0$ under P_θ. By using Taylor's formula again, we obtain

$$\delta^d[\partial_\theta^2\ell_n(\theta+\delta^{\frac{d}{2}}h)-\partial_\theta^2\ell_n(\theta)] \ = \ h\delta^{\frac{3d}{2}}\partial_\theta^3\ell_n\big|_{\theta+\psi\delta^{\frac{d}{2}}h},$$

where ψ is some $(0,1)$-valued random variable. From the condition [g4] and Lemma 6.11,

$$\left[-\frac{3(\Box\partial_\theta g)(\Box\partial_\theta^2 g)}{\delta^d\tilde{q}}-\frac{(\Box g)(\Box\partial_\theta^3 g)}{\delta^d\tilde{q}}+\partial_\theta^2\left(\frac{\partial_\theta\partial_x g}{\partial_x g}\right)\right] \ = \ R(1).$$

Thus, by (6.18),

$$\left|\delta^{\frac{3d}{2}}\partial_\theta^3\ell_n\right|_{\theta+\psi\delta^{\frac{d}{2}}h} \ = \ r(1)$$

uniformly in h, and (6.9) is proved.

(6.6): $J_n\overset{\mathrm{P}}{\to}J$.

Proof. The next formula is useful:

$$\partial_y(f\circ g^{-1}) \ = \ (\partial_x f/\partial_x g)\circ g^{-1} \tag{6.20}$$

for any function $f : x\mapsto f(x)$. By multiplying $-\delta^d$ to (6.17), substituting $X_t = g^{-1}(Y_t;t,\theta)$ and applying the formula (6.20), we obtain

$$-\delta^d\partial_\theta^2\ell_n \ = \ \sum_{t\in D_n^d}\left[\tilde{q}^{-1}(\Box F_\theta)^2+\tilde{q}^{-1}(\Box F)(\Box F_{\theta\theta})-\delta^d F_{\theta\theta}^{(1)}+\delta^d(F_\theta^{(1)})^2\right], \tag{6.21}$$

where we put $F=g\circ g^{-1}=\mathrm{id}$, $F_\theta=(\partial_\theta g)\circ g^{-1}$, $F_{\theta\theta}=(\partial_\theta^2 g)\circ g^{-1}$ and $f^{(p)}=\partial_y^p f$ for any function f, and we omit the arguments Y_t, t and θ. By using the formula (6.20) recursively and the condition [g4], it is shown that the derivatives $\partial_y^p\partial_t^q F_\theta$ and $\partial_y^p\partial_t^q F_{\theta\theta}$ for any $(p,q)\in\overline{[d{+}1]}\times\overline{[d{+}1]}^d$ are bounded over the region $\mathbb{R}\times[0,1]^d\times\Theta$. Therefore Lemma 6.11 implies

$$\begin{aligned}
\Box F \ &= \ \Delta_{\{[d]\}},\\
\Box F_\theta \ &= \ \sum_{p=1}^d F_{\theta-}^{(p)}\sum_{\{a_1,\cdots,a_p\}\in\mathcal{A}_p}\Delta_{\{a_1,\cdots,a_p\}}+r(\delta^{d/2}),\\
\Box F_{\theta\theta} \ &= \ \sum_{p=1}^d F_{\theta\theta-}^{(p)}\sum_{\{a_1,\cdots,a_p\}\in\mathcal{A}_p}\Delta_{\{a_1,\cdots,a_p\}}+r(\delta^{d/2}),
\end{aligned}$$

where $F^{(p)}_{\theta-}$ and $F^{(p)}_{\theta\theta-}$ are abbreviations of $F^{(p)}_{\theta}(Y_{t-\delta})$ and $F^{(p)}_{\theta\theta}(Y_{t-\delta})$, respectively, and $\Delta_{\{a_1,\cdots,a_p\}} = \prod_{j=1}^{p}(\Box_{a_j} Y_t)$. By the relation $Y_t - Y_{t-\delta} = r(1)$, formulas

$$\begin{aligned} F^{(1)}_{\theta\theta} &= F^{(1)}_{\theta\theta-} + r(1), \\ F^{(1)}_{\theta} &= F^{(1)}_{\theta-} + r(1) \end{aligned}$$

hold. By substituting these into (6.21), we obtain

$$-\delta^d \partial_\theta^2 \ell_n = \sum_{t\in D_n^d} \chi_t + r(1),$$

where

$$\begin{aligned} \chi_t = {} & \sum_{p,q=1}^{d} \tilde{q}^{-1} F^{(p)}_{\theta-} F^{(q)}_{\theta-} \sum_{\{a_1,\cdots,a_p\}\in\mathcal{A}_p} \sum_{\{b_1,\cdots,b_q\}\in\mathcal{A}_q} \Delta_{\{a_1,\cdots,a_p\}} \Delta_{\{b_1,\cdots,b_q\}} \\ & + \sum_{r=1}^{d} \tilde{q}^{-1} F^{(r)}_{\theta\theta-} \Delta_{\{[d]\}} \sum_{\{c_1,\cdots,c_r\}\in\mathcal{A}_p} \Delta_{\{c_1,\cdots,c_r\}} - \delta^d F^{(1)}_{\theta\theta-} + \delta^d (F^{(1)}_{\theta-})^2. \end{aligned}$$

In the following, we use Lemma 6.14 to show $\sum \chi_t \xrightarrow{\mathrm{P}} J$. The symbols used below follow from ones in Lemma 6.14. From Lemma 6.10,

$$\begin{aligned} \mathrm{E}[\Delta_{\{a_1,\cdots,a_p\}} \Delta_{\{b_1,\cdots,b_q\}} | \mathcal{F}_{t-\delta}] &= \mathbb{I}_{(\{a_1,\cdots,a_p\}=\{b_1,\cdots,b_q\})} \prod_{r=1}^{p} \mathrm{E}[(\Box_{a_r} Y_t)^2] \\ &= \mathbb{I}_{(\{a_1,\cdots,a_p\}=\{b_1,\cdots,b_q\})} \delta^d \prod_{r=1}^{p} \tilde{q}_{a_r}(t) \end{aligned}$$

This implies

$$\begin{aligned} \sum_{t\in D_n^d} \mathrm{E}[\chi_t | \mathcal{F}_{t-\delta}] &= \sum_{t\in D_n^d} \delta^d \left(\tilde{q}^{-1} \sum_{p=1}^{d} (F^{(p)}_{\theta-})^2 \prod_{r=1}^{p} \tilde{q}_{a_r}(t) + (F^{(1)}_{\theta-})^2 \right) + r(1) \\ &\overset{L^2}{\to} \int_{t\in[0,1]^d} \left(q_{[d]}(t)^{-1} \sum_{p=1}^{d} (F^{(p)}_{\theta}(Y_t; t, \theta))^2 \prod_{r=1}^{p} q_{a_r}(t) + (F^{(1)}_{\theta}(Y_t; t, \theta))^2 \right) \mathrm{d}t \\ &= J, \end{aligned}$$

where L^2-convergence comes from almost sure convergence of the Riemannian sum and L^2-boundedness of it. On the other hand, the relation $\chi_t = R(\delta^d)$ holds because of the condition [g4] and Lemma 6.10. Therefore (6.13) is satisfied. Thus $\sum \chi_t \xrightarrow{L^2} J$.

$\underline{(6.7):\ \ (\xi_n, J_n) \rightsquigarrow (\xi, J).}$

Proof. Fix $h \in \mathbb{R}$. It suffices to show the contiguity of $\mathrm{P}_{\theta+\delta^{d/2}h}$ to P_θ and the convergence $J_n \xrightarrow{\mathrm{P}} J$ under $\mathrm{P}_{\theta+\delta^{d/2}h}$ (see Lemma 3 of Section 6.6 in Le Cam & Yang (2000)).

We first prove the convergence $J_n \xrightarrow{\mathrm{P}} J$ under $\mathrm{P}_{\theta+\delta^{d/2}h}$ by assuming the contiguity of $\mathrm{P}_{\theta+\delta^{d/2}h}$ to P_θ. By replacing θ in the proof of (6.6) with $\theta + \delta^{d/2}h$ and using a fact that $F^{(p)}_{\theta+\delta^{d/2}h} = F^{(p)}_\theta + r(1)$, one shows $-\delta^d \partial^2_\theta \ell_n(\theta + \delta^{d/2}h) \xrightarrow{\mathrm{P}} J$ under $\mathrm{P}_{\theta+\delta^{d/2}h}$. On the other hand, we obtain $| - \delta^d\partial^2_\theta\ell_n(\theta + \delta^{d/2}h) - J_n| \xrightarrow{\mathrm{P}} 0$ under $\mathrm{P}_{\theta+\delta^{d/2}h}$ due to (6.19) and the contiguity of $\mathrm{P}_{\theta+\delta^{d/2}h}$ to P_θ. Thus $J_n \xrightarrow{\mathrm{P}} J$ under $\mathrm{P}_{\theta+\delta^{d/2}h}$.

Next we prove the contiguity according to the outline of the proof of Corollary 3 in Genon-Catalot & Jacod (1994). For $\theta, \theta' \in \Theta$ and $\alpha \in (0, 1)$, the Hellinger process $h(\alpha; \theta, \theta')^n = \{h(\alpha; \theta, \theta')^n_t \mid t \in D^d_n\}$ is defined by

$$h(\alpha;\theta,\theta')^n_t \;=\; \sum_{s \lhd t} \left(1 - \mathrm{E}_{\theta'}[p^\alpha_{s,\theta} p^{-\alpha}_{s,\theta'} | \mathcal{H}_s]\right).$$

Here $\lhd$ is a total order on D^d_n with a property that $s \preceq t$ implies $s \lhd t$ (e.g. any lexicographic order), $\mathcal{H}_s$ is a filtration $\bigvee_{u \lhd s, u \neq s} \mathcal{F}_u$ with respect to $\lhd$ (see Figure 6.5), and $p_{s,\theta}$ is the conditional density function of X_s given $\mathcal{H}_s$. From the commuting property, the Hellinger process at $t = (1, \cdots, 1)$ is written as

$$h(\alpha;\theta,\theta')^n_1 \;=\; \sum_{s \in D^d_n} \left(1 - \mathrm{E}_{\theta'}[p^\alpha_{s,\theta} p^{-\alpha}_{s,\theta'} | \mathcal{F}^-_s]\right), \quad \mathcal{F}^-_s = \bigvee_{j=1}^d \mathcal{F}_{s-\delta_{\{j\}}}$$

without use of $\lhd$. If one shows that

$$\limsup_{\alpha\downarrow 0} \limsup_{n\to\infty} \mathrm{P}_{\theta+\delta^{d/2}h}[h(\alpha;\theta,\theta+\delta^{d/2}h)^n_1 > \eta] \;=\; 0 \tag{6.22}$$

for any $\eta > 0$, then the contiguity of $\mathrm{P}_{\theta+\delta^{d/2}h}$ to P_θ follows from (Jacod & Shiryaev, 1987, Theorem V.2.27). It suffices to show

$$h(\alpha;\theta,\theta+\delta^{d/2}h)^n_1 \;\xrightarrow{\mathrm{P}}\; \frac{\alpha(1-\alpha)h^2}{2} J \tag{6.23}$$

under $\mathrm{P}_{\theta+\delta^{d/2}h}$. We put $\theta' = \theta + \delta^{d/2}h$. In the following, any appropriate constant independent of n is denoted by C and any appropriate $[-1, 1]$-valued random variable is denoted by $\psi = \psi(\theta, h, t)$. Although they should be denoted as C_i $(i = 1, 2, \cdots)$ and ψ_i $(i = 1, 2, \cdots)$, the indices are abbreviated for simplicity. From the definition,

$$p_{t,\theta} \;=\; \frac{\partial_x g_\theta}{\sqrt{2\pi\delta^d\tilde{q}}} \exp\left[-\frac{(\Box g_\theta)^2}{2\delta^d\tilde{q}}\right],$$

where g_θ is an abbreviation of $g(X_t; t, \theta)$. Thus

$$\begin{aligned}
&\mathrm{E}_{\theta'}[1 - p^\alpha_{t,\theta} p^{-\alpha}_{t,\theta'} | \mathcal{F}^-_t] \\
&= \; 1 - \mathrm{E}\left[\exp\left(-\frac{\alpha(\Box g_\theta \circ g^{-1}_{\theta'})^2}{2\delta^d\tilde{q}} + \frac{\alpha(\Box Y)^2}{2\delta^d\tilde{q}} + \alpha \log \frac{\partial_x g_\theta}{\partial_x g_{\theta'}} \circ g^{-1}_{\theta'}\right)\middle| \mathcal{F}^-_t\right]. \\
&= \; 1 - \mathrm{E}[\exp(K)|\mathcal{F}^-_t],
\end{aligned} \tag{6.24}$$

where

$$K = -\frac{\alpha(\square g_\theta \circ g_{\theta'}^{-1})^2}{2\delta^d\tilde{q}} + \frac{\alpha(\square Y)^2}{2\delta^d\tilde{q}} + \alpha \log\frac{\partial_x g_\theta}{\partial_x g_{\theta'}} \circ g_{\theta'}^{-1}.$$

Taylor's expansion of $\exp(K)$ is

$$\exp(K) = 1 + K + \frac{K^2}{2} + \frac{K^3 \mathrm{e}^{\psi K}}{6}. \tag{6.25}$$

The conditional expectation of each term is evaluated as follows. We first expand K with respect to θ. By Taylor's formula and the condition [g4],

$$\begin{aligned}
\square g_\theta \circ g_{\theta'}^{-1} &= \square g_{\theta'-\delta^{d/2}h} \circ g_{\theta'}^{-1} \\
&= \square Y - \delta^{d/2}h\square F_{\theta'} + \frac{\delta^d h^2}{2}\square F_{\theta'\theta'} - \frac{\delta^{3d/2}h^3}{6}\square F_{\theta'\theta'\theta'} + C\psi\delta^{2d} \\
\log\frac{\partial_x g_\theta}{\partial_x g_{\theta'}} \circ g_{\theta'}^{-1} &= -\delta^{d/2}hF_{\theta'}^{(1)} + \frac{\delta^d h^2}{2}F_{\theta'\theta'}^{(1)} - \frac{\delta^d h^2}{2}(F_{\theta'}^{(1)})^2 + C\psi\delta^{3d/2}.
\end{aligned}$$

Since $\square F_{\theta'}$, $\square F_{\theta'\theta'}$ and $\square F_{\theta'\theta'\theta'}$ are $R(\delta^{d/2})$ (Lemma 6.11),

$$\begin{aligned}
K &= -\frac{\alpha(\square g_\theta \circ g_{\theta'}^{-1})^2}{2\delta^d\tilde{q}} + \frac{\alpha(\square Y)^2}{2\delta^d\tilde{q}} + \alpha \log\frac{\partial_x g_\theta}{\partial_x g_{\theta'}} \circ g_{\theta'}^{-1} \\
&= -\frac{\alpha}{2\delta^d\tilde{q}}\left(\square Y - \delta^{d/2}h\square F_{\theta'} + \frac{\delta^d h^2}{2}\square F_{\theta'\theta'} - \frac{\delta^{3d/2}h^3}{6}\square F_{\theta'\theta'\theta'} + C\psi\delta^{2d}\right)^2 \\
&\quad + \frac{\alpha(\square Y)^2}{2\delta^d\tilde{q}} + \alpha\left(-\delta^{d/2}hF_{\theta'}^{(1)} + \frac{\delta^d h^2}{2}F_{\theta'\theta'}^{(1)} - \frac{\delta^d h^2}{2}(F_{\theta'}^{(1)})^2 + C\psi\delta^{3d/2}\right) \qquad (6.26) \\
&= \frac{\alpha h}{\delta^{d/2}\tilde{q}}(\square Y)(\square F_{\theta'}) - \frac{\alpha h^2}{2\tilde{q}}(\square Y)(\square F_{\theta'\theta'}) - \frac{\alpha h^2}{2\tilde{q}}(\square F_{\theta'})^2 \\
&\quad -\alpha\delta^{d/2}hF_{\theta'}^{(1)} + \frac{\alpha\delta^d h^2}{2}F_{\theta'\theta'}^{(1)} - \frac{\alpha\delta^d h^2}{2}(F_{\theta'}^{(1)})^2 + r(\delta^d). \qquad (6.27)
\end{aligned}$$

Let us evaluate $\mathrm{E}[K|\mathcal{F}_t^-]$. If we put $Y_t^- = Y_t - \square Y_t$ and abbreviate $\mathcal{F}_t^-$-measurable $[-1,1]$-valued random variables by ψ^-, then

$$\begin{aligned}
\square F_{\theta'} &= \sum_{a \subsetneq [d]} (-1)^{\sharp([d]\backslash a)} F_{\theta'}(Y_{t-\delta+\delta_a}) + F_{\theta'}(Y_t) \\
&= C\psi^- + F_{\theta'}(Y_t) \\
&= C\psi^- + F_{\theta'}^{(1)}(Y_t^-)\square Y + \frac{F_{\theta'}^{(2)}(Y_t^-)}{2}(\square Y)^2 + r(\delta^d).
\end{aligned}$$

From this and a relation $F_{\theta'}^{(1)} = F_{\theta'}^{(1)}(Y_t^-) + F_{\theta'}^{(2)}(Y_t^-)\square Y + r(\delta^{d/2})$, Lemma 6.10 implies

$$\mathrm{E}[(\square Y)(\square F_{\theta'}) - \delta^d\tilde{q}F_{\theta'}^{(1)}|\mathcal{F}_t^-] = r(\delta^{3d/2}).$$

Similarly,

$$\mathrm{E}[(\Box Y)(\Box F_{\theta'\theta'})|\mathcal{F}_t^-] \;=\; \delta^d \tilde{q} F^{(1)}_{\theta'\theta'-} + r(\delta^d).$$

For $f = F_\theta$ and $f = F_{\theta'}$, we define

$$S^-(f) \;=\; \sum_{p=2}^{d} \left[f_-^{(p)} \sum_{\{a_1,\cdots,a_p\}\in\mathcal{A}_p} \Delta_{\{a_1,\cdots,a_p\}} \right].$$

Then $S^-(f)$ is $\mathcal{F}_t^-$-measurable and $S^-(f) = R(\delta^{d/2})$. From the relation $\Box F_{\theta'} = S^-(F_{\theta'}) + F^{(1)}_{\theta'-}\Box Y + r(\delta^d)$ (Lemma 6.11), we obtain

$$\mathrm{E}[(\Box F_{\theta'})^2|\mathcal{F}_t^-] \;=\; (S^-(F_{\theta'}))^2 + \delta^d \tilde{q}(F^{(1)}_{\theta'-})^2 + r(\delta^d).$$

By substituting these formulas into (6.27) and using $F^{(p)}_{\theta'-} = F^{(p)}_{\theta-} + r(1)$, we obtain

$$\mathrm{E}[K|\mathcal{F}_t^-] \;=\; \frac{\alpha h^2}{2\tilde{q}} \left[-(S^-(F_\theta))^2 - 2\delta^d \tilde{q}(F^{(1)}_{\theta-})^2 \right] + r(\delta^d), \tag{6.28}$$

We next evaluate $\mathrm{E}[K^2|\mathcal{F}_t^-]$. From the relation $F^{(1)}_{\theta'} = F^{(1)}_{\theta'-} + r(1)$, we have

$$\begin{aligned}
&\mathrm{E}[\{(\Box Y)(\Box F_{\theta'}) - \delta^d \tilde{q} F^{(1)}_{\theta'}\}^2|\mathcal{F}_t^-] \\
&= \mathrm{E}[\{(\Box Y)S^-(F_{\theta'}) + ((\Box Y)^2 - \delta^d \tilde{q})F^{(1)}_{\theta'-} + r(\delta^d)\}^2|\mathcal{F}_t^-] \\
&= \delta^d \tilde{q}(S^-(F_{\theta'}))^2 + 2\delta^{2d}\tilde{q}^2(F^{(1)}_{\theta'-})^2 + r(\delta^{2d}).
\end{aligned}$$

Using this formula, (6.27) and $F^{(p)}_{\theta'-} = F^{(p)}_{\theta-} + r(1)$, we obtain

$$\mathrm{E}[K^2|\mathcal{F}_t^-] \;=\; \frac{\alpha^2 h^2}{\tilde{q}} \left[(S^-(F_\theta))^2 + 2\delta^d \tilde{q}^2 (F^{(1)}_{\theta-})^2 \right] + r(\delta^d). \tag{6.29}$$

We evaluate $\mathrm{E}[K^3 \mathrm{e}^{\psi K}|\mathcal{F}_t^-]$ finally. From the boundedness of $\Box F_{\theta'}$, $\Box F_{\theta\theta'}$, $\Box F_{\theta'\theta'\theta'}$, $F^{(1)}_{\theta'}$ and $F^{(1)}_{\theta'\theta'}$, (6.26) implies

$$|K| \;\leq\; C(1 + |U_t|),$$

where $U_t = \delta^{-d/2}\tilde{q}^{-1/2}\Box Y \sim N(0,1)$. Thus $\mathrm{e}^{\psi K} = R(1)$. From this and $K = R(\delta^{d/2})$,

$$\mathrm{E}[K^3 \mathrm{e}^{\psi K}|\mathcal{F}_t^-] \;=\; R(\delta^{3d/2}) \;=\; r(\delta^d). \tag{6.30}$$

By using (6.25), (6.28), (6.29) and (6.30), we obtain

$$\mathrm{E}[1 - \exp(K)|\mathcal{F}_t^-] \;=\; \frac{\alpha(1-\alpha)h^2}{2\tilde{q}} \left[(S^-(F_\theta))^2 + 2\delta^d \tilde{q}(F^{(1)}_{\theta-})^2 \right] + r(\delta^d).$$

If one puts

$$\chi_t \;=\; \tilde{q}^{-1}(S^-(F_\theta))^2 + 2\delta^d (F_{\theta-}^{(1)})^2,$$

then $h(\alpha;\theta,\theta')_1^n = (\alpha(1-\alpha)h^2/2)\sum_{t\in D_n^d}\chi_t + r(1)$. By Lemma 6.14, we have

$$\begin{aligned}\sum_{t\in D_n^d}\chi_t &= \sum_{t\in D_n^d}\mathrm{E}[\chi_t|\mathcal{F}_{t-\delta}] + r_2(1)\\ &= \sum_{t\in D_n^d}\delta^d\left[\tilde{q}^{-1}\sum_{p=2}^{d}(F_{\theta-}^{(p)})^2\sum_{\{a_1,\cdots,a_p\}\in\mathcal{A}_p}\prod_{j=1}^{p}\tilde{q}_{a_j}(t) + 2(F_{\theta-}^{(1)})^2\right] + r_2(1)\\ &\overset{L^2}{\to} J.\end{aligned}$$

Thus (6.23) is proved. □

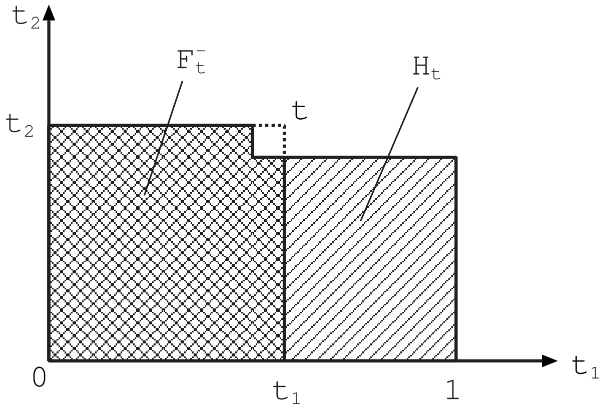


Figure 6.5: The filtration $\mathcal{H}_t$ is indicated when a lexicographic order is adopted.

6.6 Discussions

We studied the LAMN property of a class of transformed Gaussian models. We assumed that the original Gaussian process is the product of a deterministic process and a process with independent increments and that data is observed on regular lattice points. We expect that these two assumptions can be relaxed.

We concentrated to prove the LAMN property and did not discuss estimation, prediction, model selection and other statistical inference. Their asymptotic properties are also important and further investigation is required.

Chapter 7

Information criterion for LAMN models

The contents in this chapter are reported in Sei & Komaki (2004).

7.1 Introduction

Consider a model $\mathcal{P}_n = \{p_n(\cdot|\theta) \mid \theta \in \Theta \subset \mathbb{R}^k\}$ $(n = 1, 2, \cdots)$ on a sequence of measure spaces $(\Omega_n, \mathcal{F}_n, \mu_n)$, where $p_n(\cdot|\theta)$ is a probability density with a parameter θ. We recall the definition of the LAMN property (see Chapter 3).

Definition 7.1. Let $\theta \in \Theta$. A model $(\mathcal{P}_n)_{n=1}^{\infty}$ is called *locally asymptotically mixed normal* (LAMN) at θ if there exist a sequence of matrices $\gamma_n = \gamma_{n,\theta} \in \mathbb{R}^{k\times k}$, a random matrix $J = J_\theta$ and a random vector ξ with $\xi|J \sim N(0, J^{-1})$ such that for any $h \in \mathbb{R}^k$ and any convergent sequence $h_n \to h$

$$\begin{aligned} \log \frac{p_n(x|\theta + \gamma_n h_n)}{p_n(x|\theta)} &= h' J_{n,\theta}\xi_{n,\theta} - \frac{1}{2} h' J_{n,\theta} h + \mathrm{o}_{p_n}(1), \\ (\xi_{n,\theta}, J_{n,\theta}) &\rightsquigarrow (\xi, J). \end{aligned}$$

Here $'$ denotes transpose of a vector. In particular, $(\mathcal{P}_n)_{n=1}^{\infty}$ is *locally asymptotically normal* (LAN) if J is deterministic. □

The following three examples of LAMN models are analyzed later.

Example 7.2. We give a trivial example (Le Cam & Yang, 2000, p.121). Let $x_1, \cdots, x_\nu$ be an independently and identically distributed (i.i.d.) sequence subject to the probability density $p(x_1|\theta)$ and ν be a random variable independent of x_i's. If ν/n weakly converges to a non-degenerate random variable c and $p(x_1|\theta)$ satisfies some mild conditions, the model has the LAMN property with $\gamma_n = 1/\sqrt{n}$ and $J = cJ_0$, where J_0 is the Fisher information matrix of $p(x_1|\theta)$. □

Example 7.3 (Discretely observed diffusion models). Let X be a 1-dimensional diffusion process defined by the follwing stochastic differential equation

$$\mathrm{d}X_t \;=\; a(X_t,\theta)\mathrm{d}W_t + b(X_t,\theta)\mathrm{d}t, \quad X_0 \;=\; x_0, \quad t \in [0,1],$$

where x_0 is the fixed initial value of X, a and b are smooth bounded functions and W is a standard Wiener process. When θ is estimated from the discretely observed data X_{t_i}, where $t_i = i/n$ for $i = 1, \cdots, n$, it is known that the model has the LAMN property with $\gamma_n = 1/\sqrt{n}$ (Dohnal, 1987; Genon-Catalot & Jacod, 1993, 1994). The random Fisher information matrix is

$$J \;=\; 2\int_0^1 \left[\frac{\partial}{\partial\theta}\log a(X_t,\theta)\right]\left[\frac{\partial}{\partial\theta'}\log a(X_t,\theta)\right]\mathrm{d}t.$$

□

Example 7.4 (Partially explosive Gaussian AR models). Let us consider the Gaussian AR(2) model with known variance

$$\begin{aligned} X_t \;&=\; \beta_1 X_{t-1} + \beta_2 X_{t-2} + \varepsilon_t, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0,1), \quad t \in \{1,\cdots,n\}, \\ X_0 \;&=\; X_{-1} \;=\; 0. \end{aligned}$$

Let θ_1 and θ_2 be two roots of the characteristic equation $\theta^2 - \beta_1\theta - \beta_2 = 0$. Assume that $\theta_1 > 1 > |\theta_2|$. We use (θ_1,θ_2) as the parameter. Then the model is LAMN with the normalization matrix $\gamma_{n,\theta} = \mathrm{diag}(\theta_1^{-n}, n^{-1/2})$. The random Fisher information matrix is

$$J \;=\; \mathrm{diag}\left[\frac{\chi_1^2}{1-\theta_1^{-2}}, \frac{1}{1-\theta_2^2}\right],$$

where χ_1^2 is a random variable subject to the chi-square distribution with one degree of freedom. This result is generalized to any Gaussian AR(k) model for $k \geq 1$ (Jeganathan, 1988, Theorem 16). □

For examples other than described above, branching processes (See e.g. van der Vaart (1998)) and some class of semimartingale models (Luschgy, 1992) are LAMN. In Chapter 6, we have proven that the transformed Gaussian models are LAMN.

The LAMN property implies the convergence of the likelihood ratio to that of the corresponding mixed normal model (van der Vaart, 1998, Theorem 9.8). Therefore it allows us to reduce statistical problems to those of the mixed normal model. Several rigorous results including the convolution theorem and the local asymptotic minimax theorem are stated in Chapter 3.

We propose an information criterion for LAMN models by studying the corresponding mixed normal model. Since the Akaike's Information Criterion (AIC) is derived based on the LAN property (Akaike, 1974), it cannot be directly used to model selection of LAMN models. The proposed criterion *Bayes-LAMN-IC* for LAMN models is defined as an asymptotically unbiased estimator of the loss of Bayesian prediction. The loss function we adopt is equivalent to the Kullback-Leibler divergence. Here the Bayesian prediction is used since it dominates the plug-in predictive distribution as given in Section 7.3. We also give several other criteria based on other predictive distributions for comparison.

Some notations and assumptions are prepared in Section 7.2. For the mixed normal model, the Bayesian and some other predictive distributions are compared in Section 7.3. In Section 7.4, Bayes-LAMN-IC is defined for the mixed normal model. The criterion for non-limit models is given in Section 7.5. Simulation studies for the (not asymptotically) mixed normal model, the discretely observed diffusion model and the partially explosive Gaussian AR model are given in Section 7.6.

7.2 Notations and assumptions

We fix a full LAMN model $\{p_n(x|\theta) \mid \theta \in \Theta \subset \mathbb{R}^k\}$ and focus on its submodels. The corresponding full limit model is $\{p(\xi, J|h) = p(\xi|h, J)p(J) \mid h \in \mathbb{R}^k\}$, where h, ξ and J are defined in Definition 7.1. The conditional density $p(\xi|h, J)$ is $\phi(\xi|h, J^{-1})$, where $\phi(x|\mu, \Sigma)$ is the density of normal distribution with the mean vector μ and the covariance matrix Σ. The marginal density $p(J)$ of the random Fisher information matrix J does not depend on h from the definition. We use symbols indicated in Table 7.1.

Table 7.1: The symbols used in the chapter.

	full model	submodel $\alpha \in A$
non-limit model	$\{p_n(x\|\theta) \mid \theta \in \Theta \subset \mathbb{R}^k\}$	$\{p_n(x\|\theta) \mid \theta \in \Theta_\alpha\}$
limit model	$\{p(\xi\|h, J) \mid h \in \mathbb{R}^k\}$	$\{p(\xi\|h, J) \mid h \in H_\alpha\}$

In the table, A is the index set of submodels. For each $\alpha \in A$, Θ_α is a k_α-dimensional subset in Θ, where $0 \leq k_\alpha \leq k$. Let θ_α be a smooth embedding map from $\mathbb{R}^{k_\alpha}$ to Θ_α and $B_\alpha \in \mathbb{R}^{k \times k_\alpha}$ is the derivative matrix of θ_α. The subspace corresponding to α in the limit model is denoted by $H_\alpha = \{h = B_\alpha u \mid u \in \mathbb{R}^{k_\alpha}\}$.

We denote $E = E_k$ as the identity matrix of size k. We put $J_\alpha^- = B_\alpha(B_\alpha' J B_\alpha)^{-1} B_\alpha'$ and $\pi_\alpha = J_\alpha^- J$. The matrix π_α is a (random) projection operator from $\mathbb{R}^k$ to H_α. A relation $\pi_\alpha J^{-1} \pi_\alpha' = J_\alpha^- J J_\alpha^- = J_\alpha^-$ holds.

We assume that the true parameter h of the limit model is an arbitrary point in $\mathbb{R}^k$. This corresponds to a local alternate in hypothesis testing. For each submodel $\alpha \in A$, we put $h_\alpha = \pi_\alpha h$ and $\xi_\alpha = \pi_\alpha \xi$ for the true parameter h and an observation ξ. The quantity ξ_α is the maximum likelihood estimator for the subspace H_α, whose conditional mean and variance are $\mathrm{E}[\xi_\alpha | J] = h_\alpha$ and $\mathrm{Var}[\xi_\alpha | J] = J_\alpha^-$, respectively. The random variable h_α is considered as "the true parameter in H_α" because it gives the nearest distribution in H_α to the true one. The phenomenon that the true parameter is random does not appear in the LAN situation.

We assume that the prior distribution $P_\alpha(\mathrm{d}h)$ under the model α is the uniform distribution on H_α. Use of the uniform prior for the limit model is natural in the sense that any smooth prior density for the non-limit model is locally approximated by the uniform prior density. The posterior distribution $P_\alpha(\mathrm{d}h|\xi, J)$ is the degenerate normal distribution with mean ξ_α and variance J_α^-, since its characteristic function is

$$\begin{aligned}
\psi_{h|\xi,J}(\lambda) &:= \frac{\int \exp(\mathrm{i}\lambda' h)\, p(\xi|h, J)\, P_\alpha(\mathrm{d}h)}{\int p(\xi|h, J)\, P_\alpha(\mathrm{d}h)} \\
&= \frac{\int_{\mathbb{R}^{k_\alpha}} \exp\left[\mathrm{i}\lambda' B_\alpha u - \frac{1}{2}(\xi - B_\alpha u)' J (\xi - B_\alpha u)\right] \mathrm{d}u}{\int_{\mathbb{R}^{k_\alpha}} \exp\left[-\frac{1}{2}(\xi - B_\alpha u)' J (\xi - B_\alpha u)\right] \mathrm{d}u} \\
&= \exp\left[\mathrm{i}\lambda' \xi_\alpha - \frac{1}{2}\lambda' J_\alpha^- \lambda\right].
\end{aligned}$$

7.3 Risk of prediction

In this section and the next section, we consider the problem of prediction for limit models. The problem is prediction of $(\eta, \tilde{J})$ from an observation (ξ, J), where $(\eta, \tilde{J})$ and (ξ, J) are independently and identically distributed with true parameter $h \in \mathbb{R}^k$. Since the distributions of the random information matrices J and $\tilde{J}$ are independent of h, they are considered as ancillary statistics. Thus the prediction problem is reduced to that of η from ξ conditionally on J and $\tilde{J}$. When J and $\tilde{J}$ are conditioned, the arguments are usually abbreviated, for example, $q(\eta|\xi) = q(\eta|\xi, J, \tilde{J})$. Expectations are taken conditionally on J and $\tilde{J}$ unless otherwise stated.

The loss of a predictive distribution $q(\eta|\xi)$ is defined by

$$l(q(\cdot|\xi)) \;=\; -2\int p(\eta|h) \log q(\eta|\xi)\, \mathrm{d}\eta,$$

which is equivalent to the Kullback-Leibler divergence $\int p(\eta|h)\log(p(\eta|h)/q(\eta|\xi))\mathrm{d}\eta$. The risk is denoted by $r(q)=\int p(\xi|h)l(q(\cdot|\xi))\,\mathrm{d}\xi$.

We construct four predictive distributions by classifying Bayesian or plug-in, and LAMN or LAN.

Definition 7.5. The *Bayes-LAMN, plugin-LAMN, Bayes-LAN and plugin-LAN distributions* are defined by

$$\begin{aligned} q_\alpha^{\mathrm{B}}(\eta|\xi) &= \int p(\eta|h,\tilde{J})\,P_\alpha(\mathrm{d}h|\xi,J), \\ q_\alpha^{\mathrm{p}}(\eta|\xi) &= p(\eta|\xi_\alpha,\tilde{J}), \\ q_\alpha^{\mathrm{BN}}(\eta|\xi) &= \int p(\eta|h,J)\,P_\alpha(\mathrm{d}h|\xi,J), \\ q_\alpha^{\mathrm{pN}}(\eta|\xi) &= p(\eta|\xi_\alpha,J), \end{aligned}$$

respectively. □

Lemma 7.6. *The predictive distributions defined in Definition 7.5 are expressed explicitly by*

$$\begin{aligned} q_\alpha^{\mathrm{B}}(\eta|\xi) &= \phi(\eta|\xi_\alpha,\tilde{J}^{-1}+J_\alpha^{-}), \\ q_\alpha^{\mathrm{p}}(\eta|\xi) &= \phi(\eta|\xi_\alpha,\tilde{J}^{-1}), \\ q_\alpha^{\mathrm{BN}}(\eta|\xi) &= \phi(\eta|\xi_\alpha,J^{-1}+J_\alpha^{-}), \\ q_\alpha^{\mathrm{pN}}(\eta|\xi) &= \phi(\eta|\xi_\alpha,J^{-1}), \end{aligned}$$

respectively.

Proof. The first expression is obtained by using the characteristic function

$$\begin{aligned} \psi_{\eta|\xi}^{\mathrm{B}}(\lambda) &:= \int \exp(\mathrm{i}\lambda'\eta)\,q_\alpha^{\mathrm{B}}(\eta|\xi)\,\mathrm{d}\eta \\ &= \iint \exp(\mathrm{i}\lambda'\eta)\,p(\eta|h,\tilde{J})\,P_\alpha(\mathrm{d}h|\xi,J)\,\mathrm{d}\eta \\ &= \int \exp\left[\mathrm{i}\lambda'h-\frac{1}{2}\lambda'\tilde{J}^{-1}\lambda\right]\,P_\alpha(\mathrm{d}h|\xi,J) \\ &= \exp\left[\mathrm{i}\lambda'\xi_\alpha-\frac{1}{2}\lambda'(\tilde{J}^{-1}+J_\alpha^{-})\lambda\right]. \end{aligned}$$

The other expressions are also easily obtained. □

We introduce a class of predictive distributions including the four predictive distributions considered above.

Definition 7.7. Let $\Sigma_\alpha = \Sigma_\alpha(J, \tilde{J})$ be a $k \times k$ positive definite matrix. Then the *Σ-predictive distribution* is defined by

$$q_\alpha^\Sigma(\eta|\xi) \;=\; \phi(\eta|\xi_\alpha, \Sigma_\alpha).$$

□

The Bayes-LAMN, plugin-LAMN, Bayes-LAN and plugin-LAN distributions are Σ-predictive distributions with

$$\begin{aligned} \Sigma_\alpha^{\mathrm{B}} &= \tilde{J}^{-1} + J_\alpha^-, \\ \Sigma_\alpha^{\mathrm{p}} &= \tilde{J}^{-1}, \\ \Sigma_\alpha^{\mathrm{BN}} &= J^{-1} + J_\alpha^-, \\ \Sigma_\alpha^{\mathrm{pN}} &= J^{-1}, \end{aligned}$$

respectively.

The next two lemmas about the loss and risk of the Σ-predictive distributions are obtained by an elementary calculation.

Lemma 7.8. *Let $h \in \mathbb{R}^k$. The loss of the predictive distribution q_α^Σ is*

$$l(q_\alpha^\Sigma(\cdot|\xi)) \;=\; (h-\xi_\alpha)'\Sigma_\alpha^{-1}(h-\xi_\alpha) + \mathrm{tr}[\Sigma_\alpha^{-1}\tilde{J}^{-1}] + \log\det\Sigma_\alpha.$$

□

Lemma 7.9. *Let $h \in \mathbb{R}^k$. The risk of the predictive distribution q_α^Σ is*

$$r(q_\alpha^\Sigma) \;=\; (h-h_\alpha)'\Sigma_\alpha^{-1}(h-h_\alpha) + \mathrm{tr}[\Sigma_\alpha^{-1}(\tilde{J}^{-1} + J_\alpha^-)] + \log\det\Sigma_\alpha.$$

□

The next theorem reveals superiority of the Bayes-LAMN prediction q_α^{B} in a certain sense. This is a generalization of Lemma 3.17 in which only the full model $\Theta_\alpha = \Theta$ is considered. Therefore we use the Bayes-LAMN distribution for the prediction problem throughout the chapter.

Theorem 7.10. (i) *Let $h \in \mathbb{R}^k$. Then*

$$r(q_\alpha^{\mathrm{B}}) \;<\; r(q_\alpha^{\mathrm{p}}).$$

(ii) *Let $h \in H_\alpha$. Then*

$$r(q_\alpha^{\mathrm{B}}) \;\le\; r(q_\alpha^\Sigma)$$

for any $k \times k$ positive definite matrix Σ_α. The equality holds if and only if $q_\alpha^\Sigma = q_\alpha^{\mathrm{B}}$.

Proof. Let $h \in \mathbb{R}^k$ and let $\Sigma = \Sigma_\alpha$ be any $k \times k$ positive definite matrix. By Lemma 7.6 and Lemma 7.9,

$$\begin{aligned}
& r(q^{\Sigma}_\alpha) - r(q^{\mathrm{B}}_\alpha) \\
&= (h-h_\alpha)'(\Sigma^{-1}-(\tilde{J}^{-1}+J^-_\alpha)^{-1})(h-h_\alpha) \\
&\qquad +\mathrm{tr}[\Sigma^{-1/2}(\tilde{J}^{-1}+J^-_\alpha)\Sigma^{-1/2}] - k - \log\det[\Sigma^{-1/2}(\tilde{J}^{-1}+J^-_\alpha)\Sigma^{-1/2}] \\
&\geq (h-h_\alpha)'(\Sigma^{-1}-(\tilde{J}^{-1}+J^-_\alpha)^{-1})(h-h_\alpha)
\end{aligned}$$

because of an inequality

$$\mathrm{tr}\, C - k - \log\det C \geq 0$$

for any non-negative definite matrix C, where the equality holds if and only if $C = E$. If $\Sigma = \tilde{J}^{-1}$, then $(h-h_\alpha)'(\Sigma^{-1}-(\tilde{J}^{-1}+J^-_\alpha)^{-1})(h-h_\alpha) \geq 0$ for any $h \in \mathbb{R}^k$. Thus (i) holds. On the other hand, if $h \in H_\alpha$, then $h = h_\alpha$. Thus (ii) holds. □

Remark 7.11. The difference between $r(q^{\mathrm{p}}_\alpha)$ and $r(q^{\mathrm{B}}_\alpha)$ is quite large if J is close to zero relative to $\tilde{J}$. Let $h \in H_\alpha$ for simplicity. If the maximum eigenvalue of $C_\alpha = \tilde{J}^{1/2}(\tilde{J}^{-1}+J^-_\alpha)\tilde{J}^{1/2}$ is $\bar{\lambda} > 1$, then the difference is assessed as

$$\begin{aligned}
r(q^{\mathrm{p}}_\alpha) - r(q^{\mathrm{B}}_\alpha) &= \mathrm{tr} C_\alpha - k - \log\det C_\alpha \\
&\geq \bar{\lambda} - 1 - \log\bar{\lambda}.
\end{aligned}$$

On the other hand, the expectation of twice the Kullback-Leibler risk of q^{B}_α is

$$\begin{aligned}
r(q^{\mathrm{B}}_\alpha) - r(p) &= 2\iint p(\eta|h,\tilde{J}) \log\frac{p(\eta|h,\tilde{J})}{q^{\mathrm{B}}_\alpha(\eta|\xi)}\,\mathrm{d}\eta\,\mathrm{d}\xi \\
&= k + \log\det(\tilde{J}^{-1}+J^-_\alpha) - k - \log\det\tilde{J}^{-1} \\
&= \log\det C_\alpha \\
&\leq k\log\bar{\lambda}.
\end{aligned}$$

Thus

$$\frac{r(q^{\mathrm{p}}_\alpha) - r(q^{\mathrm{B}}_\alpha)}{r(q^{\mathrm{B}}_\alpha) - r(p)} \geq \frac{\bar{\lambda} - 1 - \log\bar{\lambda}}{k\log\bar{\lambda}} \to \infty$$

as $\bar{\lambda} \to \infty$. Similarly, the difference between $r(q^{\Sigma}_\alpha)$ and $r(q^{\mathrm{B}}_\alpha)$ is very large for any Σ if J is close to zero relative to $\tilde{J}$. □

7.4 Proposed information criterion

We introduce an information criterion Bayes-LAMN-IC for limit models, which forms $-2\log q^{\mathrm{B}}_{\alpha}(\xi|\xi)+c_{\alpha}$ with some correcting term c_{α}. Expectations are taken conditionally on J and $\tilde{J}$ unless otherwise stated.

We first define criteria based on Σ-predictive distributions. Bayes-LAMN-IC is a special case of them.

Definition 7.12. Fix Σ_{α}. An information criterion *Σ-IC* is defined by

$$\begin{aligned}\Sigma\text{-IC} &= \Sigma\text{-IC}(\alpha) = \Sigma\text{-IC}(\alpha,\xi,J,\tilde{J}) \\ &:= (\xi-\xi_{\alpha})'\Sigma_{\alpha}^{-1}(\xi-\xi_{\alpha})+\log\det\Sigma_{\alpha}+\mathrm{tr}[\Sigma_{\alpha}^{-1}(\tilde{J}^{-1}-J^{-1}+2J_{\alpha}^{-})]. \qquad (7.1)\end{aligned}$$

The selected model by the criterion is denoted by

$$\hat{\alpha}^{\Sigma} = \hat{\alpha}^{\Sigma}(\xi) = \hat{\alpha}^{\Sigma}(\xi,J,\tilde{J}) := \operatorname*{argmin}_{\alpha\in A}\Sigma\text{-IC}(\alpha). \qquad (7.2)$$

In particular, for $i\in\{\mathrm{B},\mathrm{p},\mathrm{BN},\mathrm{pN}\}$, *$\Sigma^{i}$-IC* is called *Bayes-LAMN-IC*, *plugin-LAMN-IC*, *Bayes-LAN-IC* and *plugin-LAN-IC*, respectively. In particular,

$$\text{Bayes-LAMN-IC}_{\alpha} = -2\log q^{\mathrm{B}}_{\alpha}(\xi|\xi)+\mathrm{tr}\left[(\tilde{J}^{-1}+J_{\alpha}^{-})^{-1}(\tilde{J}^{-1}-J^{-1}+2J_{\alpha}^{-})\right]. \quad \square$$

Theorem 7.13. *The information criterion* Σ-IC(α) *is the unique unbiased estimator of the risk* $r(q^{\Sigma}_{\alpha})$.

Proof. Put $c_{\alpha}=\mathrm{tr}[\Sigma_{\alpha}^{-1}(\tilde{J}^{-1}-J^{-1}+2J_{\alpha}^{-})]$. The expectation of Σ-IC(α) is

$$\begin{aligned}&\int p(\xi|h,J)\,\Sigma\text{-IC}(\alpha)\,\mathrm{d}\xi \\ &= \int p(\xi|h,J)\left[(\xi-\xi_{\alpha})'\Sigma_{\alpha}^{-1}(\xi-\xi_{\alpha})\right]\mathrm{d}\xi+\log\det\Sigma_{\alpha}+c_{\alpha} \\ &= (h-h_{\alpha})'\Sigma_{\alpha}^{-1}(h-h_{\alpha})+\mathrm{tr}\left[\Sigma_{\alpha}^{-1}(J^{-1}-J_{\alpha}^{-})\right]+\log\det\Sigma_{\alpha}+c_{\alpha} \\ &= (h-h_{\alpha})'\Sigma_{\alpha}^{-1}(h-h_{\alpha})+\mathrm{tr}\left[\Sigma_{\alpha}^{-1}(\tilde{J}^{-1}+J_{\alpha}^{-})\right]+\log\det\Sigma_{\alpha} \\ &= r(q^{\Sigma}_{\alpha})\end{aligned}$$

by Lemma 7.9. Uniqueness holds due to the completeness of the statistic ξ (Lehmann & Casella, 1998, p.42 and p.87). □

Proposition 7.14. *The information criterion* plugin-LAN-IC *is equivalent to* AIC.

Proof. By putting $\Sigma_{\alpha}=J^{-1}$ in (7.1), it is shown that

$$\text{plugin-LAN-IC}(\alpha) = -2\log q^{\mathrm{pN}}_{\alpha}(\xi|\xi)+2k_{\alpha}+\mathrm{tr}[J(\tilde{J}^{-1}-J^{-1})].$$

Since the last term is independent of α, plugin-LAN-IC is equivalent to AIC. □

We adopt Bayes-LAMN-IC among Σ-IC's because it is compatible with the Bayes-LAMN prediction, which has the dominating property obtained in Theorem 7.10. It should be noted that the criterion does not coincide with AIC even if a LAN model is considered. For LAN models, both Bayes-LAMN-IC and Bayes-LAN-IC coincide with

$$\mathrm{PIC} \;=\; -2\log q_{\alpha}^{\mathrm{BN}}(\xi|\xi) + k_{\alpha} \tag{7.3}$$

when the uniform prior is used (Kitagawa, 1997). It is stated in Chapter 4. The performance of PIC and AIC does not seem very different since J is deterministic for LAN models. On the other hand, for LAMN models, the difference between Bayes-LAMN-IC and AIC is quite serious as remarked after Theorem 7.10.

We compare Σ-IC's for different Σ's in Section 7.6. The risk R of the model selection procedure based on Σ-IC is defined by the risk of the Bayesian prediction based on the model selection procedure using Σ-IC, that is,

$$R \;=\; R(\Sigma, h) \;=\; \mathrm{E}[r(q_{\hat{\alpha}^{\Sigma}}^{\mathrm{B}})], \tag{7.4}$$

where E denotes the expectation with respect to J and $\tilde{J}$. Recall that $\hat{\alpha}^{\Sigma}$ is the selected model (7.2). An information criterion Σ-IC whose risk R is small is a good criterion.

In practice, a model selection is implemented without use of $\tilde{J}$. For this purpose, we define an expectation version of Σ-IC by

$$\int \Sigma\text{-IC}(\alpha, \xi, J, \tilde{J}) p(\tilde{J}) \mathrm{d}\tilde{J}.$$

We don't use the expected version of Σ-IC for the numerical examples in Section 7.6 since the performance of an information criterion is assessed by its performance of prediction.

7.5 Information criterion for non-limit models

In this section, the information criteria for limit models defined in the previous section are restored to those for the non-limit models. Only a heuristic definition is given here. The conditions for asymptotic properties such as contiguity of the selected predictive distribution are not discussed.

Let $\hat{\theta}$ and $\hat{\theta}_{\alpha}$ be the maximum likelihood estimators (or other asymptotically efficient estimators) for the full model and the submodel $\alpha \in A$, respectively. Our Σ-IC for the non-limit models is, by using Σ-IC for the limit models,

$$\Sigma\text{-IC}(\alpha)|_{J=J^{(n)}, \tilde{J}=\tilde{J}^{(n)}, (\xi-\xi_{\alpha})=(\xi-\xi_{\alpha})^{(n)}}, \tag{7.5}$$

where a matrix $J^{(n)}$ is defined by

$$J^{(n)} \;:=\; -\frac{\partial^2}{\partial u \partial u'} \log p_n(x|\hat{\theta}+\gamma_n u)\bigg|_{u=0}, \tag{7.6}$$

$\tilde{J}^{(n)}$ is given by replacing x in (7.6) with y and

$$(\xi-\xi_\alpha)^{(n)} \;:=\; \gamma_n^{-1}(\hat{\theta}-\hat{\theta}_\alpha).$$

Under mild conditions, $J^{(n)}$, $\tilde{J}^{(n)}$ and $(\xi-\xi_\alpha)^{(n)}$ converge to J, $\tilde{J}$ and $\xi-\xi_\alpha$ as $n\to\infty$, respectively.

In general, $J^{(n)}$ and $\tilde{J}^{(n)}$ may not be positive definite. Therefore some modification is needed. For the discretely observed diffusion models (Example 7.3), we can use a non-negative definite matrix $J^{\sharp(n)}$ instead of $J^{(n)}$ defined by

$$J^{\sharp(n)} \;:=\; \sum_{i=1}^{n} \frac{2}{n}\left[\frac{\partial}{\partial\theta}\log a(X_{t_i},\hat{\theta})\frac{\partial}{\partial\theta'}\log a(X_{t_i},\hat{\theta})\right]. \tag{7.7}$$

The matrix $J^{\sharp(n)}$ is used at the numerical experiments in Subsection 7.6.2.

7.6 Examples

Three examples are considered here. In Subsection 7.6.1, some theoretical and experimental results for a limit model are given. Subsection 7.6.2 is devoted to a numerical study of the discretely observed diffusion models. Subsection 7.6.3 deals with the partially explosive Gaussian AR model.

7.6.1 Scalar-randomness models

Let $J=cJ_0$ with a 1-dimensional positive random variable c and a deterministic matrix J_0 as Example 7.2 in Section 7.1. We call it a scalar-randomness model. If we consider nested submodels, the information criterion has a simple representation. Let $A=\{0,1,\cdots,k\}$. Suppose that $H_0=\{0\}\subset\mathbb{R}^k$ and H_α is an α-dimensional subspace of $\mathbb{R}^k$ including $H_{\alpha-1}$ for $1\le\alpha\le k$. Put $\tilde{J}=\tilde{c}J_0$ where $\tilde{c}$ is a random variable independent of c and has the same distribution as c.

We assume that $J_0=E$ and $H_\alpha=\{(a_1,\cdots,a_\alpha,0,\cdots,0)\mid a_1,\cdots,a_\alpha\in\mathbb{R}\}$ without loss of generality. The loss $l(q_\alpha^{\mathrm{B}}(\cdot|\xi))$ of the Bayesian prediction is, by Lemma 7.8,

$$\begin{aligned} l(q_\alpha^{\mathrm{B}}(\cdot|\xi)) \;&=\; \sum_{i=1}^{\alpha}\left[(\tilde{c}^{-1}+c^{-1})^{-1}\{(h_i-\xi_i)^2+\tilde{c}^{-1}\}+\log(\tilde{c}^{-1}+c^{-1})\right] \\ &\quad + \sum_{i=\alpha+1}^{k}\left[\tilde{c}h_i^2+1+\log\tilde{c}^{-1}\right], \end{aligned}$$

where ξ_i and h_i is the i-th component of ξ and h, respectively.

We consider only Σ-IC satisfying the condition that Σ_α is a diagonal matrix whose i-th diagonal component σ_i is $s = s(c, \tilde{c})$ if $i \le \alpha$ and $t = t(c, \tilde{c})$ otherwise, where s and t are common in all α. The four criteria (Bayes-LAMN, plugin-LAMN, Bayes-LAN and plugin-LAN) satisfy the condition as indicated in Table 7.2. The expression of Σ-IC is

$$\begin{aligned}\Sigma\text{-IC}(\alpha) &= \sum_{i=1}^{\alpha} \left[s^{-1}(\tilde{c}^{-1}+c^{-1}) + \log s\right] + \sum_{i=\alpha+1}^{k} \left[t^{-1}\xi_i^2 + t^{-1}(\tilde{c}^{-1}-c^{-1}) + \log t\right] \\ &= \sum_{i=1}^{k} \left[s^{-1}(\tilde{c}^{-1}+c^{-1}) + \log s\right] + t^{-1} \sum_{i=\alpha+1}^{k} (\xi_i^2 - \bar{\xi}^2),\end{aligned}$$

where

$$\bar{\xi}^2 = t\left[s^{-1}(\tilde{c}^{-1}+c^{-1}) + \log s - t^{-1}(\tilde{c}^{-1}-c^{-1}) - \log t\right].$$

For fixed $\alpha \in A$, the set Ξ_α of ξ such that $\hat{\alpha}^\Sigma(\xi) = \alpha$ is given by

$$\Xi_\alpha = L_\alpha \cap G_\alpha,$$

where

$$L_\alpha = \{\xi \mid \forall j \le \alpha, \sum_{i=j+1}^{\alpha} (\xi_i^2 - \bar{\xi}^2) > 0\}$$

and

$$G_\alpha = \{\xi \mid \alpha < \forall j \le k, \sum_{i=\alpha+1}^{j} (\xi_i^2 - \bar{\xi}^2) < 0\}.$$

The quantities s, t and $\bar{\xi}^2$ corresponding to the four criteria are summarized in Table 7.2, where we put $r = \tilde{c}/c$.

If $s = t$, then $\bar{\xi}^2 = 2c^{-1}$, which is independent of s. Thus, by Proposition 7.14, the next proposition holds.

Proposition 7.15. *Suppose that the model is a scalar-randomness model and Σ_α is common in all $\alpha \in A$. Then* Σ-IC *is equivalent to* AIC. □

Consider the scalar-randomness model with dimension $k = 10$. Assume that c takes only two values $\sqrt{10}$ and $1/\sqrt{10}$ with the same probability. We numerically evaluate the risk R of the model selection procedures (eq. (7.4)).

Figure 7.1 indicates R of FULL (which always selects the full model: $\bar{\xi}^2 = 0$), Bayes-LAMN-IC, Bayes-LAN-IC, AIC and BOUND (which is a lower bound based on the "best

Table 7.2: The quantities s, t and $\bar{\xi}^2$ corresponding to the four predictive distributions. $(r = \tilde{c}/c)$

prediction	s	t	$\bar{\xi}^2$
Bayes-LAMN	$\tilde{c}^{-1} + c^{-1}$	$\tilde{c}^{-1}$	$c^{-1}(r^{-1}\log(1+r)+1)$
plugin-LAMN	$\tilde{c}^{-1}$	$\tilde{c}^{-1}$	$2c^{-1}$
Bayes-LAN	$2c^{-1}$	c^{-1}	$c^{-1}(\log 2 + \frac{3}{2} - \frac{1}{2r})$
plugin-LAN	c^{-1}	c^{-1}	$2c^{-1}$

selection" $\hat{\alpha} = \text{argmin}\ l(q_\alpha^{\text{B}})$ using the true h), respectively. The true parameter h takes its value in $D_i = \{d_i e_i \mid d_i \in [0, 10]\}$ for $i \in \{1, \cdots, 10\}$, where $e_i = (0, \cdots, 0, 1, 0, \cdots, 0)$ is the i-th unit vector in $\mathbb{R}^{10}$. The horizontal axis denotes d_i such that $h = d_i e_i$.

The figure shows that Bayes-LAMN-IC is better than AIC especially for $h \in D_i$ $(i = 4, \cdots, 10)$. Although the minimax criterion is FULL, the difference between risks of Bayes-LAMN-IC and BOUND is stable throughout the parameter space compared to that of FULL and BOUND. This kind of stability is considered important from the view point of model selection. The difference between risks of an information criterion and BOUND is nothing else the regret defined in Chapter 4. From the above numerical experiment, the maximum regret over $h \in \cup_{i=1}^{10} D_i$ of Bayes-LAMN-IC, Bayes-LAN-IC, AIC and FULL is 5.17, 14.0, 12.7 and 9.7, respectively. Thus Bayes-LAMN-IC is best. However, if we replace the distribution of c by a two-point distribution on 10 and 1/10 (with prob. 1/2 each), the result is changed: the maximum regret of Bayes-LAMN-IC, Bayes-LAN-IC, AIC and FULL over $h \in \cup_{i=1}^{10} D_i$ is 44, 160, 141 and 15, respectively. The criterion FULL is best in this example. Nevertheless, Bayes-LAMN-IC remains to have better performance than Bayes-LAN-IC and AIC.

7.6.2 Discretely observed diffusion models

Let us consider the discretely observed diffusion model stated in Example 7.3 of Section 7.1. The example used here is

$$\mathrm{d}X_t = \frac{1 + \theta X_t^2}{1 + X_t^2}\mathrm{d}W_t, \quad X_0 = 0, \quad \theta > 0, \tag{7.8}$$

which satisfies the regularity condition in Genon-Catalot & Jacod (1994). Two submodels $\Theta_{\text{I}} = \{\theta \mid \theta = 1\}$ and $\Theta_{\text{II}} = \{\theta \mid \theta > 0\}$ are compared, where $A = \{\text{I}, \text{II}\}$ is the index set. If $\theta = 1$, $X_t = W_t$.

Figure 7.2 gives a numerical result about the risk R of the model selection procedure based on Bayes-LAMN-IC and AIC. The simulation algorithm for Σ-IC is as follows.

1. Fix integers L and $\tilde{L}$. For each $l = 1, 2, \cdots, L$,
   (a) Generate a path $\{X_t(l) \mid t \in \{\frac{1}{n}, \cdots, \frac{n}{n}\}\}$ according to the true parameter θ. Calculate the maximum likelihood estimator $\hat{\theta}(l)$ for the full model and the random Fisher information $J^{\sharp(n)}(l)$ by the formula (7.7).
   (b) For $\tilde{l} = 1, 2, \cdots, \tilde{L}$, generate $\tilde{L}$ paths $\{Y_t(l, \tilde{l}) \mid t \in \{\frac{1}{n}, \cdots, \frac{n}{n}\}\}$ according to the estimated parameter $\hat{\theta}(l)$. Calculate the random Fisher information $\tilde{J}^{\sharp(n)}(l, \tilde{l})$.
   (c) Select one of the submodels according to Σ-IC determined by eq. (7.5) and calculate the loss $\ell(l, \tilde{l})$ of the selected predictive distribution, where the loss is also calculated by LAMN approximation for simplicity.
2. Calculate $R = (L\tilde{L})^{-1} \sum_{l=1}^{L} \sum_{\tilde{l}=1}^{\tilde{L}} \ell(l, \tilde{l})$.

The number of sampling points is $n = 100$. The number of loops is $L = \tilde{L} = 1000$ for each true $\theta \in \{0.25, 0.50, \cdots, 3.00\}$. In the example, Bayes-LAMN-IC is better than AIC in the minimax sense.

7.6.3 A partially explosive Gaussian AR model

Let us consider the partially explosive Gaussian AR(2) model stated in Example 7.4 of Section 7.1. Two submodels $\Theta_{\mathrm{I}} = \{\theta \mid \theta_1 > 1,\ \theta_2 = 0\}$ and $\Theta_{\mathrm{II}} = \{\theta \mid \theta_1 > 1,\ |\theta_2| < 1\}$ are compared, where $A = \{\mathrm{I}, \mathrm{II}\}$ is the index set. Let $J = \mathrm{diag}(J_{11}, J_{22})$, $\tilde{J} = \mathrm{diag}(\tilde{J}_{11}, \tilde{J}_{22})$ and $\xi = (\xi_1, \xi_2)'$.

Since $\tilde{J}_{22} = J_{22}$, Bayes-LAMN-IC's for the two submodels are

$$\begin{aligned}
\text{Bayes-LAMN-IC(I)} &= \xi_2^2 (2J_{22}^{-1})^{-1} + \log(\tilde{J}_{11}^{-1} + J_{11}^{-1}) + \log(J_{22}^{-1}) + 1, \\
\text{Bayes-LAMN-IC(II)} &= \log(\tilde{J}_{11}^{-1} + J_{11}^{-1}) + \log(2J_{22}^{-1}) + 2.
\end{aligned}$$

Their difference is $\xi_2^2 J_{22}/2 - (\log 2 + 1)$. On the other hand,

$$\begin{aligned}
\text{Bayes-LAN-IC(I)} &= \xi_2^2 (2J_{22}^{-1})^{-1} + \log(2J_{11}^{-1}) + \log(J_{22}^{-1}) + 1, \\
\text{Bayes-LAN-IC(II)} &= \log(2J_{11}^{-1}) + \log(2J_{22}^{-1}) + 2.
\end{aligned}$$

Their difference is $\xi_2^2 J_{22}/2 - (\log 2 + 1)$. Therefore both Bayes-LAMN-IC and Bayes-LAN-IC are equivalent to PIC (eq. (7.3)). In particular, $\tilde{J}$ is not needed in order to calculate them. Similarly, both plugin-LAMN-IC and plugin-LAN-IC are equivalent to AIC. These properties hold for any AR(k) model if we consider only submodels where some of the stationary components of θ are restricted to zero.

We now compare Bayes-LAMN-IC and AIC by finite-sample experiments. Figure 7.3 gives a numerical result about the risk R of the model selection procedure based on Bayes-LAMN-IC and AIC. The simulation algorithm for Bayes-LAMN-IC is as follows. A similar algorithm is used for AIC.

1. Fix L and $\tilde{L}$. For each $l = 1, \cdots, L$,
   (a) Generate a path $\{X_t(l) \mid t \in \{1, \cdots, n\}\}$ according to the true parameter θ. Calculate the maximum likelihood estimator $\hat{\theta}_\alpha(l)$ and Bayes-LAMN-IC(α) for each model $\alpha \in A$.
   (b) Calculate the loss $\ell(l)$ by the Monte-Carlo method, that is, generate $\tilde{L}$ paths $\{Y_t(l, \tilde{l}) \mid t \in \{1, \cdots, n\}\}$ ($\tilde{l} = 1, \cdots, \tilde{L}$) according to the true parameter θ and take the sample mean: $\ell(l) = \tilde{L}^{-1} \sum_{\tilde{l}=1}^{\tilde{L}} 2 \log\{p_n(Y|\theta)/q_n^{\mathrm{B}}(Y|X)\}$, where $q_n^{\mathrm{B}}(Y|X)$ is the selected predictive distribution by Bayes-LAMN-IC.
2. Calculate $R = L^{-1} \sum_{l=1}^{L} \ell(l)$.

The number of sampling points is $n = 100$. The number of loops is $L = \tilde{L} = 1000$ for each $\theta = (\theta_1, \theta_2) \in \{1.03\} \times \{0.00, 0.05, 0.10, \cdots, 1.00\}$. In the example, Bayes-LAMN-IC is slightly better than AIC in the minimax sense.

7.7 Discussions

We proposed an information criterion Bayes-LAMN-IC for LAMN models. It is the unique unbiased estimator of the risk of the Bayesian prediction. We numerically compared it with other criteria including AIC. For the scalar-randomness model, the risk of the model selection procedures based on Bayes-LAMN-IC was relatively stable over the true parameter space. For the discretely observed diffusion model and the partially explosive Gaussian AR model, the maximum risk of the model selection procedure based on Bayes-LAMN-IC was less than the maximum risk of the procedure based on the other criteria. These numerical results show that Bayes-LAMN-IC is better than the other criteria.

The remaining tasks are to give many numerical experiments, real data analysis, theoretical evaluation of the risk and characterization of Bayes-LAMN-IC. Another future work is to construct a version of Bayes-LAMN-IC like Takeuchi's information criterion (TIC; see Chapter 4). It is naturally constructed for the i.i.d. models with random number of samples. We believe that it is also available for the discretely observed diffusion models.

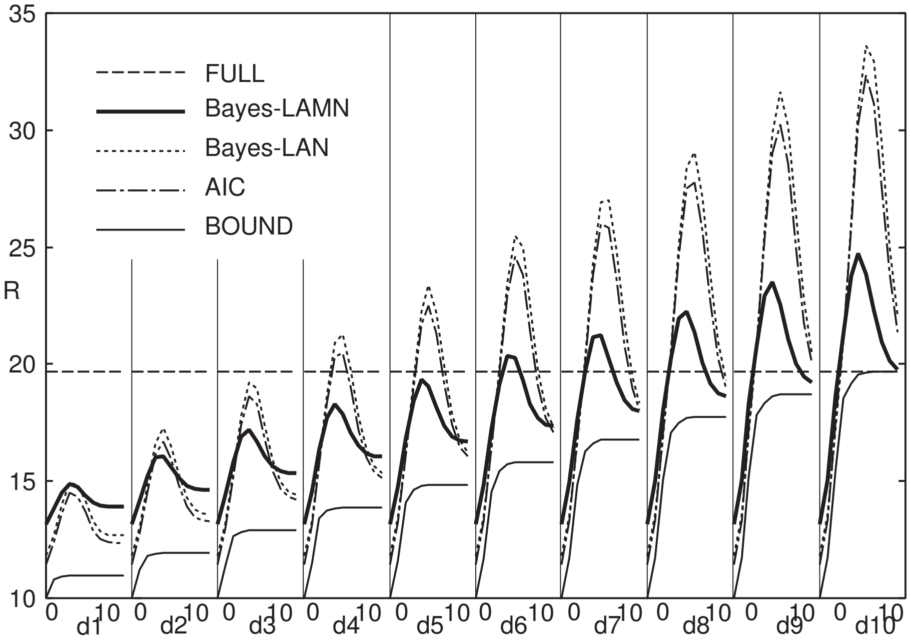


Figure 7.1: The risk R of the model selection procedures for the scalar-randomness model. The true parameter h takes its value in $D_i = \{(0, \cdots, 0, d_i, 0, \cdots, 0) \in \mathbb{R}^{10} \mid d_i \in [0, 10]\}$ for each $i \in \{1, \cdots, 10\}$, where $(0, \cdots, 0, d_i, 0, \cdots, 0)$ denotes the vector whose i-th coordinate is d_i. The horizontal axis denotes d_i such that $h = (0, \cdots, 0, d_i, 0, \cdots, 0)$.

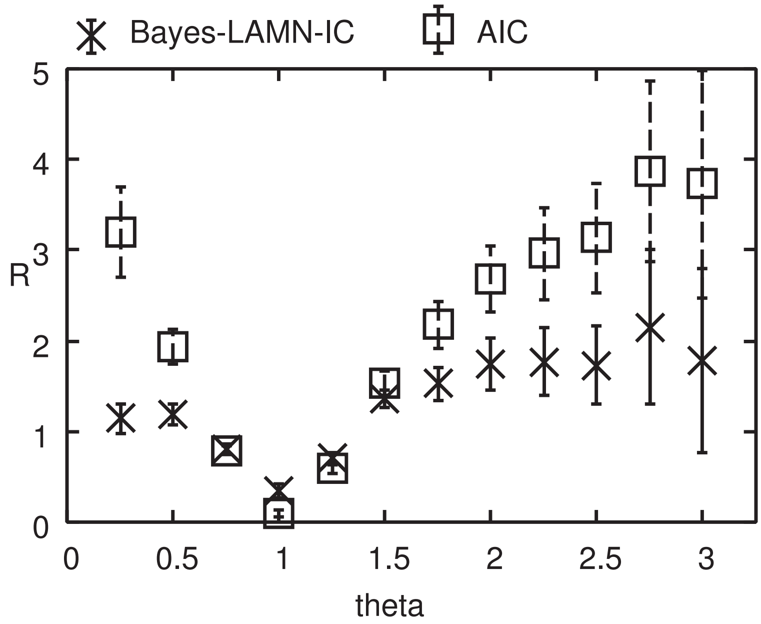


Figure 7.2: The risk R of the model selection procedures for the discretely observed diffusion model (eq. (7.8)). The confidence interval is based on 3 times of the standard deviation.

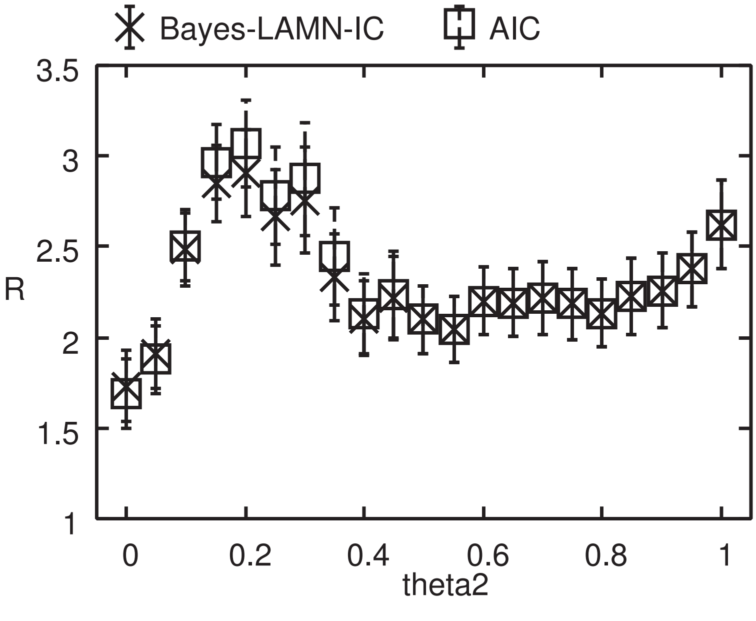


Figure 7.3: The risk R of the model selection procedures for the partially explosive Gaussian AR(2) model. The horizontal axis denotes θ_2. The value of θ_1 is fixed to 1.03. The confidence interval is based on 3 times of the standard deviation.

Chapter 8

Conclusion

We showed consistency of the quasi maximum likelihood estimator under the semiparametric setting in Chapter 5. We proved the LAMN property of a restricted class of transformed Gaussian models in Chapter 6. We proposed an information criterion Bayes-LAMN-IC for LAMN models in Chapter 7. By combining latter two results, we obtain a prediction procedure based on Bayes-LAMN-IC for transformed Gaussian models. From the first result, we believe that the LAMN property for more general transformed Gaussian models than the restricted class in Chapter 6 is proved in the future.

Acknowledgments

I thank my supervisor Prof. Fumiyasu Komaki for all of his help and lots of encouragement to me during the master's and doctoral courses. I also thank Prof. Akimichi Takemura for his many valuable comments and suggestions. Prof. Yasunori Okabe, Prof. Yoshihiro Yajima and Prof. Takehiko Horita gave me many helpful comments about the draft of this thesis. Dr. Satoshi Aoki and Mr. Kei Kobayashi made my student life very meaningful and enjoyable in various forms. I also feel grateful for the discussion with Prof. Hiroyuki Takeuchi, Prof. Motoaki Kawanabe, Dr. Hisayuki Hara, Dr. Takafumi Kanamori, Dr. Yoshiyuki Ninomiya, Mr. Hideo Iwasaki, Mr. Kentaro Tanaka, Mr. Ichiro Fujisawa, Mr. Fuyuhiko Tanaka and all of my academic colleagues. I was financially aided by Superrobust Computation Project of the 21st century COE project "Information Science and Technology Strategic Core". Lastly, I thank my father, mother, grand mother, brother and his family, for their support.

Dec. 16, 2004.

Bibliography

Adler, R. J. (1981). *The Geometry of Random Fields.* Chichester: John Wiley & Sons.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19**.

Akaike, H. (1980). On the use of the predictive likelihood of a Gaussian model. *Ann. Inst. Statist. Math.* **32**, 311–324.

Amari, S. (1985). *Differential Geometrical methods in statistics.* Lecture Notes in Statistics 28. Berlin: Springer-Verlag.

Amari, S. (1987). Differential geometry of a parametric family of invertible linear systems - Riemannian metric, dual affine connections, and divergence. *Math. Systems Thoery* **20**, 53–82.

Arato, M., Pap, G. & van Zuijlen, M. C. A. (2001). Asymptotic inference for spatial autoregression and orthogonality of Ornstein-Uhlenbeck sheets. *Comput. Math. Appl.* **42**, 219–229.

Beran, J. (1994). *Statistics for Long-Memory Processes.* New York: Chapman & Hall.

Billingsley, P. (1999). *Convergence of Probability Measures.* New York: John Wiley & Sons, 2nd ed.

Bingham, N. H., Goldie, C. M. & Teugels, J. L. (1987). *Regular Variation.* Cambridge: Cambridge University Press.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc.* **26**, 211–252.

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference - A Practical Information-Theoretic Approach.* New York: Springer-Verlag, 2nd ed.

Chan, G., Hall, P. & Poskitt, D. S. (1995). Periodogram-based estimators of fractal properties. *Ann. Statist.* **23**, 1684–1711.

Chan, G. & Wood, A. T. A. (2000). Increment-based estimators of fractal dimension for two-dimensional surface data. *Statist. Sinica* **10**, 343–376.

Chan, G. & Wood, A. T. A. (2004). Estimation of fractal dimension for a class of non-Gaussian stationary processes and fields. *Ann. Statist.* **32**, 1222–1260.

Chilès, J.-P. & Delfiner, P. (1999). *Geostatistics - Modeling Spatial Uncertainty.* New York: John Wiley & Sons.

Constantine, A. G. & Hall, P. (1994). Characterizing surface smoothness via estimation of effective fractal dimension. *J. Roy. Statist. Soc. Ser. B* **56**, 97–113.

Cressie, N. A. C. (1993). *Statistics for Spatial Data.* New York: John Wiley & Sons.

Dahlhaus, R. (1989). Efficient parameter estimation for self-similar processes. *Ann. Statist.* **17**, 1749–1766.

Davies, S. & Hall, P. (1999). Fractal analysis of surface roughness by using spatial data. *J. Roy. Statist. Soc. Ser. B* **61**, 3–37.

De Oliveira, V., Kedem, B. & Short, D. A. (1997). Bayesian prediction of transformed Gaussian random fields. *J. Amer. Statist. Assoc.* **92**, 1422–1433.

Dohnal, G. (1987). On estimating the diffusion coefficient. *J. Appl. Probab.* **24**, 105–114.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, vol. 2. New York: John Wiley & Sons, 2nd ed.

Feuerverger, A., Hall, P. & Wood, A. T. A. (1994). Estimation of fractal index and fractal dimension of a Gaussian process by counting the number of level crossings. *J. Time Series Anal.* **15**, 587–606.

Fox, R. & Taqqu, M. S. (1986). Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Ann. Statist.* **14**, 517–532.

Fox, R. & Taqqu, M. S. (1987). Central limit theorems for quadratic forms in random variables having long-range dependence. *Probab. Theory Related fields* **74**, 213–240.

Genon-Catalot, V. & Jacod, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. Inst. H. Poincaré Probab. Statist.* **29**, 119–151.

Genon-Catalot, V. & Jacod, J. (1994). Estimation of the diffusion coefficient for diffusion processes: random sampling. *Scand. J. Statist.* **21**, 193–221.

Gneiting, T. (2002). Compactly supported correlation functions. *J. Multivariate Anal.* **83**, 493–508.

Guyon, X. (1982). Parameter estimation for a stationary process on d-dimensional lattice. *Biometrika* **69**, 95–105.

Hall, P. & Roy, R. (1994). On the relationship between fractal dimension and fractal index for stationary stochastic processes. *Ann. Appl. Probab.* **4**, 241–253.

Hall, P. & Wood, A. T. A. (1993). On the performance of Box-Counting estimators of fractal dimension. *Biometrika* **80**, 246–252.

Ibragimov, I. A. & Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* New York: Springer-Verlag.

Ishiguro, M., Sakamoto, Y. & Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an estension of AIC. *Ann. Inst. Statist. Math.* **49**, 411–434.

Istas, J. & Lang, G. (1997). Quadratic variations and estimation of the local Hölder index of a Gaussian process. *Ann. Inst. H. Poincaré Probab. Statist.* **33**, 407–436.

Jacod, J. & Shiryaev, A. N. (1987). *Limit Theorems for Stochastic Processes.* Berlin: Springer-Verlag.

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

Jeganathan, P. (1983). Some asymptotic properties of risk functions when the limit of the experiment is mixed normal. *Sankhyā Ser. A* **45**, 66–87.

Jeganathan, P. (1988). On the strong approximation of the distributions of estimators in linear stochastic models, I and II: stationary and explosive AR models. *Ann. Statist.* **16**, 1283–1314.

KENT, J. & WOOD, A. T. A. (1997). Estimating the fractal dimension of a locally self-similar Gaussian process by using increments. *J. Roy. Statist. Soc. Ser. B* **59**, 679–699.

KHOSHNEVISAN, D. (2002). *Multiparameter Processes.* New York: Springer-Verlag.

KITAGAWA, G. (1997). Information criteria for the predictive evaluation of Bayesian models. *Comm. Statist. Theory Methods* **26**, 2223–2246.

KONISHI, S. & KITAGAWA, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875–890.

KUTOYANTS, Y. (1994). *Identification of Dynamical Systems with Small Noise.* Mathematics and Its Applications. Boston: Kluwer Academic Publishers.

KUTOYANTS, Y. A. (1984). Expansion of a maximum likelihood estimate by diffusion powers. *Theory Probab. Appl.* **29**, 465–477.

LE CAM, L. & YANG, G. L. (2000). *Asymptotics in Statistics.* New York: Springer-Verlag, 2nd ed.

LEHMANN, E. L. & CASELLA, G. (1998). *Theory of Point Estimation.* New York: Springer-Verlag, 2nd ed.

LUSCHGY, H. (1992). Local asymptotic mixed normality for semimartingale experiments. *Probab. Theory Related Fields* **92**, 151–176.

MANDELBROT, B. B. & VAN NESS, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Rev.* **10**, 422–437.

PITMAN, E. J. G. (1968). On the behaviour of the characteristic function of a probability distribution in the neighbourhood of the origin. *J. Aust. Math. Soc.* **10**, 423–443.

PRAKASA RAO, B. L. S. (1999). *Statistical Inference for Diffusion Type Processes.* London: Arnold.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

SEI, T. (2004). Local asymptotic mixed normality of transformed Gaussian models for random fields. Tech. Rep. METR2004-37, The University of Tokyo.

SEI, T. & KOMAKI, F. (2003). Information geometry of small diffusions. Submitted.

Sei, T. & Komaki, F. (2004). Bayesian prediction and model selection for locally asymptotically mixed normal models. Tech. Rep. METR2004-25, The University of Tokyo.

Shibata, R. (1986). Selection of the number of regression variables; a minimax choice of generalized FPE. *Ann. Inst. Statist. Math.* **38**, 459–474.

Sørensen, M. & Uchida, M. (2003). Small-diffusion asymptotics for discretely sampled stochastic differential equations. *Bernoulli* **9**, 1051–1069.

Stein, M. L. (1995). Fixed-domain asymptotics for spatial periodograms. *J. Amer. Statist. Assoc.* **90**, 1277–1288.

Stein, M. L. (1999). *Interpolation of Spatial Data.* New York: Springer-Verlag.

Stein, M. L. (2001). Local stationarity and simulation of self-affine intrinsic random functions. *IEEE Trans. Inform. Theory* **47**, 1385–1390.

Stein, M. L. (2002). Fast and exact simulation of fractional Brownian surfaces. *J. Comput. Graph. Statist.* **11**, 587–599.

Stone, C. J. (1982). Local asymptotic admissibility of a generalization of Akaike's model selection rule. *Ann. Inst. Statist. Math.* **34**, 123–133.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B* **39**, 44–47.

Takeuchi, K. (1976). Distribution of information statistics and a criterion of model fitting. *Suri-Kagaku (Mathematic Sciences)* **153**, 12–18. (In Japanese).

Uchida, M. (2003). Estimation for dynamical systems with small noise from discrete observations. *J. Japan Statist. Soc.* **33**, 157–168.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge: Cambridge University Press.

Wang, Y. (1997). Fractal function estimation via wavelet shrinkage. *J. Roy. Statist. Soc. Ser. B* **59**, 603–613.

Yajima, Y. (2003). *Statistics of Economic Time Series (in Japanese)*, chap. II, Time Series Model with Long Memory. No. 8 in Fronteer in Statistical Science. Tokyo: Iwanami, pp. 103–202.

YING, Z. (1993). Maximum likelihood estimation of parameters under a spatial sampling scheme. *Ann. Statist.* **21**, 1567–1590.

YOSHIDA, N. (1992a). Asymptotic expansion for statistics related to small diffusions. *J. Japan Statist. Soc.* **22**, 139–159.

YOSHIDA, N. (1992b). Asymptotic expansions of maximum likelihood estimators for small diffusions via the theory of Malliavin-Watanabe. *Probab. Theory Related Fields* **92**, 275–311.

YOSHIDA, N. (1997). Malliavin calculus and asymptotic expansion for martingales. *Probab. Theory Related Fields* **109**, 301–342.

ZHU, Z. & STEIN, M. L. (2002). Parameter estimation for fractional Brownian surfaces. *Statist. Sinica* **12**, 863–883.

ZYGMUND, A. (2002). *Trigonometric Series.* Cambridge University Press, 3rd ed.

Index