From prediction and interpolation problem to parameter estimation problem of time series

Yan Liu (Kyoto University)*

Abstract

We associate the prediction and interpolation problem with parameter estimation problem for stationary time series. To expand the scope of prediction and interpolation problem for Gaussian stationary time series, we discuss a concept to minimize the prediction (or interpolation) error, which is also available for the harmonizable stable process. The concept is readily transplanted to a disparity between the spectral density and a parametric model for the parameter estimation. We define an estimator which minimizes the disparity as a minimum contrast estimator. We study the asymptotic distributions of minimum contrast estimators in cases of a stationary process with finite variance innovations and infinite variance innovations. We also discuss the asymptotic efficiency and the robustness of our estimator based on the disparity.

1. Introduction

The prediction and interpolation problem for the stationary process was introduced by Kolmogorov (1941a). The problem has been thoroughly investigated by Grenander and Rosenblatt (1954), Yaglom (1962), Rozanov (1967), just to name a few. By a work of Urbanik (1967), a formulation of prediction problem from operator theory was proposed. He assumed that the predictor and the prediction error are independent and find that this formulation is only valid for the Gaussian stationary process. As a natural extension instead of independence, a concept to minimize the prediction error which is evaluated in an adequate normed space was proposed in Hosoya (1982). This refinement of the definition makes the prediction problem available for a much richer class of stationary processes, such as harmonizable stable process (See Schilder (1970), Samorodnitsky and Taqqu (1994)). The interpolation problem for the class is considered along the same line.

The original form of the Whittle estimator was proposed for parameter estimation of time series in Whittle (1952). He discussed the error bound derived in Kolmogorov (1941a) of the prediction problem and suggested using the bound in parameter estimation. Later on, a statistical theory for multivariate time series was considered in Hannan (1970), including ideas of approximate maximum likelihood estimation. Bloomfield (1973) proposed an exponential model for the spectral estimation, which contains elements of both parametric and nonparametric procedures. Minimum contrast estimations with location disparity and scale disparity were separately discussed in Taniguchi (1981) and Taniguchi (1987), respectively. We proposed a new dispar-

This work was supported by JSPS KAKENHI (17K12652).
2010 Mathematics Subject Classification: MSC-62M10, MSC-62M15, MSC-62M20.
Keywords: stationary process, prediction problem, interpolation problem, infinite variance, spectral analysis, parameter estimation, minimum contrast estimation, asymptotically efficiency, robustness.
* Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan
e-mail: yan.liu@i.kyoto-u.ac.jp
web: https://yliustat.github.io/

ity from the concept to minimize the interpolation error in Suto, Liu and Taniguchi (2016), and to minimize the prediction and interpolation error in a general norm in Liu (2017a). In addition, a new method to nonparametrically estimate the spectral density was proposed in Liu (2017b). A general discussion on the statistical inference for time series can be found in Taniguchi and Kakizawa (2000). The new disparity based on the minimization of the prediction and interpolation error in Liu (2017a) relates to the power divergence in Rényi (1961), Csiszár (1975) and Fujisawa and Eguchi (2008), which was proposed for the density estimation of independent and identically distributed random variables. The disparity has the robustness to the outliers and heavy contamination. To precisely understand the estimation procedure by our new disparity, we first investigate the fundamental properties of the contrast functions. The new functions are not contained in the class of either location or scale disparities. We focus on the asymptotic behaviors of the minimum contrast estimator based on the new disparity when it is applied to the case of the stationary process with finite variance innovations and infinite variance innovations. Under both situations, the estimator is shown to be asymptotically consistent. The asymptotic distribution of the estimator depends on the assumptions of the stochastic process. In particular, the estimator is robust against the fourth order cumulant when the process is Gaussian. Although it is shown that the Whittle estimator is asymptotically efficient in the sense that the family of parametric spectral densities is truly specified, the new class contains robust members to the randomly missing observations from the stationary process.

In Section 2, we discuss the prediction and interpolation problem for the stationary process. In Section 3, we introduce the minimum disparity functionals for parameter estimation of parametric models. In Section 4, the asymptotic behavior of the estimator minimizing the disparity is studied. In Section 5, we discuss the robustness and asymptotic efficiency case of the estimator based on the disparity.

2. Prediction and interpolation problem

Let us consider a harmonizable symmetric α-stable process $\{X(t);\ t \in \mathbb{Z}\}$. A stochastic process is called harmonizable if it has the representation

$$X(t) = \int_{-\pi}^{\pi} e^{it\lambda} dz(\lambda), \tag{2.1}$$

where $z(\lambda)$ is an independent increment process. Let Λ be $\Lambda = (-\pi, \pi]$. In general, the process $\{z(\lambda);\ \lambda \in \Lambda\}$ is a complex-valued stable independent increment process such that the characteristic function $\phi_{\lambda_1,\dots,\lambda_l}(s_1,\dots,s_l)$ of the joint distribution $(z(\lambda_1), z(\lambda_2), \cdots, z(\lambda_l))$ is expressed by

$$\phi_{\lambda_1,\dots,\lambda_l}(s_1,\dots,s_l) = \exp\Big\{-\sum_{j=1}^{l}\int_{-\pi}^{\pi}\Big|\mathrm{Re}\Big(\sum_{k=j}^{l} s_k e^{-i\theta}\Big)\Big|^{\alpha}\Big(F(\lambda_j, d\theta) - F(\lambda_{j-1}, d\theta)\Big)\Big\},$$

for $s_1,\dots,s_l \in \mathbb{C}$. If for any fixed $\lambda \in \Lambda$, the random variables $z(\lambda)$ and $z(\lambda)e^{i\omega}$ have the same distribution for any $\omega \in \Lambda$, then the harmonizable stable process $\{X(t)\}$ is stationary and its joint characteristic function is

$$\varphi_{t_1,\dots,t_l}(s_1,\dots,s_l) = \exp\Big\{-\int_{-\pi}^{\pi}\Big|\sum_{j=1}^{l} s_j e^{it_j\lambda}\Big|^{\alpha} G(d\lambda)\Big\} \tag{2.2}$$

for a non-negative bounded non-decreasing function G.

As a well-known result of (2.2), the Gaussian process is considered so far. Let $\alpha = 2$. It is easy to see that for any $l \in \mathbb{Z}$, the joint characteristic function $\varphi_{t_1,\ldots,t_l}(s_1, \ldots, s_l)$ is that of multivariate normal if $s_1, \ldots, s_l \in \mathbb{R}$, and that of multivariate complex normal if $s_1, \ldots, s_l \in \mathbb{C}$. In addition, for any $t_1, t_2 \in \mathbb{Z}$, $s \in \mathbb{C}$,

$$\varphi_{t_1}(s) = \varphi_{t_2}(s), \tag{2.3}$$

and for any $t_1, t_2 \in \mathbb{Z}$ such that $t_1 - t_2 = k$, there exists a function H such that

$$H(k) = \varphi_{t_1,t_2}(s_1, s_2). \tag{2.4}$$

From equations (2.3) and (2.4), we can see that the Gaussian process is second order stationary.

First, we focus on the prediction problem of stationary process. Let $[X(t)]$ denote the linear space spanned by all random variables $X(t)$ $(-\infty < t < \infty)$ and closed with respect to the mean convergence. Further, let $[X(t);\, t \leq a]$ be the subspace of $[X(t)]$ spanned by all random variables $X(t)$ such that $t \leq a$.

Definition 2.1 *A stationary process $\{X(t)\}$ is said to admit a prediction if there exists a linear operator A_0 from $[X(t)]$ onto $[X(t);\, t \leq 0]$ such that*

(i) *$A_0X = X$ whenever $X \in [X(t);\, t \leq 0]$,*

(ii) *if for every $Y \in [X(t);\, t \leq 0]$, the random variables X and Y are independent, then $A_0X = 0$,*

(iii) *for every $X \in [X(t)]$ and $Y \in [X(t);\, t \leq 0]$, the random variables $X - A_0X$ and Y are independent.*

This definition could be regarded as an extension from the original prediction problem (or interpolation problem) considered in Kolmogorov (1941b) and Kolmogorov (1941a).

A stationary process $\{X(t)\}$ admitting a prediction is called *deterministic* if $A_0X = X$ for every $X \in [X(t)]$. In addition, a stationary process $\{X(t)\}$ admitting a prediction is called *completely nondeterministic* if $\lim_{t\to-\infty} A_tX = 0$ for every $X \in [X(t)]$.

Theorem 2.2 (Urbanik (1967)) *A nontrivial completely nondeterministic process $\{X(t)\}$ is Gaussian if and only if for every $Y \in [X(t)]$ the process $\{T_tY\}$ admits a prediction, where T_t denotes a shift preserving the probability distribution such that $X(t) = T_tX(0)$.*

Theorem 2.2 shows that the definition of "admit a prediction" is restrictive and only useful for the Gaussian stationary process. To expand the scope of the prediction problem, a prediction is instead defined by a concept of the least prediction error, which is evaluated by an adequate norm.

A one-step ahead predictor of $X(0)$ is defined as a random variable $Y \in [X(t);\, t \leq -1]$ and the predictor error of the predictor Y is defined as $\|X(0) - Y\|$. An optimal predictor $Z \in [X(t);\, t \leq -1]$ is defined as the one which satisfies

$$\|X(0) - Z\| \leq \|X(0) - Y\|, \qquad \forall Y \in [X(t);\, t \leq -1]. \tag{2.5}$$

The optimal predictor defined in (2.5) is not the same as the one defined in Definition 2.1 except for the Gaussian process, since the "independence" described in (ii) and (iii) is too strong.

To see how the new definition (2.5) works for the harmonizable stable process (2.1), we have to characterize the intrinsic properties of the harmonizable stable process. Remember that the characteristic function of a complex-valued isotropic random variable which has a symmetric stable distribution with exponent α is expressed as $\exp\big(-b|s|^\alpha\big)$. A natural way to define the length $\|X\|$ of the random variable X is to let $\|X\| = b^{1/\alpha}$ for $1 \leq \alpha \leq 2$; $\|X\| = b$ for $0 < \alpha < 1$. If two random variables X_1 and X_2 are independent in this metric space, then the following equalities hold:

$$
\begin{aligned}
\|X_1 + X_2\| &= \|X_1\| + \|X_2\|, && \text{if } 0 < \alpha < 1,\\
\|X_1 + X_2\|^\alpha &= \|X_1\|^\alpha + \|X_2\|^\alpha, && \text{if } 1 \leq \alpha \leq 2.
\end{aligned}
$$

Let L^α be the space of αth power integrable Borel functions with respect to $d\|z\|$ for $0 < \alpha < 1$ and to $d\|z\|^\alpha$ for $1 \leq \alpha \leq 2$. For any function $h \in L^\alpha$, it holds that

$$
\Big\|\int_{-\pi}^{\pi} h(\lambda)dz(\lambda)\Big\| = \begin{cases} \int_{-\pi}^{\pi}|h(\lambda)|^\alpha d\|z\|, & \text{for } 0 < \alpha < 1,\\[1em] \int_{-\pi}^{\pi}|h(\lambda)|^\alpha d(\|z\|^\alpha), & \text{for } 1 \leq \alpha \leq 2. \end{cases} \tag{2.6}
$$

Denote by $L^\alpha(S)$ the completion of the linear hull of the set S in the space L^α. Especially, $L^\alpha(e^{it};\ t \in \mathbb{Z})$ is the completion of the linear hull of the set $\{e^{it};\ t \in \mathbb{Z}\}$. We see that the spaces of $L^\alpha(e^{it};\ t \in \mathbb{Z})$ and $\{X(t);\ t \in Z\}$ are isometric from (2.6). To interpret the prediction problem in the frequency domain, we use the notation $f(\lambda)$ in general as the spectral density induced by the measure of the space L^α.

Now let us formulate the prediction problem to find the optimal predictor along the equation (2.5). Let $Z_1 = \{x \in \mathbb{Z};\ x \leq -1\}$. From the isometry property above, we have

$$
\|X(0) - Y\| = \int_{-\pi}^{\pi} |1 - \psi(\lambda)|^\alpha f(\lambda)d\lambda,
$$

where $Y \in [X(t);\ t \leq -1]$ and $\psi(\lambda)$ is in the space $L^\alpha(S_1)$, which is generated by the set $S_1 = \{e^{ij\lambda};\ j \in Z_1\}$. The prediction problem is to solve the following minimization problem:

$$
\min_{\psi(\lambda)\in L^\alpha(S_1)} \int_{-\pi}^{\pi} |1 - \psi(\lambda)|^\alpha f(\lambda)d\lambda. \tag{2.7}
$$

When $\alpha = 2$, the problem (2.7) is surprisingly solved in 1915 by Szegö (1915), in which the formula is given by

$$
\min_{\phi(\lambda)\in L^2(S_1)} \int_{-\pi}^{\pi} |1 - \phi(\lambda)|^2 f(\lambda)d\lambda = 2\pi \exp\Big(\frac{1}{2\pi}\int_{-\pi}^{\pi} \log f(\lambda)d\lambda\Big), \tag{2.8}
$$

although the problem is later reformulated in time domain by Kolmogorov in his papers Kolmogorov (1941a) and Kolmogorov (1941b). A simpler analytic derivation of the formula (2.8), for example, can be found in Brockwell and Davis (1991). A general proof, discussed under the condition that $\int_{-\pi}^{\pi} \log f(\lambda)d\lambda > -\infty$, can be found in Hannan (1970) and Ash and Gardner (1975).

Now, we discuss the problem (2.7) in general. Under the condition that $\int_{-\pi}^{\pi} \log f(\lambda)d\lambda > -\infty$, the function $(1/\alpha)\log f$ is integrable and thus has the formal Fourier series representation:

$$
\frac{1}{\alpha}\log f(\lambda) \sim \sum_{j=-\infty}^{\infty} a_j e^{ij\lambda}, \tag{2.9}
$$

where

$$a_j = \frac{1}{2\pi\alpha} \int_{-\pi}^{\pi} \log f(\lambda) e^{-ij\lambda} d\lambda.$$

The key representation of (2.9) makes the linear model of time series popular and meaningful. The statistical inference for the spectral density is discussed in the next section. In a similar manner to the usual arguments for the prediction problem in L^2, the following result holds.

Theorem 2.3 (Hosoya (1982)) *Let $\{X(t)\}$ be a stationary harmonizable stable process with the spectral density $f(\lambda)$. If $\int_{-\pi}^{\pi} \log f(\lambda) d\lambda > -\infty$, there exists an optimal predictor $Z \in [X(t); t \leq -1]$ such that*

$$\|X(0) - Z\|^\alpha = 2\pi \exp\Big(\frac{1}{2\pi} \int_{-\pi}^{\pi} \log f(\lambda) d\lambda\Big).$$

If $\int_{-\pi}^{\pi} \log f(\lambda) d\lambda = -\infty$, there exists a predictor $Z \in [X(t); t \leq -1]$ such that $Z = X(0)$ a.e.

Theorem 2.3 shows that the optimal prediction error evaluated by the natural norm for the harmonizable stable process can be expressed by the spectral density $f(\lambda)$ induced by the natural measure of the space L^α. In addition, the αth power of the optimal prediction error does only depend on a functional expression of the spectral density $f(\lambda)$ but not on the index α.

Let us next introduce the interpolation problem along the same line with the prediction problem as a minimization problem. For the interpolation problem, it is first formulated in Kolmogorov (1941a) for the Gaussian stationary process. A more general interpolation problem is considered by Yaglom (1963) and Salehi (1979). However, from the point of view of the minimization problem, we only need to substitute an integer set S_2 for the set S_1 in equation (2.7). To be specific, the interpolation problem is formulated by

$$\min_{\psi(\lambda) \in L^\alpha(S_2)} \int_{-\pi}^{\pi} |1 - \psi(\lambda)|^\alpha f(\lambda) d\lambda, \tag{2.10}$$

where $S_2 = \{e^{ij\lambda}; j \neq 0\}$. The minimization problem (2.10) is considered in Weron (1985) and Miamee and Pourahmadi (1988).

Theorem 2.4 (Miamee and Pourahmadi (1988)) *Let $\{X(t)\}$ be a stationary harmonizable stable process with the spectral density $f(\lambda)$. If $0 \leq f \in L^1$ and $f^{-1/(\alpha-1)} \in L^1$, then there exists an optimal interpolator $Z \in [X(t); t \neq 0]$ such that*

$$\|X(0) - Z\|^\alpha = \Big(\int_{-\pi}^{\pi} f(\lambda)^{-1/(\alpha-1)} d\lambda\Big)^{1-\alpha}. \tag{2.11}$$

In addition, the optimal interpolator $\psi(\lambda)$ is expressed as

$$\psi(\lambda) = 1 - \Big(\int_{-\pi}^{\pi} f(\lambda)^{-1/(\alpha-1)} d\lambda\Big)^{-1} f(\lambda)^{-1/(\alpha-1)}. \tag{2.12}$$

Theorem 2.4 shows the optimal interpolator $\psi(\lambda)$ and its interpolation error. The similarity between the prediction problem and the interpolation problem is that both optimal errors can be expressed in the functional form of the spectral density. The difference between them is that the interpolation error depends on the power considered in the norm for the minimization problem.

Let $\alpha = 2$. From equations (2.11) and (2.12) in Theorem 2.4, we can see that the optimal interpolator for the Gaussian process is

$$\psi(\lambda) = 1 - \Big(\int_{-\pi}^{\pi} f(\lambda)^{-1} d\lambda\Big)^{-1} f(\lambda)^{-1}, \tag{2.13}$$

and its interpolation error is

$$\|X(0) - Z\|^2 = \Big(\int_{-\pi}^{\pi} f(\lambda)^{-1} d\lambda\Big)^{-1}. \tag{2.14}$$

As a remark, these results are possible to be extended to the interpolation problem for multiple missing points. We describe this problem in the following.

Suppose there are p ($p < n$) missing points $X(t)$, $t \in Z_p = \{1, \dots, p\}$. Let $\mathcal{M}_p$ denote the closed linear manifold generated by $e^{-ij\lambda}$, $j \notin Z_p$. To formulate the problem evidently, define $\boldsymbol{e}(\lambda) = (e^{-i\lambda}, \dots, e^{-ip\lambda})'$ and $F(\lambda) = \boldsymbol{e}(\lambda)\boldsymbol{e}(\lambda)^*$. To find the optimal interpolator for the missing points is equivalent to seek a response function $\boldsymbol{h}(\lambda)$ minimizing

$$\mathrm{tr} \int_{-\pi}^{\pi} \big(\boldsymbol{e}(\lambda) - \boldsymbol{h}(\lambda)\big) f(\lambda) \big(\boldsymbol{e}(\lambda) - \boldsymbol{h}(\lambda)\big)^* d\lambda,$$

where $h_i(\lambda) \in \mathcal{M}_p$ ($i = 1, \dots, p$). From the results in L^2, $\big(\boldsymbol{e}(\lambda) - \boldsymbol{h}(\lambda)\big) f(\lambda)$ is orthogonal to $\mathcal{M}_p$, that is to say,

$$\int_{-\pi}^{\pi} \big(\boldsymbol{e}(\lambda) - \boldsymbol{h}(\lambda)\big) f(\lambda) e^{ik\lambda} d\lambda = \boldsymbol{0}, \quad k \notin Z_p.$$

Therefore, there exists a constant matrix C such that

$$\big(\boldsymbol{e}(\lambda) - \boldsymbol{h}(\lambda)\big) f(\lambda) = C\boldsymbol{e}(\lambda).$$

From the orthogonality in L^2 again, the interpolation error matrix Σ_f, i.e.,

$$\Sigma_f = \int_{-\pi}^{\pi} \big(\boldsymbol{e}(\lambda) - \boldsymbol{h}(\lambda)\big) f(\lambda) \boldsymbol{e}(\lambda)^* d\lambda = C,$$

which shows the matrix C real. Also, we have

$$\Sigma_f = \int_{-\pi}^{\pi} \big(\boldsymbol{e}(\lambda) - \boldsymbol{h}(\lambda)\big) f(\lambda) \big(\boldsymbol{e}(\lambda) - \boldsymbol{h}(\lambda)\big)^* d\lambda = C\Big(\int_{-\pi}^{\pi} f(\lambda)^{-1} F(\lambda)\, d\lambda\Big) C.$$

If the process is nondeterministic, then $\Sigma_f = C$ is nonsingular. Therefore we obtain

$$\Sigma_f = \Big(\int_{-\pi}^{\pi} f(\lambda)^{-1} F(\lambda) d\lambda\Big)^{-1}, \tag{2.15}$$

and

$$\boldsymbol{h}(\lambda) = \Big(I_p - \Big(\frac{1}{2\pi} \int_{-\pi}^{\pi} f(\lambda)^{-1} F(\lambda)\, d\lambda\Big)^{-1} f(\lambda)^{-1}\Big) \boldsymbol{e}(\lambda). \tag{2.16}$$

Therefore, both equations (2.15) and (2.16) in the interpolation problem for multiple missing points are extensions of equations (2.13) and (2.14). The disparity based on (2.15) is discussed in Suto, Liu and Taniguchi (2016).

3. Minimum disparity functionals

In this section, let us consider the disparity to estimate the parameter θ in our parametric model f_θ, which is motivated by the prediction and interpolation problem of stationary process. Hereafter we only consider the optimal predictor and interpolator as discussed in Section 2 and omit "optimal" for brevity.

In the case of the Gaussian stationary process, denote by σ^2 the one-step ahead prediction error, that is,

$$\sigma^2 = 2\pi \exp\Big(\frac{1}{2\pi}\int_{-\pi}^{\pi} \log f(\lambda) d\lambda\Big).$$

Formulating it in time domain, we have

$$X(t) - \sum_j a_j X(t-j) = \epsilon(t), \qquad \epsilon(t) \sim \text{i.i.d. } \mathcal{N}(0, \sigma^2). \tag{3.1}$$

This is generally called the autoregressive (AR) process in time series analysis. In practice, it is more natural to use for statistical inference a parametric submodel whose spectral density is f_θ, $\theta \in \Theta \subset \mathbb{R}^d$, instead of the direct use of the model (3.1). Still, it can be regarded as an extrapolation problem, an extension of the prediction problem, in the sense of the minimization problem (2.7). Consequently, the parameter θ can be estimated by minimizing the disparity D^P, i.e.,

$$D^P(f_\theta, f) = \int_{-\pi}^{\pi} f_\theta(\lambda)^{-1} f(\lambda) d\lambda. \tag{3.2}$$

The usage of this form (3.2) to estimate the parameter of time series is proposed in Whittle (1952) and it is well-known as the Whittle estimator.

Next we derive other disparities from the result (2.12) of the interpolation problem in L^α. Instead of the true spectral density $f(\lambda)$, it is natural to use a parametric spectral density f_θ for the optimal interpolator $\psi(\lambda)$. That is, the parametric interpolator ψ_θ is

$$\psi_\theta(\lambda) = 1 - \Big(\int_{-\pi}^{\pi} f_\theta(\lambda)^{-1/(\alpha-1)} d\lambda\Big)^{-1} f_\theta(\lambda)^{-1/(\alpha-1)}. \tag{3.3}$$

Plugging (3.3) in equation (2.10), we obtain that

$$\int_{-\pi}^{\pi} |1-\psi_\theta(\lambda)|^\alpha f(\lambda) d\lambda = \Big(\int_{-\pi}^{\pi} f_\theta(\lambda)^{-1/(\alpha-1)} d\lambda\Big)^{-\alpha} \int_{-\pi}^{\pi} f_\theta(\lambda)^{-\alpha/(\alpha-1)} f(\lambda) d\lambda.$$

For simplicity, let $p = -\alpha/(\alpha-1)$. Another disparity D^I can be defined by

$$D^I(f_\theta, f) = \Big(\int_{-\pi}^{\pi} f_\theta(\lambda)^{p+1} d\lambda\Big)^{-p/(p+1)} \int_{-\pi}^{\pi} f_\theta(\lambda)^p f(\lambda) d\lambda. \tag{3.4}$$

To unify the notation of (3.2) and (3.4), we introduce the disparity D as

$$D(f_\theta, f) = \int_{-\pi}^{\pi} a_\theta f_\theta(\lambda)^p f(\lambda) d\lambda, \tag{3.5}$$

where a_θ is expressed as

$$a_\theta = \begin{cases} \Big(\int_{-\pi}^{\pi} f_\theta^{p+1}(\lambda) d\lambda\Big)^{-p/(p+1)}, & \text{if } p \neq -1, \\ 1, & \text{if } p = -1. \end{cases}$$

In summary, the parameter θ in the parametric model f_θ can be estimated by optimizing the disparity (3.5).

As a remark, several types of disparities have been proposed in the literature. We review two main types developed for the parameter estimation of time series in the following:

- Location disparity $D^L(f_\theta, f)$.

  The location disparity $D^L(f_\theta, f)$ is defined as

  $$D^L(f_\theta, f) = \int_{-\pi}^{\pi} \Psi(f_\theta(\lambda))^2 - 2\Psi(f_\theta(\lambda))\Psi(f(\lambda))d\lambda,$$

  where the function Ψ is some appropriate bijective function. Special cases (i) $\Psi(x) = \log x$; (ii) $\Psi(x) = 1$ are some choice for practical use.

- Scale disparity $D^S(f_\theta, f)$.

  The scale disparity $D^S(f_\theta, f)$ is defined as

  $$D^S(f_\theta, f) = \int_{-\pi}^{\pi} K\left(\frac{f_\theta(\lambda)}{f(\lambda)}\right)d\lambda,$$

  where K is sufficiently smooth with its minimum at 1. Without loss of generality, the function K can be replaced by some function $\tilde{K}$ such that

  $$\tilde{K}\left(\frac{f_\theta(\lambda)}{f(\lambda)} - 1\right) = K\left(\frac{f_\theta(\lambda)}{f(\lambda)}\right).$$

  In this case, the minimizer is 0 for $\tilde{K}$. Examples of K are

  (i) $K(x) = \log x + 1/x$;
  (ii) $K(x) = -\log x + x$;
  (iii) $K(x) = (\log x)^2$;
  (iv) $K(x) = (x^\alpha - 1)^2$;
  (v) $K(x) = x \log x - x$;
  (vi) $K(x) = \log\big((1-\alpha) + \alpha x\big) - \alpha \log x, \quad \alpha \in (0, 1)$.

  Some of these examples are applied to the parameter estimation problem and others are applied to the discriminant analysis.

It is easy to see that the new disparity (3.5) based on the prediction and interpolation problem is quite different from the location disparity $D^L(f_\theta, f)$ and the scale disparity $D^S(f_\theta, f)$. Although some special case of new disparity may be contained in these two types of disparities, the disparity in general is not contained in the class of either location or scale disparities. This motivates us to investigate (3.5) further.

We discuss the fundamental properties of the disparity under the following assumptions. Denote by $\mathcal{F}(\Theta)$ the set of spectral densities indexed by parameter θ. Let θ_0 be the true parameter in the parameter space Θ, i.e., $f = f_{\theta_0} \in \mathcal{F}(\Theta)$.

Assumption 1

(i) *The parameter space* Θ *is a compact subset of* $\mathbb{R}^d$.

(ii) *If* $\theta_1 \neq \theta_2$, *then* $f_{\theta_1} \neq f_{\theta_2}$ *on a set of positive Lebesgue measure.*

(iii) *The parametric spectral density* $f_\theta(\lambda)$ *is three times continuously differentiable with respect to* θ *and the second derivative* $\frac{\partial^2}{\partial\theta\partial\theta^T} f_\theta(\lambda)$ *is continuous in* λ.

Theorem 3.1 *Under Assumption 1 (i) and (ii), we have the following results:*

(i) *If* $p > 0$, *then* θ_0 *maximizes the disparity* $D(f_\theta, f)$.

(ii) *If* $p < 0$, *then* θ_0 *minimizes the disparity* $D(f_\theta, f)$.

Theorem 3.1 can be shown by Hölder's inequality. The maximizer (or minimizer) θ_0 in case (i) (or case (ii)) is found by the equality case and Assumption 1 (ii).

The convexity of the disparity (3.5) can be shown by the analytic approach under Assumption 1 (iii). To present the result, we need some notation preparation. Denote by ∂ the partial derivative with respect to θ. For $1 \leq i \leq d$, let ∂_i be the partial derivative with respect to θ_i. We simplify the notation for differentiation as follows:

$$A_1(\theta) = \int_{-\pi}^{\pi} f_\theta(\lambda)^{p+1} d\lambda, \qquad B_1(\theta)_i = f_\theta(\omega)^{p-1}\partial_i f_\theta(\omega),$$
$$A_2(\theta)_i = \int_{-\pi}^{\pi} f_\theta(\lambda)^{p}\partial_i f_\theta(\lambda) d\lambda, \qquad B_2(\theta) = f_\theta(\omega)^p,$$
$$A_3(\theta)_{ij} = \int_{-\pi}^{\pi} f_\theta(\lambda)^{p-1}\partial_i f_\theta(\lambda)\partial_j f_\theta(\lambda) d\lambda, \quad C_1(\theta) = -\frac{p}{p+1}\left(\int_{-\pi}^{\pi} f_\theta(\lambda)^{p+1} d\lambda\right)^{-\frac{2p+1}{p+1}}.$$

Theorem 3.2 *Under Assumption 1, we have the following results:*

(i) *if* $p > 0$, *then the disparity* $D(f_\theta, f)$ *is convex upward with respect to* θ.

(ii) *if* $p < 0$, *then the disparity* $D(f_\theta, f)$ *is convex downward with respect to* θ.

Theorem 3.2 is shown by noting that

$$\partial_i\partial_j D(f_\theta, f)\Big|_{\theta=\theta_0} = (p+1)C_1(\theta_0)\Big(A_1(\theta_0)A_3(\theta_0)_{ij} - A_2(\theta_0)_i A_2(\theta_0)_j\Big).$$

By the Cauchy-Bunyakovsky inequality, the matrix $\left(\partial_i\partial_j D(f_\theta, f)\Big|_{\theta=\theta_0}\right)$ is shown to be positivity definite if $p < 0$ and negative definite if $p > 0$. Thus, we obtain the conclusion.

4. Parameter estimation problem

In this section, we discuss the parameter estimation problem for our parametric model f_θ by optimizing the disparity discussed before. We adopt the case $p < 0$ in which the true parameter minimizes the disparity since the discussion for the other case is parallel. Let the functional T be defined as

$$D(f_{T(f)}, f) = \min_{\theta\in\Theta} D(f_\theta, f), \qquad \text{for every } f \in \mathcal{F}(\Theta). \tag{4.1}$$

The analytic properties of the functional T can be found in the following result.

Theorem 4.1 *Under Assumption 1, the following results hold.*

(i) *For every $f \in \mathcal{F}(\Theta)$, there exists a value $T(f) \in \Theta$.*

(ii) *If $T(g)$ is unique and if $f_n \xrightarrow{L^2} f$, then $T(f_n) \to T(f)$ as $n \to \infty$.*

(iii) *$T(f_\theta) = \theta$ for every $\theta \in \Theta$.*

Define $h(\theta)$ as $h(\theta) = D(f_\theta, f)$. The assertion (i) is almost obvious but we have to note that

$$h(\theta) \le \Big(\int_{-\pi}^{\pi} f(\lambda)^{p+1} d\lambda\Big)^{\frac{1}{p+1}} \le C,$$

and by Lebesgue's dominated convergence theorem, it holds that

$$|h(\theta_n) - h(\theta)| \le \Big|\int_{-\pi}^{\pi} (a_{\theta_n} f_{\theta_n}^{\alpha}(\lambda) - a_\theta f_\theta^{\alpha}(\lambda)) g(\lambda) d\lambda\Big| \to 0$$

for any convergence sequence $\{\theta_n \in \Theta; \theta_n \to \theta\}$. The existence of the minimizer follows from the continuity of $h(\theta)$.

To prove the assertion (ii), let h_n be $h_n(\theta) = D(f_\theta, f_n)$. Note that as $f_n \xrightarrow{L^2} f$, it holds that

$$\begin{aligned}
\lim_{n\to\infty} \sup_{\theta\in\Theta} |h_n(\theta) - h(\theta)| &= \lim_{n\to\infty} \sup_{\theta\in\Theta} \Big|\int_{-\pi}^{\pi} a_\theta f_\theta(\lambda)^p (f_n(\lambda) - f(\lambda)) d\lambda\Big| \\
&\le \lim_{n\to\infty} \sup_{\theta\in\Theta} \Big|\int_{-\pi}^{\pi} a_\theta^2 f_\theta(\lambda)^{2p} d\lambda \int_{-\pi}^{\pi} (f_n(\lambda) - f(\lambda))^2 d\lambda\Big|^{1/2} \\
&\le C \lim_{n\to\infty} \sup_{\theta\in\Theta} \Big|\int_{-\pi}^{\pi} (f_n(\lambda) - f(\lambda))^2 d\lambda\Big|^{1/2} \\
&= 0.
\end{aligned}$$

The assertion (iii) follows from Theorem 3.1 (ii).

Now we move on to the Hessian matrix of the estimation procedure (4.1). This can be regarded as a continuation of Theorem 3.2 from the view of estimation.

Theorem 4.2 *Under Assumption 1, we have*

$$T(g_n) = T(g) - \int_{-\pi}^{\pi} \rho(\lambda)(f_n(\lambda) - f(\lambda)) d\lambda$$

for every spectral density sequence $\{f_n\}$ satisfying $f_n \xrightarrow{L^2} f$, where

$$\rho(\omega) = \Big(A_1(\theta_0)A_3(\theta_0) - A_2(\theta_0)A_2(\theta_0)^T\Big)^{-1} \Big(A_1(\theta_0)B_1(\theta_0) - A_2(\theta_0)B_2(\theta_0)\Big).$$

Theorem 4.2 can be shown by the mean value theorem. Note that there exists a $\theta^* \in \mathbb{R}^d$ on the line joining θ_n and θ_0 such that

$$T(f_n) - T(f) = \Big\{(\alpha + 1)C_1(\theta)\Big(A_1(\theta)A_3(\theta) - A_2(\theta)A_2(\theta)^T\Big)\Big|_{\theta=\theta^*}\Big\}^{-1} \int_{-\pi}^{\pi} \Big(A_1(\theta_0)B_1(\theta_0) - A_2(\theta_0)B_2(\theta_0)\Big)(f_n - f) d\lambda.$$

From the fact that

$$(\alpha+1)C_1(\theta)\Big(A_1(\theta)A_3(\theta)-A_2(\theta)A_2(\theta)^T\Big)\Big|_{\theta=\theta^*} \\ -\Big((\alpha+1)C_1(\theta_0)(A_1(\theta_0)A_3(\theta_0)-A_2(\theta_0)A_2(\theta_0)^T\Big)$$

is bounded by $C|T(f_n)-T(f)|$, the conclusion is obtained.

The essence of the estimation has been considered in Theorems 4.1 and 4.2. Below we construct the nonparametric estimator for the spectral density f in both cases of a stochastic process with finite variance innovations and infinite variance innovations.

4.1. Finite variance innovations

We suppose the stationary process has the following representation:

$$X(t)=\sum_{j=0}^{\infty}g_j\epsilon(t-j),\quad t\in\mathbb{Z},$$

where $\{\epsilon(t)\}$ is a stationary innovation process with finite fourth moment $E\epsilon(t)^4<\infty$ and satisfies $E[\,\epsilon(t)\,]=0$ and $\mathrm{Var}[\,\epsilon(t)\,]=\sigma^2$ with $\sigma^2>0$. We impose the following regularity conditions.

Assumption 2 *For all $|z|\le 1$, there exist $C<\infty$ and $\delta>0$ such that*

(i) $\sum_{j=0}^{\infty}(1+j^2)|g_j|\le C$;

(ii) $\left|\sum_{j=0}^{\infty}g_jz^j\right|\ge\delta$;

(iii) $\sum_{t_1,t_2,t_3=-\infty}^{\infty}|Q_\epsilon(t_1,t_2,t_3)|<\infty$, *where $Q_\epsilon(t_1,t_2,t_3)$ is the fourth order cumulant of $\epsilon(t)$, $\epsilon(t+t_1)$, $\epsilon(t+t_2)$ and $\epsilon(t+t_3)$.*

Assumption 2 (iii) guarantees the existence of a fourth order spectral density

$$\tilde{Q}_\epsilon(\omega_1,\omega_2,\omega_3)=\Big(\frac{1}{2\pi}\Big)^3\sum_{t_1,t_2,t_3=-\infty}^{\infty}Q_\epsilon(t_1,t_2,t_3)e^{-i(\omega_1t_1+\omega_2t_2+\omega_3t_3)}.$$

Denote by $(X(1),\dots,X(n))$ the observations from the process $\{X(t)\}$. Let $I_{n,X}(\omega)$ be the periodogram of observations, that is,

$$I_{n,X}(\omega)=\frac{1}{2\pi n}\Big|\sum_{t=1}^{n}X(t)e^{it\omega}\Big|^2,\quad -\pi\le\omega\le\pi.$$

Now, under Assumption 1 we can define the estimator $\hat{\theta}_n$ based on (4.1) as

$$\hat{\theta}_n=\arg\min_{\theta\in\Theta}D(f_\theta,I_{n,X}).\tag{4.2}$$

Now we state the regularity conditions for the parameter estimation by $\hat{\theta}_n$.

Let $\mathcal{B}(t)$ denote the σ-field generated by $\epsilon(s)$ $(-\infty<s\le t)$.

Assumption 3

(i) *For each nonnegative integer* m *and* $\eta_1 > 0$,

$$\mathrm{Var}[E\big(\epsilon(t)\epsilon(t+m)|\mathcal{B}(t-\tau)\big)] = O(\tau^{-2-\eta_1})$$

uniformly in t.

(ii) *For any* $\eta_2 > 0$,

$$E|E\{\epsilon(t_1)\epsilon(t_2)\epsilon(t_3)\epsilon(t_4)|\mathcal{B}(t_1-\tau)\} - E\big(\epsilon(t_1)\epsilon(t_2)\epsilon(t_3)\epsilon(t_4)\big)| = O(\tau^{-1-\eta_2}),$$

uniformly in t_1, *where* $t_1 \le t_2 \le t_3 \le t_4$.

(iii) *For any* $\eta_3 > 0$ *and for any fixed integer* $L \ge 0$, *there exists* $B_{\eta_3} > 0$ *such that*

$$E[T(n,s)^2 \mathbb{1}\{T(n,s) > B_{\eta_3}\}] < \eta_3$$

uniformly in n *and* s, *where*

$$T(n,s) = \Big[\frac{1}{n}\sum_{r=0}^{L}\Big\{\sum_{t=1}^{n}\epsilon(t+s)\epsilon(t+s+r) - \sigma^2\delta(0,r)\Big\}^2\Big]^{1/2}.$$

Here, $\delta(s,t)$ *is the Kronecker delta.*

Assumption 3 is a sufficient condition under which the estimator (4.2) is asymptotically normal. The details for Assumption 3 can be found in Hosoya and Taniguchi (1982).

Theorem 4.3 *Suppose Assumptions 1–3 hold. As for the spectral density* $f \in \mathcal{F}(\Theta)$, *the estimator* $\hat{\theta}_n$ *defined by* (4.2) *has the following asymptotic properties.*

(i) $\hat{\theta}_n$ *converges to* θ_0 *in probability;*

(ii) *The distribution of* $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *is asymptotically normal with mean 0 and covariance matrix* $H(\theta_0)^{-1}V(\theta_0)H(\theta_0)^{-1}$, *where*

$$\begin{aligned}
H(\theta_0) = &\Big(\int_{-\pi}^{\pi} f_{\theta_0}(\omega)^p \partial f_{\theta_0}(\omega) d\omega\Big)\Big(\int_{-\pi}^{\pi} f_{\theta_0}(\omega)^p \partial f_{\theta_0}(\omega) d\omega\Big)^T \\
&- \int_{-\pi}^{\pi} f_{\theta_0}(\omega)^{p+1} d\omega \int_{-\pi}^{\pi} f_{\theta_0}(\omega)^{p-1}\Big(\partial f_{\theta_0}(\omega)\Big)\Big(\partial f_{\theta_0}(\omega)\Big)^T d\omega, \\
V(\theta_0) = 4\pi &\int_{-\pi}^{\pi}\Big(f_{\theta_0}(\omega)^p \partial f_{\theta_0}(\omega)\int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p+1} d\lambda - f_{\theta_0}(\omega)^{p+1}\int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^p \partial f_{\theta_0}(\lambda) d\lambda\Big) \\
&\Big(f_{\theta_0}(\omega)^p \partial f_{\theta_0}(\omega)\int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p+1} d\lambda - f_{\theta_0}(\omega)^{p+1}\int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^p \partial f_{\theta_0}(\lambda) d\lambda\Big)^T d\omega \\
&+ 2\pi \iint_{-\pi}^{\pi}\Big(f_{\theta_0}(\omega_1)^{p-1}\partial f_{\theta_0}(\omega_1)\int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p+1} d\lambda - f_{\theta_0}(\omega_1)^p \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^p \partial f_{\theta_0}(\lambda) d\lambda\Big) \\
&\Big(f_{\theta_0}(\omega_2)^{p-1}\partial f_{\theta_0}(\omega_2)\int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p+1} d\lambda - f_{\theta_0}(\omega_2)^p \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^p \partial f_{\theta_0}(\lambda) d\lambda\Big)^T \\
&\tilde{Q}_X(-\omega_1, \omega_2, -\omega_2) d\omega_1 d\omega_2. \qquad (4.3)
\end{aligned}$$

Here, $\tilde{Q}_X(\omega_1,\omega_2,\omega_3) = A(\omega_1)A(\omega_2)A(\omega_3)A(-\omega_1-\omega_2-\omega_3)\tilde{Q}_\epsilon(\omega_1,\omega_2,\omega_3)$ *and* $A(\omega) = \sum_{j=0}^{\infty} g_j \exp(ij\omega)$.

The asymptotic variance (4.3) of the estimator $\hat{\theta}_n$ seems extremely complex. Sometimes we are not interested in all disparities (3.5) for different p but some. Especially, when $p = 1$, the disparity corresponds to the prediction error.

Let us give the result for the Gaussian stationary process under the special case $p = -1$. From Theorem 2.3, there exists a real constant $\tilde{\sigma}^2 > 0$ such that

$$\tilde{\sigma}^2 = 2\pi \exp\Big(\frac{1}{2\pi}\int_{-\pi}^{\pi} \log f_\theta(\lambda) d\lambda\Big).$$

Thus we have the following result that if the parametric model f_θ is innovation-free, i.e., independent of $\tilde{\sigma}^2$, then

$$\int_{-\pi}^{\pi} f_\theta(\lambda)^{-1} \partial f_\theta(\lambda) d\lambda = \partial \int_{-\pi}^{\pi} \log f_\theta(\lambda) d\lambda = 0. \tag{4.4}$$

When $p = -1$, by (4.4), we have

$$H(\theta_0) = 2\pi \int_{-\pi}^{\pi} f_{\theta_0}(\omega)^{-2} \Big(\partial f_{\theta_0}(\omega)\Big)\Big(\partial f_{\theta_0}(\omega)\Big)^T d\omega.$$

Note that $\tilde{Q}_X(-\omega_1, \omega_2, -\omega_2) = 0$ for the Gaussian process. By (4.4) again, we have

$$V(\theta_0) = 16\pi^3 \int_{-\pi}^{\pi} f_{\theta_0}(\omega)^{-2} \Big(\partial f_{\theta_0}(\omega)\Big)\Big(\partial f_{\theta_0}(\omega)\Big)^T d\omega.$$

Therefore, the asymptotic covariance matrix for $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is

$$H(\theta_0)^{-1} V(\theta_0) H(\theta_0)^{-1} = 4\pi \Big(\int_{-\pi}^{\pi} f_{\theta_0}(\omega)^{-2} \Big(\partial f_{\theta_0}(\omega)\Big)\Big(\partial f_{\theta_0}(\omega)\Big)^T d\omega\Big)^{-1}. \tag{4.5}$$

Generally, the inverse of the right hand side of (4.5) is called the Gaussian Fisher information matrix in time series analysis. Let us denote it by $\mathcal{F}(\theta_0)$, i.e.,

$$\mathcal{F}(\theta_0) = \frac{1}{4\pi} \int_{-\pi}^{\pi} f_{\theta_0}^{-2}(\lambda) \partial f_{\theta_0}(\lambda) \partial f_{\theta_0}(\lambda)^T d\lambda. \tag{4.6}$$

An estimator $\hat{\theta}_n$ is said to be Gaussian asymptotically efficient if $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{F}(\theta_0)^{-1})$.

4.2. Infinite variance innovations

Next we consider the linear processes with infinite variance innovations. Suppose $\{X(t); t \in \mathbb{Z}\}$ is a stationary process

$$X(t) = \sum_{j=0}^{\infty} g_j \epsilon(t - j), \quad t \in \mathbb{Z},$$

where i.i.d. symmetric innovation process $\{\epsilon(t)\}$ satisfy the following assumptions.

Assumption 4 *For some $k > 0$, $\delta = 1 \wedge k$ and positive sequence a_n satisfying $a_n \uparrow \infty$, the coefficient g_j and the innovation process $\{\epsilon(t)\}$ have the following properties:*

(i) $\sum_{j=0}^{\infty} |j||g_j|^\delta < \infty$;

(ii) $E|\epsilon(t)|^k < \infty$;

(iii) *as* $n \to \infty$, $n/a_n^{2\delta} \to 0$;

(iv) $\lim_{x\to 0} \limsup_{n\to\infty} P\Big(a_n^{-2} \sum_{t=1}^n \epsilon(t)^2 \le x\Big) = 0$.

(v) *For some* $0 < \alpha < 2$*, the distribution of* $\epsilon(t)$ *is in the domain of normal attraction of a symmetric* α*-stable random variable* Y.

For Assumption 4, note that the positive sequence a_n can be specified from (v) by choosing $a_n = n^{1/\alpha}$ for $n \ge 1$. (See Feller (1968), Bingham et al. (1987).)

An issue concerning the infinite variance innovations is the periodogram $I_{n,X}(\omega)$ is not well-defined in this case. For this type of stationary process, the self-normalized periodogram $\tilde{I}_{n,X}(\omega) = I_{n,X}(\omega)/\sum_{t=1}^n X(t)^2$ is substituted for the periodogram $I_{n,X}(\omega)$. Let us define the power transfer function $f(\omega)$ by

$$f(\omega) = \Big|\sum_{j=0}^{\infty} g_j e^{ij\omega}\Big|^2, \quad \omega \in [-\pi, \pi].$$

Again it is possible to formulate the approach to estimate parameter by minimizing the disparity (4.1). That is, we fit a parametric model f_θ to the self-normalized periodogram $\tilde{I}_{n,X}(\omega)$:

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} D(f_\theta, \tilde{I}_{n,X}). \tag{4.7}$$

For the case of infinite variance innovations, we introduce the scale constant C_α appearing in the asymptotic distribution, i.e.,

$$C_\alpha = \begin{cases} \frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)}, & \text{if } \alpha \ne 1, \\ \frac{2}{\pi}, & \text{if } \alpha = 1. \end{cases}$$

Theorem 4.4 *Suppose Assumptions 1, 2 and 4 hold. As for the power transfer function* $f \in \mathcal{F}(\Theta)$*, the estimator* $\hat{\theta}_n$ *defined by* (4.7) *has the following asymptotic properties.*

(i) $\hat{\theta}_n$ *converges to* θ_0 *in probability;*

(ii) *It holds that*

$$\Big(\frac{n}{\log n}\Big)^{1/\alpha}(\hat{\theta}_n - \theta_0) \to 4\pi H^{-1}(\theta_0)\sum_{k=1}^{\infty}\frac{Y_k}{Y_0}V_k(\theta_0)$$

in law, where $H(\theta_0)$ *is the same as in Theorem 4.3,*

$$\begin{aligned} V_k(\theta_0) = \Big(\int_{-\pi}^{\pi} f_{\theta_0}(\omega)^p \partial f_{\theta_0}(\omega)d\omega\Big)\Big(\int_{-\pi}^{\pi} f_{\theta_0}(\omega)^{p+1}\cos(k\omega)d\omega\Big) \\ - \Big(\int_{-\pi}^{\pi} f_{\theta_0}(\omega)^{p+1}d\omega\Big)\Big(\int_{-\pi}^{\pi} f_{\theta_0}(\omega)^p \partial f_{\theta_0}(\omega)\cos(k\omega)d\omega\Big), \end{aligned}$$

and $\{Y_k\}_{k=0,1,\ldots,}$ *are mutually independent random variables.* Y_0 *is* $\alpha/2$*-stable with scale* $C_{\alpha/2}^{-2/\alpha}$ *and* Y_k $(k \ge 1)$ *is* α*-stable with scale* $C_\alpha^{-1/\alpha}$.

Theorem 4.4 shows the asymptotic distribution of the estimator $\hat{\theta}_n$. This is quite different from that in the case of finite variance innovation case.

5. Robustness and Efficiency

In this section, we discuss the robustness and asymptotic efficiency of the estimator $\hat{\theta}_n$ in (4.2). Especially, the estimator $\hat{\theta}_n$ is robust in their asymptotic distribution in the sense that it does not depend on the fourth cumulant under some suitable conditions; and robust from its intrinsic feature against randomly missing observations from time series. On the other hand, we discuss the asymptotic efficient estimator in the new class of disparities and illustrate it by some examples.

5.1. Robustness

The estimator $\hat{\theta}_n$ is said to be robust against the fourth cumulant if the asymptotic variance of $\hat{\theta}_n$ does not depend on $\tilde{Q}_X(-\omega_1, \omega_2, -\omega_2)$. The estimator $\hat{\theta}_n$ is robust against the fourth cumulant under the following assumption.

Assumption 5 *For the innovation process $\{\epsilon(t)\}$, suppose the fourth order cumulant* $\mathrm{cum}\big(\epsilon(t_1), \epsilon(t_2), \epsilon(t_3), \epsilon(t_4)\big)$ *satisfies*

$$\mathrm{cum}\big(\epsilon(t_1), \epsilon(t_2), \epsilon(t_3), \epsilon(t_4)\big) = \begin{cases} \kappa_4 & \textit{if } t_1 = t_2 = t_3 = t_4, \\ 0 & \textit{otherwise.} \end{cases}$$

If Assumption 5 holds, the fourth order spectral density $\tilde{Q}_X(\omega_1, \omega_2, \omega_3)$ of $\{X(t)\}$ becomes

$$\tilde{Q}_X(\omega_1, \omega_2, \omega_3) = (2\pi)^{-3}\kappa_4 A(\omega_1)A(\omega_2)A(\omega_3)A(-\omega_1 - \omega_2 - \omega_3). \tag{5.1}$$

Thus, we obtain the following theorem.

Theorem 5.1 *Suppose Assumptions 1, 2, 3 and 5 hold. As for the spectral density $f \in \mathcal{F}(\Theta)$, The distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically Gaussian with mean 0 and variance $H(\theta_0)^{-1}\tilde{V}(\theta_0)H(\theta_0)^{-1}$, where*

$$\tilde{V}(\theta_0) = 4\pi \int_{-\pi}^{\pi} \Big(f_{\theta_0}(\omega)^p \partial f_{\theta_0}(\omega) \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p+1} d\lambda - f_{\theta_0}(\omega)^{p+1} \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^p \partial f_{\theta_0}(\lambda) d\lambda\Big)$$
$$\Big(f_{\theta_0}(\omega)^p \partial f_{\theta_0}(\omega) \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p+1} d\lambda - f_{\theta_0}(\omega)^{p+1} \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^p \partial f_{\theta_0}(\lambda) d\lambda\Big)^T d\omega. \tag{5.2}$$

Assumptions 5 seems strong. However, for example, the Gaussian process always satisfies Assumption 5. In practice, modeling a process is usually up to second order. Making an assumption on simultaneous fourth cumulants covers a sufficiently large family of models.

Let us compare equation (5.2) with equation (4.3) in Theorem 4.3. The term with the fourth order spectral density $\tilde{Q}_X(\omega_1, \omega_2, \omega_3)$ of $\{X(t)\}$ vanishes. This fact is well known for the case $p = -1$, i.e., the Whittle likelihood estimator is robust against the fourth cumulant. We have shown that the robustness against the fourth cumulant also holds for any $p \in \mathbb{R} \setminus \{0\}$.

Now let us explain how to show Theorem 5.1. By the expression (5.1), we have

$$\tilde{Q}_X(-\omega_1, \omega_2, -\omega_2) = \frac{\kappa_4}{2\pi\sigma^4}\Big(\frac{\sigma^2}{2\pi}\Big)^2 A(-\omega_1)A(\omega_2)A(-\omega_2)A(\omega_1) = \frac{\kappa_4}{2\pi\sigma^4} f_{\theta_0}(\omega_1) f_{\theta_0}(\omega_2). \tag{5.3}$$

By (5.3), the second term in (4.3) can be evaluated by

$$
\begin{aligned}
& 2\pi \iint_{-\pi}^{\pi} \Big(f_{\theta_0}(\omega_1)^{p-1} \partial f_{\theta_0}(\omega_1) \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p+1} d\lambda - f_{\theta_0}(\omega_1)^{p} \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p} \partial f_{\theta_0}(\lambda) d\lambda \Big) \\
& \Big(f_{\theta_0}(\omega_2)^{p-1} \partial f_{\theta_0}(\omega_2) \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p+1} d\lambda - f_{\theta_0}(\omega_2)^{p} \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p} \partial f_{\theta_0}(\lambda) d\lambda \Big)^T \\
& \tilde{Q}_X(-\omega_1, \omega_2, -\omega_2) d\omega_1 d\omega_2 \\
= \; & \frac{\kappa_4}{\sigma^4} \Big(\int_{-\pi}^{\pi} f_{\theta_0}(\omega_1)^p \partial f_{\theta_0}(\omega_1) d\omega_1 \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p+1} d\lambda - \int_{-\pi}^{\pi} f_{\theta_0}(\omega_1)^{p+1} d\omega_1 \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p} \partial f_{\theta_0}(\lambda) d\lambda \Big) \\
& \Big(\int_{-\pi}^{\pi} f_{\theta_0}(\omega_2)^p \partial f_{\theta_0}(\omega_2) d\omega_2 \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p+1} d\lambda - \int_{-\pi}^{\pi} f_{\theta_0}(\omega_2)^{p+1} d\omega_2 \int_{-\pi}^{\pi} f_{\theta_0}(\lambda)^{p} \partial f_{\theta_0}(\lambda) d\lambda \Big)^T \\
= \; & O_d,
\end{aligned}
$$

where O_d denotes the $d \times d$ zero matrix. This shows why the second term in equation (4.3) vanishes when we take the prediction and interpolation error as a disparity.

Not only robust against the fourth cumulant, the estimation by the disparity (4.2) is also robust against randomly missing observations. This property can be illustrated by some numerical simulations. For conceptual understanding, let $\{Y(t)\}$ be an amplitude modulated series, that is,

$$Y(t) = X(t)Z(t),$$

where

$$Z(t) = \begin{cases} 1, & Y(t) \text{ is observed} \\ 0, & \text{otherwise.} \end{cases}$$

If we define $P(Z(t) = 1) = q$ and $P(Z(t) = 0) = 1 - q$, then the spectral density f_Y for the series $\{Y(t)\}$ is represented by

$$f_Y(\omega) = q^2 f_X(\omega) + q \int_{-\pi}^{\pi} a(\omega - \alpha) f_X(\alpha) d\alpha, \tag{5.4}$$

where $a(\omega) = (2\pi)^{-1} \sum_r a_r e^{ir\omega}$ with $a_r = q^{-1}\mathrm{Cov}(Z(t), Z(t+r))$. The spectral density f_Y from equation (5.4) can be considered as the original spectral density f_X heavily contaminated by a missing spectral density $\int_{-\pi}^{\pi} a(\omega - \alpha) f_X(\alpha) d\alpha$.

5.2. Asymptotic efficiency

As shown in (4.6), the variance of the estimator $\hat{\theta}_n$ minimizing prediction error is asymptotically $\mathcal{F}(\theta_0)^{-1}$. Actually, it is well known in time series analysis that the Fisher information matrix for the Gaussian process is

$$\mathcal{F}(\theta_0) = \frac{1}{4\pi} \int_{-\pi}^{\pi} f_{\theta_0}^{-2}(\lambda) \partial f_{\theta_0}(\lambda) \partial f_{\theta_0}(\lambda)^T d\lambda,$$

which can be derived from the approximate maximum likelihood estimation. When the asymptotic variance of the estimator $\hat{\theta}_n$ attaining the Cramer-Rao lower bound, that is, the inverse matrix of Fisher information matrix $\mathcal{F}(\theta)^{-1}$, the estimator $\hat{\theta}_n$ is called asymptotically Gaussian efficient. We compare the asymptotic variance the estimator $\hat{\theta}_n$ based on the disparity (4.5) in the following. In addition an analytic lower bound for the estimator $\hat{\theta}_n$ is found in the following theorem.

Theorem 5.2 *Suppose Assumptions 1, 2, 3 and 5 hold. We obtain the following inequality in the matrix sense*

$$H(\theta_0)^{-1}\tilde{V}(\theta_0)H(\theta_0)^{-1} \geq \mathcal{F}(\theta_0)^{-1}. \tag{5.5}$$

The equality holds if $p = -1$ *or the spectral density* $f(\omega)$ *is a constant function.*

The inequality (5.5) in Theorem 5.2 can be shown by the Cauchy-Bunyakovsky inequality. The equality holds if and only if there exists a constant C such that

$$\left(\int_{-\pi}^{\pi} f_\theta(\lambda)^{p+1} d\lambda\right) f_\theta(\omega)^{p+1}\partial f_\theta(\omega) - \left(\int_{-\pi}^{\pi} f_\theta(\lambda)^{p}\partial f_\theta(\lambda) d\lambda\right) f_\theta(\omega)^{p+2} - C\partial f_\theta(\omega)\bigg|_{\theta=\theta_0} = 0. \tag{5.6}$$

If $p = -1$, then the left hand side of (5.6) is

$$\begin{aligned}&\left(\int_{-\pi}^{\pi} f_\theta(\lambda)^{p+1} d\lambda\right) f_\theta(\omega)^{p+1}\partial f_\theta(\omega) - \left(\int_{-\pi}^{\pi} f_\theta(\lambda)^{p}\partial f_\theta(\lambda) d\lambda\right) f_\theta(\omega)^{p+2} - C\partial f_\theta(\omega)\bigg|_{\theta=\theta_0} \\ &= 2\pi\partial f_\theta(\omega) - \left(\int_{-\pi}^{\pi} f_\theta(\lambda)^{-1}\partial f_\theta(\lambda) d\lambda\right) f_\theta(\omega) - C\partial f_\theta(\omega)\bigg|_{\theta=\theta_0}.\end{aligned}$$

From equation (4.4), we see that if we take $C = 2\pi$, then the equation (5.6) holds. When $p \neq -1$, then (5.6) does not hold in general. Still we can see that (5.6) holds if the spectral density is a constant function independent of ω.

In the following, let us give close this section with two examples for Theorem 5.2. Especially, we compare the asymptotic variance of the estimator $\hat{\theta}_n$ based on the disparity (4.5) when $p = -2$ with that when $p = -1$.

Example 1 Let $\{X(t)\}$ be generated by the AR(1) model as follows.

$$X(t) = \theta X(t-1) + \epsilon(t), \quad |\theta| < 1, \quad \epsilon(t) \sim \text{i.i.d.}\,\mathcal{N}(0, \sigma^2).$$

The spectral density $f_\theta(\omega)$ of $\{X(t)\}$ is expressed as

$$f_\theta(\omega) = \frac{\sigma^2}{2\pi}\frac{1}{|1-\theta e^{i\omega}|^2}.$$

From Theorem 4.3, let $p = -2$ and we obtain

$$H(\theta) = 2\cdot(2\pi)^4(1-\theta^2), \quad V(\theta) = 4\cdot(2\pi)^8(1-\theta^2)^2.$$

Thus, the asymptotic variance when $p = -2$ is

$$H(\theta)^{-1}V(\theta)H(\theta)^{-1} = 1. \tag{5.7}$$

On the other hand, by (4.6), it holds that

$$\mathcal{F}(\theta) = \frac{1}{1-\theta^2}. \tag{5.8}$$

Then comparing the equation (5.7) with (5.8), we have

$$1 = H(\theta)^{-1}V(\theta)H(\theta)^{-1} \geq \mathcal{F}(\theta)^{-1} = 1 - \theta^2. \tag{5.9}$$

From (5.9) we can see that $\hat{\theta}_n$ is not asymptotically efficient except for $\theta = 0$.

Example 2 Let $\{X(t)\}$ be generated by the MA(1) model as follows.

$$X(t) = \epsilon(t) + \theta\,\epsilon(t-1), \quad |\theta| < 1, \quad \epsilon_t \sim \text{i.i.d.}\,\mathcal{N}(0, \sigma^2).$$

The spectral density $f_\theta(\omega)$ of $\{X_t\}$ is

$$f_\theta(\omega) = \frac{\sigma^2}{2\pi}|1 + \theta e^{i\omega}|^2.$$

From Theorem 4.3, let $p = -2$ and we obtain

$$H(\theta)^{-1}V(\theta)H(\theta)^{-1} = 1 - \theta^4 = (1 - \theta^2)(1 + \theta^2). \tag{5.10}$$

On the other hand, by (4.6), it holds that

$$\mathcal{F}(\theta) = \frac{1}{1 - \theta^2}. \tag{5.11}$$

Then comparing the equation (5.10) with (5.11), we have

$$(1 - \theta^2)(1 + \theta^2) = H(\theta)^{-1}V(\theta)H(\theta)^{-1} \geq \mathcal{F}(\theta)^{-1} = 1 - \theta^2.$$

Therefore $\hat{\theta}_n$ is not asymptotically efficient except for $\theta = 0$.

References

Ash, R. and Gardner, M. (1975). *Topics in Stochastic Processes.* Academic Press, New York.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85** 549–559.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1987). *Regular Variation.* Cambridge University Press.

Bloomfield, P. (1973). An exponential model for the spectrum of a scalar time series. *Biometrika* **60** 217–226.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods.* Springer-Verlag.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability* **3** 146–158.

Feller, W. (1968). *An introduction to probability theory and its applications.* John Wiley & Sons.

Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis* **99** 2053–2081.

GRENANDER, U. and ROSENBLATT, M. (1954). An extension of a theorem of G. Szegö and its application to the study of stochastic processes. *Transactions of the American Mathematical Society* **76** 112–126.

HANNAN, E. (1970). *Multiple Time Series.* John Wiley & Sons.

HOSOYA, Y. (1982). Harmonizable stable processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **60** 517–533.

HOSOYA, Y. and TANIGUCHI, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *The Annals of Statistics* **10** 132–153.

KOLMOGOROV, A. N. (1941a). Interpolation and extrapolation of stationary random sequences. *Bulletin de l'academie des sciences de U.R.S.S., Ser. Math* **5** 3–14.

KOLMOGOROV, A. N. (1941b). Stationary sequences in Hilbert space. *Bull. Math. Univ. Moscow* **2** 1–40.

LIU, Y. (2017a). Robust parameter estimation for stationary processes by an exotic disparity from prediction problem. *Statistics & Probability Letters* **129** 120–130.

LIU, Y. (2017b). Statistical inference for quantiles in the frequency domain. *Statistical Inference for Stochastic Processes* **20** 369–386.

MIAMEE, A. and POURAHMADI, M. (1988). Best Approximations in $\mathrm{L}^p(\mathrm{d}\mu)$ and Prediction Problems of Szegö, Kolmogorov, Yaglom, and Nakazi. *Journal of the London Mathematical Society* **2** 133–145.

RÉNYI, A. (1961). *On measures of entropy and information.* In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 547–561.

ROZANOV, Y. (1967). *Stationary random processes.* Holden Day, San Francisco.

SALEHI, H. (1979). Algorithms for linear interpolator and interpolation error for minimal stationary stochastic processes. *The Annals of Probability* 840–846.

SAMORODNITSKY, G. and TAQQU, M. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance.* CRC Press.

SCHILDER, M. (1970). Some structure theorems for the symmetric stable laws. *The Annals of Mathematical Statistics* **41** 412–421.

SUTO, Y., LIU, Y., and TANIGUCHI, M. (2016). Asymptotic theory of parameter estimation by a contrast function based on interpolation error. *Statistical Inference for Stochastic Processes* **19** 93–110.

SZEGÖ, G. (1915). Ein Grenzwertsatz über die Toeplitzschen Determinanten einer reellen positiven Funktion. *Math. Ann.* **76** 490–503.

TANIGUCHI, M. and KAKIZAWA, Y. (2000). *Asymptotic theory of statistical inference for time series.* New York: Springer-Verlag.

TANIGUCHI, M. (1981). An estimation procedure of parameters of a certain spectral density model. *Journal of the Royal Statistical Society. Series B* **43** 34–40.

TANIGUCHI, M. (1987). Minimum contrast estimation for spectral densities of stationary processes. *Journal of the Royal Statistical Society. Series B* **49** 315–325.

URBANIK, K. (1967). *Some prediction problems for strictly stationary processes.* In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.

WERON, A. (1985). Harmonizable stable processes on groups: spectral, ergodic and interpolation properties. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **68** 473–491.

WHITTLE, P. (1952). Some results in time series analysis. *Scandinavian Actuarial Journal* 48–60.

YAGLOM, A. M. (1962). *An introduction to the theory of stationary random functions.* Prentice-Hall.

YAGLOM, A. (1963). On a problem of linear interpolation of stationary random sequences and processes. *Amer. Math. Soc. Selected Translations Math. Statist* **4** 330–344.