Exact Maximum Likelihood Estimation for Non-Gaussian Non-invertible Moving Averages

Nan-Jung Hsu
Institute of Statistics
National Tsing-Hua University
F. Jay Breidt
Department of Statistics
Colorado State University

January 31, 2005

Abstract

A procedure for solving exact maximum likelihood estimation (MLE) is proposed for non-invertible non-Gaussian MA processes. By augmenting certain latent variables, the exact likelihood of all relevant innovations can be expressed explicitly according to a set of recursions (Breidt and Hsu, 2005). Then, the exact MLE is solved numerically by EM algorithm. Two alternative estimators are proposed subject to different treatments to the latent variables. We show that these three estimators and the approximate MLE proposed by Lii and Rosenblatt (1992) are asymptotically equivalent. In simulations, the MLE solved by EM performs better than other estimators in terms of smaller root mean square errors for small samples.

Keywords: EM algorithm, Monte Carlo, non-invertible, non-minimum phase, non-Gaussian.

1 Introduction

Consider a qth order moving average process (MA(q))

$$X_t = \theta(B) Z_t, \tag{1}$$

where $\{Z_t\}$ is an independent and identically distributed (iid) sequence of random variables with zero mean and finite variance, B is the backshift operator ($B^k Y_t = Y_{t-k}$ for $k = 0, \pm 1, \pm 2, \ldots$),

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q,$$

and $\theta_q \neq 0$. The moving average polynomial, $\theta(z)$, is said to be invertible if all the roots of $\theta(z) = 0$ are outside the unit circle in the complex plane, and non-invertible (or non-minimum phase) otherwise (Brockwell and Davis, 1991, Theorem 3.1.2).

Invertibility is a standard assumption in the analysis of moving average time series models, because without this assumption the model (1) is not identifiable using estimation methods based on second-order moments of the process. Such methods include Gaussian likelihood, least-squares, and various spectral-based methods (see, for example, Brockwell and Davis, 1991). But in the non-Gaussian case, invertible and non-invertible moving averages are distinguishable on the basis of higher-order cumulants or likelihood functions. The invertibility assumption in the non-Gaussian case is entirely artificial, and removing this assumption leads to a broad class of useful models.

Indeed, non-invertible moving averages and the broader class of non-minimum phase autoregressive moving average (ARMA) models are important tools in a number of applications, including seismic and other deconvolution problems (Wiggins, 1978; Ooe and Ulrych, 1979; Blass and Halsey, 1981; Donoho, 1981; Godfrey and Rocca, 1981; Hsueh and Mendel, 1985; Scargle, 1981), design of communication systems (Benveniste, Goursat, and Roget, 1980), processing of blurry images (Donoho, 1981; Chien, Yang, and Chi, 1997), and modeling of vocal tract filters (Rabiner and Schafer, 1978; Chien, Yang, and Chi, 1997).

Estimation methods for general moving average processes include cumulant-based estimators using cumulants of order greater than two (Wiggins, 1978; Donoho, 1981; Lii and Rosenblatt, 1982; Giannakis and Swami, 1990; Chi and Kung, 1995; Chien, Yang, and Chi, 1997); quasi-likelihood methods which lead to least absolute deviation-type estimators (Huang and Pawitan, 2000; Breidt, Davis, and Trindade, 2001); and the approximate maximum likelihood estimation (Lii and Rosenblatt, 1992). The approach by Huang and Pawitan (2000) is a semi-parametric method which uses the Laplacian likelihood and does not need the assumption for the distribution of innovations. However, their likelihood is actually a conditional likelihood which is conditional on some initial variables setting to be zero. This estimator is proved to be consistent only for heavy tailed errors. The alternative approach by Lii and Rosenblatt (1992) uses an appropriate truncation in the representation of the innovations in terms of the observations and approximates the likelihood function based on these truncated innovations. This approximate MLE is shown to be asymptotically equivalent to the exact MLE under mild conditions.

In this work, we incorporate q latent variables into the likelihood function and solve the exact MLE of the model parameters by EM algorithm. In addition, we propose two alternative estimators subject to different treatments to the latent variables. One is the conditional MLE in which all of the latent variables are set to be zero. The other is the joint MLE in which the latent variables and the model parameters are estimated simultaneously by maximizing the joint likelihood. We show that these two alternative estimators have the same asymptotic distribution as the MLE.

The rest of the paper is organized as follows. In Section 2, we first review the recursions given by Breidt and Hsu (2005) for computing residuals from a realization of a general moving average process. Then, the likelihood function is written in terms of these residuals and the specified distribution of the innovations. In Section 3, the procedures for solving exact MLE by EM algorithm and two alternative estimators are introduced. In Section 4, numerical simulations are conducted for evaluating the performance of different estimators in finite samples for a non-invertible MA(2) with non-Gaussian noise under various parameter settings. As a result, the exact MLE performs better than other estimators in terms of smaller root mean square errors for all cases. A brief discussion follows in Section 5. Proofs are given in the Appendix.

2 MA processes and the likelihoods

Rewrite (1) as

$$X_t = \theta(B)Z_t = \theta^{\dagger}(B)\theta^*(B)Z_t, \tag{2}$$

with

$$\theta^*(z) = 1 + \theta_1^* z + \cdots + \theta_s^* z^s \neq 0 \text{ for } |z| > 1,$$

and

$$\theta^{\dagger}(z) = 1 + \theta_1^{\dagger} z + \cdots + \theta_r^{\dagger} z^r \neq 0 \text{ for } |z| < 1,$$

where $r + s = q$, $\theta_s^* \neq 0$, and $\theta_r^{\dagger} \neq 0$. We further assume that any unit roots of $\theta^*(z) = 0$ are not repeated roots. The moving average polynomial, $\theta(z)$, is said to be invertible if $s = 0$ and non-invertible (or non-minimum phase) if $s \neq 0$. We will refer to the case in which $r = 0$ and $s > 0$ as purely non-invertible.

According to Breidt and Hsu (2005), define

$$W_t = \theta^{\dagger}(B)Z_t, \tag{3}$$

so that

$$Z_t = W_t - (\theta^{\dagger}(B) - 1)Z_t. \tag{4}$$

Then,

$$\begin{aligned} X_t &= \theta^*(B)W_t \\ &= (1 + \theta_1^* B + \cdots + \theta_s^* B^s)W_t \\ &= \theta_s^* \tilde{\theta}(B^{-1})W_{t-s}, \end{aligned} \tag{5}$$

where

$$\tilde{\theta}(z) = 1 + \frac{\theta_{s-1}^*}{\theta_s^*} z + \cdots + \frac{\theta_1^*}{\theta_s^*} z^{s-1} + \frac{1}{\theta_s^*} z^s.$$

Thus,

$$W_{t-s} = \frac{X_t}{\theta_s^*} - \left(\tilde{\theta}(B^{-1}) - 1\right) W_{t-s}. \tag{6}$$

By incorporating the latent variables $\boldsymbol{Z}_r = (Z_{-q+1}, \ldots, Z_{-q+r})'$ and $\boldsymbol{W}_s = (W_{n-s+1}, \ldots, W_n)'$ to the observations $\boldsymbol{X}_n = (X_1, \ldots, X_n)'$, the joint distribution of $(\boldsymbol{X}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s)$ can be represented by the joint distribution of $(Z_{-q+1}, \ldots, Z_{-1}, Z_0, Z_1, \ldots, Z_n)$ via the following two linear transformations:

$$\begin{pmatrix} Z_{-q+1} \\ \vdots \\ Z_{-q+r} \\ W_{-s+1} \\ \vdots \\ W_0 \\ W_1 \\ \vdots \\ W_n \end{pmatrix} = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ \theta_r^\dagger & \cdots & \theta_1^\dagger & 1 & & & & & \\ & \ddots & \ddots & \ddots & \ddots & & & & \\ & & \ddots & \ddots & \ddots & \ddots & & & \\ & & & \ddots & \ddots & \ddots & \ddots & & \\ & & & & \ddots & \ddots & \ddots & \ddots & \\ & & & & & \theta_r^\dagger & \cdots & \theta_1^\dagger & 1 \end{pmatrix} \begin{pmatrix} Z_{-q+1} \\ \vdots \\ Z_{-q+r} \\ Z_{-s+1} \\ \vdots \\ Z_0 \\ Z_1 \\ \vdots \\ Z_n \end{pmatrix},$$

and

$$\begin{pmatrix} Z_{-q+1} \\ \vdots \\ Z_{-q+r} \\ X_1 \\ \vdots \\ X_n \\ W_{n-s+1} \\ \vdots \\ W_n \end{pmatrix} = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & \theta_s^* & \cdots & \theta_1^* & 1 & \\ & & & & \ddots & \ddots & \ddots & \ddots \\ & & & & & \theta_s^* & \cdots & \theta_1^* \quad 1 \\ & & & & & & 1 & \\ & & & & & & & \ddots \\ & & & & & & & 1 \end{pmatrix} \begin{pmatrix} Z_{-q+1} \\ \vdots \\ Z_{-q+r} \\ W_{-s+1} \\ \vdots \\ W_{n-s} \\ W_{n-s+1} \\ \vdots \\ W_n \end{pmatrix},$$

which can be obtained directly from (4) and (6). Consequently, the joint distribution of $(\boldsymbol{X}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s)$ satisfies

$$p(\boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s; \boldsymbol{\theta}, \sigma) = |\theta_s^*|^{-n} \prod_{t=-q+1}^{n} f_\sigma(z_t), \tag{7}$$

where $f_\sigma(z) = \sigma^{-1} f(z/\sigma)$ is the probability density function of Z_t with scale parameter σ and $\boldsymbol{\theta} = (\boldsymbol{\theta}^*, \boldsymbol{\theta}^\dagger)$ with $\boldsymbol{\theta}^* \equiv (\theta_1^*, \theta_2^*, \ldots, \theta_s^*)'$ and $\boldsymbol{\theta}^\dagger \equiv (\theta_1^\dagger, \theta_2^\dagger, \ldots, \theta_r^\dagger)'$. Also, note that $\{z_t : t = -q+1, \ldots, n\}$ solved by (4) and (6) are functions of $\boldsymbol{\theta}$, $\boldsymbol{X}_n$, $\boldsymbol{Z}_r$ and $\boldsymbol{W}_s$, denoted as $z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s)$ for completeness in the following context.

Since the latent variables $\boldsymbol{Z}_r$ and $\boldsymbol{W}_s$ are unobserved in practice, we proposed three estimators of parameters by maximizing the joint distribution (7) subject to different treatments to the latent variables. The first estimator is the exact MLE solved by EM algorithm in which the latent variables are treated as missing data. The second estimator is the conditional MLE in which all of the latent variables are set to be zero. The third estimator is the joint MLE in which the latent variables and the model parameters are estimated simultaneously. The details are described in the next section.

3 Estimation methods

In the following, three estimators of $\boldsymbol{\psi} = (\boldsymbol{\theta}, \sigma)'$ are introduced.

First, the exact MLE is considered. The latent variables $(\boldsymbol{Z}_r, \boldsymbol{W}_s)$ are unobserved and treated as missing data in the joint distribution (7). Then, the MLE of parameters is solved recursively by the EM algorithm given in the following two steps:

- E-step: Compute the conditional expectation of the log-likelihood which satisfies

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}_{old}) = -n \log |\theta_s^*| + \sum_{t=-q+1}^{n} E_{old}\left[\log f_\sigma\left(z_t(\boldsymbol{\theta}, \boldsymbol{X}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s)\right) | \boldsymbol{X}_n\right],$$

  where the expectation E_{old} is taken on the latent variables $(\boldsymbol{Z}_r, \boldsymbol{W}_s)$ with respect to the conditional distribution given $\boldsymbol{X}_n$ at current parameter estimates $\boldsymbol{\psi}_{old}$.

- M-step: Solve

$$\boldsymbol{\psi}_{new} = \text{argmax}_{\boldsymbol{\psi}}\, Q(\boldsymbol{\psi}|\boldsymbol{\psi}_{old}).$$

These two steps are repeated until the convergence achieved. In general, the expectation in the E-step does not have close form. Therefore, we adopt a Monte Carlo method to compute the expectation numerically. The same idea is used in Breidt and Hsu (2005) for obtaining the best mean square prediction for non-Gaussian MA processes. More precisely, the expectation in the E-step is re-written as

$$\begin{aligned} Q_t(\boldsymbol{\psi}|\boldsymbol{\psi}_{old}) &\equiv E_{old}\left[\log f_\sigma\left(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s)\right) | \boldsymbol{X}_n = \boldsymbol{x}_n\right] \\ &= \frac{\int \log f_\sigma\left(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s)\right) p(\boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s; \boldsymbol{\psi}_{old}) d\boldsymbol{z}_r d\boldsymbol{w}_s}{\int p(\boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s; \boldsymbol{\psi}_{old}) d\boldsymbol{z}_r d\boldsymbol{w}_s} \\ &= \frac{\int \log f_\sigma\left(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s)\right) \left[\prod_{k=-q+1}^{n} f_{\sigma_{old}}(z_k(\boldsymbol{\theta}_{old}, \boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s))\right] d\boldsymbol{z}_r d\boldsymbol{w}_s}{\int \prod_{k=-q+1}^{n} f_{\sigma_{old}}(z_k(\boldsymbol{\theta}_{old}, \boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s)) d\boldsymbol{z}_r d\boldsymbol{w}_s}. \end{aligned} \quad (8)$$

These q-dimensional integrals can be evaluated numerically via importance sampling as follows. Let $h(\boldsymbol{z}_r, \boldsymbol{w}_s)$ be a q-dimensional joint density (called importance sampler) satisfying

$$\text{supp}(h) \supset \text{supp}\left(\int p(\boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s; \boldsymbol{\psi}_{old})\, d\boldsymbol{x}_n\right),$$

where supp(h) is the support of h. For any $\left(\boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)}\right) \in \text{supp}(h)$, define the importance weight

$$A\left(\boldsymbol{x}_n, \boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)}; \boldsymbol{\psi}_{old}\right) = \frac{\prod_{k=-q+1}^{n} f_{\sigma_{old}}\left(z_k(\boldsymbol{\theta}_{old}, \boldsymbol{x}_n, \boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)})\right)}{h\left(\boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)}\right)}.$$

Then, (8) can be written as

$$\begin{aligned} Q_t(\boldsymbol{\psi}|\boldsymbol{\psi}_{old}) &= \frac{\int \log f_\sigma\left(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s)\right) A\left(\boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s; \boldsymbol{\psi}_{old}\right) h\left(\boldsymbol{z}_r, \boldsymbol{w}_s\right) d\boldsymbol{z}_r d\boldsymbol{w}_s}{\int A\left(\boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s; \boldsymbol{\psi}_{old}\right) h\left(\boldsymbol{z}_r, \boldsymbol{w}_s\right) d\boldsymbol{z}_r d\boldsymbol{w}_s} \\ &= \frac{E_h\left[\log f_\sigma\left(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s)\right) A\left(\boldsymbol{x}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s; \boldsymbol{\psi}_{old}\right)\right]}{E_h\left[A\left(\boldsymbol{x}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s; \boldsymbol{\psi}_{old}\right)\right]}, \end{aligned} \quad (9)$$

where the expectation in (9) is taken with respect to $h(\boldsymbol{z}_r, \boldsymbol{w}_s)$. The ratio in (9) can be approximated via Monte Carlo (MC) method as

$$\hat{Q}_t(\boldsymbol{\psi}|\boldsymbol{\psi}_{old}) \quad = \quad \frac{\sum_{i=1}^{M} \log f_\sigma \left(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)}) \right) A \left(\boldsymbol{x}_n, \boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)}; \boldsymbol{\psi}_{old} \right)}{\sum_{i=1}^{M} A \left(\boldsymbol{x}_n, \boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)}; \boldsymbol{\psi}_{old} \right)}, \tag{10}$$

where M is the number of draws in the importance sampling and $\{(\boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)}); i = 1, 2, \ldots, M\}$ are q-dimensional random vectors drawn from the importance sampler $h(\boldsymbol{z}_r, \boldsymbol{w}_s)$. Consequently, the MC approximation of $Q\left(\boldsymbol{\psi}|\boldsymbol{\psi}_{old}\right)$ is

$$\hat{Q}\left(\boldsymbol{\psi}|\boldsymbol{\psi}_{old}\right) = -n \log |\theta_s^*| + \sum_{t=-q+1}^{n} \hat{Q}_t(\boldsymbol{\psi}|\boldsymbol{\psi}_{old}),$$

which is actually used in the implementation of the EM algorithm to solve the MLE.

The second estimator is the conditional MLE in which the latent variables are set to be zero. Namely,

$$\hat{\boldsymbol{\psi}}_c = \operatorname{argmax}_{\boldsymbol{\psi}} \left\{ -n \log |\theta_s^*| + \sum_{t=-q+1}^{n} \log f_\sigma \left(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \mathbf{0}_r, \mathbf{0}_s) \right) \right\}. \tag{11}$$

The third estimator is the joint MLE in which the latent variables and the model parameters are estimated simultaneously, which is

$$(\hat{\boldsymbol{\psi}}_J, \hat{\boldsymbol{z}}_r, \hat{\boldsymbol{w}}_s) = \operatorname{argmax}_{(\boldsymbol{\psi}, \boldsymbol{z}_r, \boldsymbol{w}_s)} \left\{ -n \log |\theta_s^*| + \sum_{t=-q+1}^{n} \log f_\sigma \left(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s) \right) \right\}. \tag{12}$$

In Lii and Rosenblatt (1992), they derived the asymptotic distribution for the "quasi-estimator" of parameters which maximizing the "pseudo" likelihood defined in (1.18) in their paper. This "quasi-estimator" is the MLE if all $\{X_t : t \in \mathbb{Z}\}$ are observed. However, the "pseudo" likelihood depends on unobserved $\{X_t\}$ and cannot be computed in practice. Therefore, they further show that the approximate MLE of parameters, which maximizes a truncated version of the "pseudo" likelihood, is asymptotically equivalent to the "quasi-estimator". In this work, we further show that the conditional MLE $\hat{\boldsymbol{\psi}}_c$ and the joint MLE $\hat{\boldsymbol{\psi}}_J$ defined in (11) and (12) are also equivalent to the "quasi-estimator" asymptotically. The proof is given in the Appendix.

4 Simulation

To investigate the performance of the proposed estimators for finite samples, the following simulation study is conducted. We consider two kinds of innovation distributions; one is Laplace distribution and the other is Student's t distribution with four degrees of freedom, which are the same as the setting given in Lii and Rosenblatt (1992). The models considered in our simulation study include an invertible MA(1) ($s = 0, r = 1$), a non-invertible MA(1) ($s = 1, r = 0$) and a non-purely non-invertible MA(2) ($s = r = 1$). For each MA process,

Table 1: Model settings for simulations.

Models	Parameter Values	Innovation Distributions
invertible MA(1)	$\theta_1^\dagger = 0.9$ or $\theta_1^\dagger = 0.8$	Laplace
non-invertible MA(1)	$\theta_1^* = 1/0.9$ or $\theta_1^* = 1/0.8$	Laplace
non-invertible MA(2)	$\theta_1^* = -1/0.9$ and $\theta_1^\dagger = 1/1.1$	Laplace or $t(4)$
	$\theta_1^* = -1/0.8$ and $\theta_1^\dagger = 1/1.2$	Laplace or $t(4)$

two parameters settings are considered in which one is closer to the unit root case and the other is not. The detail settings are given in Table 1. For each process, 200 samples with size $n = 50$ and $n = 100$ are generated and four estimators, including the MLE by EM, conditional MLE, joint MLE and the approximate MLE proposed by Lii and Rosenbaltt (abbreviated by L&R MLE), are computed for each sample. Similar to Lii and Rosenblatt (1992) in their simulations, we use $q_{trunc} = 10$ as the number of truncation for computing the approximate MLE for both sample sizes, which leads to the effective data length $(n-2q_{trunc})$ being 30 for the case with $n = 50$ and being 80 for the case with $n = 100$.

The performance of four estimators under various settings are summarized in Tables **??**–5 in terms of biases, standard errors and root mean square errors. The standard error based on the asymptotic distribution (ASD) for each process is also given at the bottom of each table for reference. Note that, in Table 4 and 5, the estimation results contain two parts (a) and (b) in which the parameters are estimated under the true distribution ($t(4)$) with unknown σ in part (a) and the parameters are estimated under the Laplace distribution with unknown σ in part (b) (which leads to the least absolute deviation estimation).

Add results for MA(1) here

Generally speaking, for MA(2) processes, the MLE solved by EM outperforms the other estimators in terms of smaller root mean square errors when $n = 50$, especially for the scale parameter σ. Moreover, the second best estimator is the joint MLE for most of the cases with $n = 50$. For the rest of two estimators, the L&R MLE is better on estimating the MA coefficients and the conditional MLE is better on estimating the scale parameter. Note that, compared to other estimators, the L&R MLE produces very poor estimation for the scale parameter σ for small samples which is probably due to small effective data length subject to the truncation in the objective function. This situation is improved when the sample size increases. Roughly speaking, these four estimators are fairly competitive when $n = 100$ which confirms their asymptotic equivalence. Between different parameter settings, the efficiency gain of the MLE solved by EM relative to the L&R MLE is larger when the roots of MA polynomials are closer to the unit circle.

To sum up, the MLE solved by EM has best performance among four considered estimators for small samples. But, from the computational point of view, the joint MLE is a good alternative for small samples, especially when the roots of MA polynomials are closer to the unit circle.

Table 2: Biases, standard errors and root mean square errors of various likelihood-based estimators for a non-invertible MA(2) satisfying $X_t = (1 - 0.9^{-1}B)(1 + 1.1^{-1}B)Z_t$ where $\{Z_t\}$ are iid Laplace with $\sigma = 1$. (200 replicates)

		$n = 50$			$n = 100$		
Estimator		$r_1 = 0.9$	$r_2 = -1.1$	$\sigma = 1$	$r_1 = 0.9$	$r_2 = -1.1$	$\sigma = 1$
L&R MLE	bias	-0.0228	-0.0477	-0.0747	-0.0051	-0.0139	-0.0280
	s.d.	0.0990	0.1329	0.1948	0.0456	0.0612	0.1176
	rmse	0.1014	0.1412	0.2087	0.0459	0.0628	0.1209
MLE by EM	bias	0.0022	-0.0125	-0.0464	0.0029	-0.0008	-0.0224
	s.d.	0.0765	0.1017	0.1410	0.0494	0.0559	0.1122
	rmse	0.0765	0.1025	0.1485	0.0495	0.0560	0.1144
Joint MLE	bias	-0.0116	-0.0155	-0.0742	-0.0022	0.0018	-0.0357
	s.d.	0.1003	0.1189	0.1604	0.0575	0.0631	0.1168
	rmse	0.1010	0.1199	0.1767	0.0575	0.0631	0.1219
Cond. MLE	bias	-0.0374	-0.0790	-0.0616	-0.0224	-0.0434	-0.0295
	s.d.	0.1015	0.1501	0.1674	0.0624	0.0951	0.1217
	rmse	0.1082	0.1696	0.1784	0.0661	0.1043	0.1249
ASD		0.0566	0.0653	0.2731	0.0346	0.0401	0.1672

Table 3: Biases, standard errors and root mean square errors of various likelihood-based estimators for a non-invertible MA(2) satisfying $X_t = (1 - 0.8^{-1}B)(1 + 1.2^{-1}B)Z_t$ where $\{Z_t\}$ are iid Laplace with $\sigma = 1$. (200 replicates)

		$n = 50$			$n = 100$		
Estimator		$r_1 = 0.8$	$r_2 = -1.2$	$\sigma = 1$	$r_1 = 0.8$	$r_2 = -1.2$	$\sigma = 1$
L&R MLE	bias	-0.0054	-0.0210	-0.0481	0.0065	-0.0038	0.0008
	s.d.	0.1134	0.1582	0.2146	0.0608	0.0806	0.1308
	rmse	0.1136	0.1596	0.2199	0.0611	0.0807	0.1308
MLE by EM	bias	0.0185	0.0054	-0.0135	0.0088	-0.0010	-0.0032
	s.d.	0.0937	0.1554	0.1613	0.0574	0.0756	0.1161
	rmse	0.0955	0.1555	0.1619	0.0581	0.0757	0.1161
Joint MLE	bias	-0.0085	-0.0016	-0.0566	0.0023	-0.0007	-0.0201
	s.d.	0.1083	0.1520	0.1863	0.0637	0.0855	0.1214
	rmse	0.1086	0.1520	0.1947	0.0637	0.0855	0.1231
Cond. MLE	bias	-0.0231	-0.0421	-0.0498	-0.0071	-0.0271	-0.0157
	s.d.	0.1133	0.1538	0.1932	0.0590	0.0915	0.1212
	rmse	0.1156	0.1595	0.1995	0.0594	0.0955	0.1222
ASD		0.0988	0.0874	0.2936	0.0605	0.0535	0.1798

Table 4: Biases, standard errors and root mean square errors of various likelihood-based estimators for a non-invertible MA(2) satisfying $X_t = (1 - 0.9^{-1}B)(1 + 1.1^{-1}B)Z_t$ where $\{Z_t\}$ are iid $t(4)$ with $\sigma = 1$. (200 replicates)

(a) Parameters are estimated under $f_\sigma \sim t(4)$

		$n = 50$			$n = 100$		
Estimator		$r_1 = 0.9$	$r_2 = -1.1$	$\sigma = 1$	$r_1 = 0.9$	$r_2 = -1.1$	$\sigma = 1$
L&R MLE	bias	-0.0164	-0.0367	-0.0615	-0.0017	-0.0098	-0.0242
	s.d.	0.0985	0.1537	0.1937	0.0504	0.0517	0.1111
	rmse	0.0998	0.1580	0.2032	0.0505	0.0527	0.1137
MLE by EM	bias	0.0102	-0.0072	-0.0402	0.0080	0.0056	-0.0202
	s.d.	0.0767	0.1114	0.1338	0.0524	0.0541	0.1040
	rmse	0.0774	0.1117	0.1397	0.0531	0.0544	0.1059
Joint MLE	bias	0.0188	0.0123	-0.0528	0.0140	0.0126	-0.0239
	s.d.	0.0928	0.1103	0.1432	0.0540	0.0554	0.1033
	rmse	0.0946	0.1110	0.1526	0.0558	0.0568	0.1060
Cond. MLE	bias	-0.0361	-0.0611	-0.0622	-0.0135	-0.0270	-0.0211
	s.d.	0.1030	0.1582	0.1536	0.0568	0.0692	0.1062
	rmse	0.1092	0.1696	0.1657	0.0584	0.0743	0.1083
ASD		0.0672	0.0778	0.2637	0.0412	0.0476	0.1615

(b) Parameters are estimated under $f_\sigma \sim Lap$

		$n = 50$			$n = 100$		
Estimator		$r_1 = 0.9$	$r_2 = -1.1$	$\sigma = 1$	$r_1 = 0.9$	$r_2 = -1.1$	$\sigma = 1$
L&R MLE	bias	-0.0225	-0.0419	-0.0694	-0.0059	-0.0201	-0.0296
	s.d.	0.0944	0.1250	0.2008	0.0559	0.0666	0.1199
	rmse	0.0972	0.1319	0.2125	0.0562	0.0696	0.1235
MLE by EM	bias	0.0025	-0.0235	-0.0527	0.0060	0.0030	-0.0238
	s.d.	0.0801	0.1534	0.1481	0.0544	0.0583	0.1094
	rmse	0.0801	0.1552	0.1572	0.0548	0.0584	0.1120
Joint MLE	bias	-0.0049	-0.0155	-0.0780	0.0050	0.0068	-0.0336
	s.d.	0.0987	0.1624	0.1514	0.0572	0.0672	0.1083
	rmse	0.0988	0.1631	0.1703	0.0574	0.0675	0.1134
Cond. MLE	bias	-0.0326	-0.0851	-0.0654	-0.0151	-0.0347	-0.0268
	s.d.	0.0991	0.2166	0.1611	0.0620	0.0779	0.1142
	rmse	0.1043	0.2327	0.1739	0.0638	0.0852	0.1173

Table 5: Biases, standard errors and root mean square errors of various likelihood-based estimators for a non-invertible MA(2) satisfying $X_t = (1 - 0.8^{-1}B)(1 + 1.2^{-1}B)Z_t$ where $\{Z_t\}$ are iid $t(4)$ with $\sigma = 1$. (200 replicates)

(a) Parameters are estimated under $f_\sigma \sim t(4)$

		$n = 50$			$n = 100$		
Estimator		$r_1 = 0.8$	$r_2 = -1.2$	$\sigma = 1$	$r_1 = 0.8$	$r_2 = -1.2$	$\sigma = 1$
L&R MLE	bias	0.0010	-0.0304	-0.0212	0.0004	-0.0049	-0.0151
	s.d.	0.1212	0.2286	0.2249	0.0608	0.0964	0.1270
	rmse	0.1212	0.2306	0.2259	0.0608	0.0965	0.1279
MLE by EM	bias	0.0245	0.0194	-0.0003	0.0064	0.0089	-0.0123
	s.d.	0.0987	0.1352	0.1806	0.0593	0.0819	0.1188
	rmse	0.1017	0.1366	0.1806	0.0596	0.0824	0.1194
Joint MLE	bias	0.0154	0.0199	-0.0315	0.0027	0.0095	-0.0256
	s.d.	0.1110	0.1441	0.1827	0.0644	0.0891	0.1210
	rmse	0.1121	0.1454	0.1854	0.0645	0.0896	0.1236
Cond. MLE	bias	-0.0215	-0.0694	-0.0398	-0.0112	-0.0318	-0.0268
	s.d.	0.1037	0.1788	0.1836	0.0649	0.1100	0.1255
	rmse	0.1059	0.1918	0.1878	0.0658	0.1145	0.1283
ASD		0.1193	0.1055	0.2946	0.0730	0.0646	0.1193

(b) Parameters are estimated under $f_\sigma \sim Lap$

		$n = 50$			$n = 100$		
Estimator		$r_1 = 0.8$	$r_2 = -1.2$	$\sigma = 1$	$r_1 = 0.8$	$r_2 = -1.2$	$\sigma = 1$
L&R MLE	bias	-0.0044	-0.0467	-0.0309	-0.0059	-0.0147	-0.0242
	s.d.	0.1171	0.1890	0.2365	0.0698	0.1173	0.1435
	rmse	0.1172	0.1947	0.2385	0.0700	0.1182	0.1455
MLE by EM	bias	0.0157	-0.0020	-0.0174	-0.0004	0.0012	-0.0243
	s.d.	0.1054	0.1467	0.1950	0.0636	0.0997	0.1310
	rmse	0.1066	0.1467	0.1958	0.0636	0.0997	0.1332
Joint MLE	bias	-0.0039	-0.0292	-0.0566	-0.0116	-0.0127	-0.0430
	s.d.	0.1212	0.2017	0.2022	0.0684	0.1124	0.1349
	rmse	0.1212	0.2038	0.2100	0.0693	0.1131	0.1416
Cond. MLE	bias	-0.0291	-0.0978	-0.0591	-0.0174	-0.0418	-0.0380
	s.d.	0.1189	0.2462	0.2053	0.0696	0.1272	0.1395
	rmse	0.1224	0.2649	0.2137	0.0717	0.1339	0.1446

5 Discussion

We proposed an algorithm to compute the exact MLE for invertible or non-invertible moving averages with non-Gaussian noise. Instead of truncating unobserved $\{X_t\}$ in the residuals for approximating the likelihood (Lii and Rosenblatt, 1992), we augment suitable latent variables to construct a joint likelihood. We, then solve the exact MLE by the EM algorithm. In addition, two alternative estimators are suggested subject to different treatments to the latent variables. Both of the alternative estimators are much easier to implement but equivalent to the MLE asymptotically. Simulation results show that the exact MLE solved by EM algorithm performs better than other estimators in terms of smaller root mean square errors in small samples for non-invertible non-Gaussian MA processes.

6 Appendix

In Lii and Rosenblatt (1992), they first derive the asymptotic distribution for the "quasi-estimator" of parameters which maximizing the "pseudo" likelihood defined in (1.18) in Lii and Rosenblatt (1992). However, the "pseudo" likelihood depends on unobserved $\{X_t\}$ and cannot be computed in practice, they further show that the approximate MLE of parameters, which maximizing the truncated version of "pseudo" likelihood defined in (1.19), is asymptotically equivalent to the "quasi-estimator".

In the following, we shall show that the conditional MLE and the joint MLE defined in (11) and (12) are also asymptotically equivalent to the "quasi-estimator". For notational simplicity, we only give the proof for a non-invertible MA(2) with $r = s = 1$ satisfying $\theta^{\dagger}(B) = 1 - \theta_1 B$ and $\theta^*(B) = 1 - \theta_2 B$ with $|\theta_1| < 1$, $|\theta_2| > 1$. The parameters considered in this particular case are $\boldsymbol{\psi} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma)'$.

First, some assumptions and notations are given. The assumptions for the density of the innovations $\{Z_t\}$ satisfying $f_\sigma(z) = \sigma^{-1} f(z/\sigma)$ are given below:

$A1 : f(x) > 0$ for all x.

$A2 : f \in C^2$.

$A3 : f' \in L^1$ with $\int f'(x)dx = 0$.

$A4 : \int xf'(x)dx = -1$.

$A5 : \int f''(x)dx = 0$.

$A6 : \int xf''(x)dx = 0$.

$A7 : \int x^2 f''(x)dx = 2$.

$A8 : \int (1 + x^2)[f'(x)]^2/f(x)dx < \infty$.

$A9 : |u(z + h) - u(z)| \leq A\left((1 + |z|^k)|h| + |h|^\ell\right)$ for all z, h with positive constants k, ℓ, A, where $u(\cdot) = (f'/f)\,, (f'/f)'$.

Before the proof, We first give two lemmas (without proofs).

Lemma 1 (Exercise 8.8 in Chapter 1, Durrett (2005)) *Assume $\{X_t; t = 1, 2, \ldots\}$ are iid non-zero random variables, then $\sum_{t=1}^n X_t z^t$ converges almost surely for $|z| < 1$ if and only if $E \log^+ |X_t| < \infty$, where $\log^+(x) = \max\{\log(x), 0\}$.*

Lemma 2 (Proposition 3.1.1 in Brockwell and Davis (1996)) *Assume $\{X_t\}$ are any sequence of random variables such that $\sup_t E|X_t| < \infty$. If $\sum_j |\psi_j| < \infty$, then $\sum_j \psi_j X_{t-j}$ converges almost surely.*

Proof: Define $W_t = \theta^\dagger(B)Z_t$ and $V_t = \theta^*(B)Z_t$. Consequently, $X_t = \theta^*(B)W_t$. Given the initials (w_n, z_{-1}), according to the backward and forward recursions in (6) and (4), we have

$$\begin{aligned} w_{n-j} &= -\sum_{\ell=1}^{j} \left(\frac{1}{\theta_2}\right)^{\ell} X_{n-j+\ell} + \left(\frac{1}{\theta_2}\right)^{j} w_n, \quad j = 1, 2, \ldots, n, \\ z_k &= \sum_{i=0}^{k} \theta_1^i w_{k-i} + \theta_1^{k+1} z_{-1} \\ &= -\sum_{i=0}^{k} \sum_{\ell=1}^{n-k+i} \theta_1^i \left(\frac{1}{\theta_2}\right)^{\ell} X_{k-i+\ell} + \left[\sum_{i=0}^{k} \theta_1^i \left(\frac{1}{\theta_2}\right)^{n-k+i}\right] w_n + \theta_1^{k+1} z_{-1} \\ &= -\sum_{i=0}^{k} \sum_{\ell=1}^{n-k+i} \theta_1^i \left(\frac{1}{\theta_2}\right)^{\ell} X_{k-i+\ell} + \left(\frac{1}{\theta_2}\right)^{n-k} \frac{1 - (\theta_1/\theta_2)^{k+1}}{1 - \theta_1/\theta_2} w_n + \theta_1^{k+1} z_{-1}, \end{aligned} \tag{13}$$

$k = 0, 1, \ldots, n$. Since the residuals $\{z_t\}$ are functions of $\boldsymbol{\theta}$, $\boldsymbol{X}_n$, w_n and z_{-1}, they are expressed as $z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, z_{-1}, w_n)$ for completeness and abbreviated as $z_t(\boldsymbol{\theta}) \equiv z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, z_{-1}, w_n)$ and $\hat{z}_t(\boldsymbol{\theta}) \equiv z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \hat{z}_{-1}, \hat{w}_n)$ for simplicity, where (z_{-1}, w_n) is the actual initial variables associated with $\boldsymbol{X}_n$ and $(\hat{z}_{-1}, \hat{w}_n)$ is some estimator of (z_{-1}, w_n). For Instance, $(\hat{z}_{-1}, \hat{w}_n) = \mathbf{0}$ for the conditional MLE case and $(\hat{z}_{-1}, \hat{w}_n)$ is a function of $\boldsymbol{X}_n$ for the joint MLE case.

Similar to Lii and Rosenblatt (1992), let

$$Q_\epsilon = \left\{\boldsymbol{\psi} \in \mathrm{I\!R}^{q+1} : |\boldsymbol{\psi} - \boldsymbol{\psi}_0| \le \epsilon\right\},$$

be a neighborhood of $\boldsymbol{\psi}_0 \equiv (\boldsymbol{\theta}_0, \sigma_0)'$, where $\boldsymbol{\psi}_0$ is the true parameter vector and $|\cdot|$ is the max norm on $\mathrm{I\!R}^{q+1}$. Then, for small $\epsilon > 0$, there exist a $d \in (0, 1)$ such that

$$\max_{\boldsymbol{\psi} \in Q_\epsilon} \left\{|\theta_1|, |\theta_2|^{-1}\right\} < d, \tag{14}$$

which is obtained directly from (3.2) in Lii and Rosenblatt (1992). Moreover, according to (3.3)' in Lii and Rosenblatt (1992), we have

$$\sup_{\boldsymbol{\psi} \in Q_\epsilon} |z_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}_0)| \le C\, \epsilon^{1/2} \sum_{j=-\infty}^{\infty} d^{|j|} |X_{t-j}|,$$

for some $C > 0$.

According to (13) and (14), we have

$$\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}) = \frac{1 - (\theta_1/\theta_2)^{t+1}}{1 - \theta_1/\theta_2} \left(\frac{1}{\theta_2}\right)^{n-t} (\hat{w}_n - w_n) + \theta_1^{t+1}(\hat{z}_{-1} - z_{-1}),$$

and

$$\sup_{\boldsymbol{\psi}\in Q_\epsilon} |\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta})| \le \frac{2}{1-d^2} d^{n-t}|\hat{w}_n - w_n| + d^{t+1}|\hat{z}_{-1} - z_{-1}|. \tag{15}$$

Also, the first derivatives of $(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}))$ satisfy

$$\frac{\partial}{\partial\theta_1}(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta})) = \left(\frac{1}{\theta_2}\right)^{n-t+1} \frac{\left[1-(t+1)\left(\frac{\theta_1}{\theta_2}\right)^t + t\left(\frac{\theta_1}{\theta_2}\right)^{t+1}\right]}{\left(1-\frac{\theta_1}{\theta_2}\right)^2}(\hat{w}_n - w_n) + (t+1)\theta_1^t(\hat{z}_{-1} - z_{-1}),$$

$$\begin{aligned}
&\frac{\partial}{\partial\theta_2}(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta})) \\
= &\left\{\theta_1\left(\frac{1}{\theta_2}\right)^{n-t+2} \frac{\left[-1+(t+1)\left(\frac{\theta_1}{\theta_2}\right)^t - t\left(\frac{\theta_1}{\theta_2}\right)^{t+1}\right]}{\left(1-\frac{\theta_1}{\theta_2}\right)^2} - (n-t)\left(\frac{1}{\theta_2}\right)^{n-t+1}\frac{1-\left(\frac{\theta_1}{\theta_2}\right)^{t+1}}{1-\theta_1/\theta_2}\right\}(\hat{w}_n - w_n),
\end{aligned}$$

and therefore,

$$\sup_{\boldsymbol{\psi}\in Q_\epsilon}\left|\frac{\partial}{\partial\boldsymbol{\theta}}(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}))\right| \le \frac{2(n-t+1)d^{n-t+1} + 2td^{n+t+1}}{(1-d^2)^2}|\hat{w}_n - w_n| + (t+1)d^t|\hat{z}_{-1} - z_{-1}|. \tag{16}$$

Similarly, the second derivative of $(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}))$ satisfies

$$\sup_{\boldsymbol{\psi}\in Q_\epsilon}\left|\frac{\partial^2}{\partial\theta_1\partial\theta_2}(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}))\right| \le \frac{4\,|\hat{w}_n - w_n|}{(1-d^2)^4}\left\{(n-t+2)d^{n-t+2} + [(2t+1)(n+2)+t]d^{n+t}\right\}. \tag{17}$$

In the following proof, we also need the derivatives of $z_t(\boldsymbol{\theta})$ which satisfy

$$\begin{aligned}
\frac{\partial z_t(\boldsymbol{\theta})}{\partial\theta_1} &= \frac{\partial}{\partial\theta_1}[\theta^\dagger(B)]^{-1}W_t = \frac{-1}{(\theta^\dagger(B))^2}(-B)W_t = \frac{1}{\theta^\dagger(B)}\left(\frac{1}{\theta^\dagger(B)}W_{t-1}\right) \\
&= \frac{1}{\theta^\dagger(B)}Z_{t-1} = \sum_{j=0}^\infty \theta_1^j Z_{t-1-j}, \\
\frac{\partial z_t(\boldsymbol{\theta})}{\partial\theta_2} &= \frac{\partial}{\partial\theta_1}[\theta^*(B)]^{-1}V_t = \frac{1}{(\theta^*(B))}Z_{t-1} = \left[-\theta_2 B\left(1-\frac{1}{\theta_2}B^{-1}\right)\right]^{-1} Z_{t-1} \\
&= -\frac{1}{\theta_2}B^{-1}\left(\sum_{j=0}^\infty\left(\frac{1}{\theta_2}\right)^j B^{-j}\right)Z_{t-1} = -\sum_{j=0}^\infty\left(\frac{1}{\theta_2}\right)^{j+1} Z_{t+j}, \\
\frac{\partial^2 z_t(\boldsymbol{\theta})}{\partial\theta_1\partial\theta_2} &= \frac{\partial}{\partial\theta_2}\frac{1}{\theta^\dagger(B)}Z_{t-1} = \frac{1}{\theta^\dagger(B)}\frac{\partial}{\partial\theta_2}\left(\frac{1}{(\theta^*(B))}V_{t-1}\right) \\
&= \frac{1}{\theta^\dagger(B)}\frac{1}{\theta^*(B)}Z_{t-2} = -\left(\sum_{j=0}^\infty \theta_1^j B^j\right)\left(\sum_{k=0}^\infty\left(\frac{1}{\theta_2}\right)^{k+1} B^{-k}\right)Z_{t-1},
\end{aligned}$$

in which all of the partial derivatives with respect to $\boldsymbol{\theta}$ have the following form

$$\sum_{j=-\infty}^{\infty} \gamma_j Z_{t-j}, \tag{18}$$

with some $\{\gamma_j\}$ where $|\gamma_j|$ decay exponentially.

In Lii and Rosenblatt (1992), the asymptotic distribution of the maximizer is derived based on the following "pseudo" likelihood function

$$\ell_n(\boldsymbol{\psi}) = -n \log|\theta_2| + \sum_{t=-q+1}^{n} \log f_\sigma(z_t(\boldsymbol{\theta})),$$

in which the residuals are functions of $(\boldsymbol{X}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s)$ and can be represented as infinite series of $\{X_t\}$. This maximizer is ideal but cannot be computed in practice since only finite number of $\{X_t\}$ are observable. They further show that the maximizer of a truncated version of $\ell_n(\boldsymbol{\psi})$ (denoted as $\ell_q(\boldsymbol{\psi})$, which is computable) is equivalent to the "quasi-MLE" asymptotically. In this paper, we consider the following objective function:

$$\hat{\ell}_n(\boldsymbol{\psi}) = -n \log|\theta_2| + \sum_{t=-q+1}^{n} \log f_\sigma(\hat{z}_t(\boldsymbol{\theta})),$$

where $\hat{z}_t$ depends on some initials $(\hat{z}_{-1}, \hat{w}_n)$. When $(\hat{z}_{-1}, \hat{w}_n) = \mathbf{0}$, the corresponding estimator is the conditional MLE defined in (11). One the other hand, the joint MLE defined in (12) is also a special case which maximizes $\hat{\ell}_n(\boldsymbol{\psi})$ where $(\hat{z}_{-1}, \hat{w}_n)$ are functions of $\boldsymbol{X}_n$. Our goal is to show that the conditional MLE, joint MLE and the quasi-MLE are asymptotically equivalent. Following the proof given by Lii and Rosenblatt (1992), we first show that

$$\sup_{\boldsymbol{\psi}\in Q_\epsilon} \frac{1}{n} |\ell_n(\boldsymbol{\psi}) - \hat{\ell}_n(\boldsymbol{\psi})| \to 0, \quad \text{a.s.}, \tag{19}$$

and then show that

$$n^{-1/2} \sum_{t=-1}^{n} \left[\frac{\partial}{\partial \boldsymbol{\psi}} \log f_\sigma(\hat{z}_t(\boldsymbol{\theta})) - \frac{\partial}{\partial \boldsymbol{\psi}} \log f_\sigma(z_t(\boldsymbol{\theta})) \right]_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} \to \mathbf{0}, \tag{20}$$

$$n^{-1} \left(\hat{\boldsymbol{B}}(\boldsymbol{\psi}^*) - \boldsymbol{B}(\boldsymbol{\psi}_0) \right) \to \mathbf{0}, \tag{21}$$

in probability, where $\boldsymbol{\psi} = (\boldsymbol{\theta}, \sigma)'$, $\boldsymbol{\psi}_0 = (\boldsymbol{\theta}_0, \sigma_0)'$,

$$\boldsymbol{B}(\boldsymbol{\psi}) \equiv \begin{pmatrix} \boldsymbol{B}_{\theta\theta}(\boldsymbol{\psi}) & \boldsymbol{B}_{\theta\sigma}(\boldsymbol{\psi}) \\ \boldsymbol{B}_{\sigma\theta}(\boldsymbol{\psi}) & \boldsymbol{B}_{\sigma\sigma}(\boldsymbol{\psi}) \end{pmatrix} \equiv \sum_{t=-1}^{n} \frac{\partial^2}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} \log f_\sigma(z_t(\boldsymbol{\theta})),$$

and $\hat{\boldsymbol{B}}(\boldsymbol{\psi})$ is defined similarly but subject to $\hat{z}_t(\boldsymbol{\theta})$.

For showing (19),

$$
\begin{aligned}
\frac{1}{n}|\ell_n(\boldsymbol{\psi}) - \hat{\ell}_n(\boldsymbol{\psi})| &= \frac{1}{n}\left|\sum_{t=-1}^{n} \log f_\sigma(z_t(\boldsymbol{\theta})) - \sum_{t=-1}^{n} \log f_\sigma(\hat{z}_t(\boldsymbol{\theta}))\right| \\
&\le \frac{1}{n}\sum_{t=-1}^{n} |\log f_\sigma(z_t(\boldsymbol{\theta})) - \log f_\sigma(\hat{z}_t(\boldsymbol{\theta}))| \\
&= \frac{1}{n}\sum_{t=-1}^{n} \left|(z_t(\boldsymbol{\theta}) - \hat{z}_t(\boldsymbol{\theta}))\frac{f'_\sigma}{f_\sigma}(\hat{z}_t^*(\boldsymbol{\theta}))\right| \\
&\le \frac{1}{n}\sum_{t=-1}^{n} |z_t(\boldsymbol{\theta}) - \hat{z}_t(\boldsymbol{\theta})| \left|\frac{f'_\sigma}{f_\sigma}(z_t(\boldsymbol{\theta}_0)) + \left(\frac{f'_\sigma}{f_\sigma}(\hat{z}_t^*(\boldsymbol{\theta})) - \frac{f'_\sigma}{f_\sigma}(z_t(\boldsymbol{\theta}_0))\right)\right| \\
&\le \frac{1}{n}\sum_{t=-1}^{n} |z_t(\boldsymbol{\theta}) - \hat{z}_t(\boldsymbol{\theta})| \left|\frac{f'_\sigma}{f_\sigma}(z_t(\boldsymbol{\theta}_0))\right| + \frac{1}{n}\sum_{t=-1}^{n} |z_t(\boldsymbol{\theta}) - \hat{z}_t(\boldsymbol{\theta})| \left|\frac{f'_\sigma}{f_\sigma}(\hat{z}_t^*(\boldsymbol{\theta})) - \frac{f'_\sigma}{f_\sigma}(z_t(\boldsymbol{\theta}_0))\right| \\
&\equiv \frac{1}{n}(A_1 + A_2),
\end{aligned}
$$

where the first-order Taylor expansion is used with $\hat{z}_t^*(\boldsymbol{\theta}) = z_t(\boldsymbol{\theta}) + u_t(\boldsymbol{\theta})(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}))$ for some $u_t(\boldsymbol{\theta}) \in [0, 1]$ (i.e., $\hat{z}_t^*(\boldsymbol{\theta})$ is between $z_t(\boldsymbol{\theta})$ and $\hat{z}_t(\boldsymbol{\theta})$). According to (15), we have

$$
\begin{aligned}
\sup_{\boldsymbol{\psi}\in Q_\epsilon} A_1 &= \sum_{t=-1}^{n} \left|\frac{f'_\sigma}{f_\sigma}(z_t(\boldsymbol{\theta}_0))\right| \sup_{\boldsymbol{\psi}\in Q_\epsilon} |z_t(\boldsymbol{\theta}) - \hat{z}_t(\boldsymbol{\theta})| \\
&\le \sum_{t=-1}^{n} \left|\frac{f'_\sigma}{f_\sigma}(z_t(\boldsymbol{\theta}_0))\right| \left\{\frac{2}{1-d^2} d^{n-t}|\hat{w}_n - w_n| + d^{t+1}|\hat{z}_{-1} - z_{-1}|\right\} \\
&= \frac{2}{1-d^2}|\hat{w}_n - w_n| \left\{\sum_{t=-1}^{n} d^{n-t}\left|\frac{f'_\sigma}{f_\sigma}(z_t(\boldsymbol{\theta}_0))\right|\right\} \\
&\quad + d|\hat{z}_{-1} - z_{-1}| \left\{\sum_{t=-1}^{n} d^{t}\left|\frac{f'_\sigma}{f_\sigma}(z_t(\boldsymbol{\theta}_0))\right|\right\}. \qquad (22)
\end{aligned}
$$

Since $\left\{\frac{f'_\sigma}{f_\sigma}(z_t(\boldsymbol{\theta}_0))\,; t = -1, \ldots, n\right\}$ are iid random variables with

$$
E\left[\frac{f'_\sigma}{f_\sigma}(z_t(\boldsymbol{\theta}_0))\right]^2 = \int \left[\frac{f'_\sigma}{f_\sigma}(z)\right]^2 f_\sigma(z)dz = \int \left[\frac{1}{\sigma}\frac{f'}{f}\left(\frac{z}{\sigma}\right)\right]^2 \frac{1}{\sigma} f\left(\frac{z}{\sigma}\right) dz = \frac{1}{\sigma^2}\int \frac{[f'(z)]^2}{f(z)} dz < \infty,
$$

because of Assumption (A8), by Lemma 2, both series in (22) converge almost surely. Therefore, A_1 converges almost surely.

On the other hand,

$$
\hat{z}_t^*(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}_0) = (z_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}_0)) + u_t(\boldsymbol{\theta})\left(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta})\right)
$$

implies

$$
|\hat{z}_t^*(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}_0)| \le |z_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}_0)| + |\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta})|,
$$

and from Assumption (A9) with $u(\cdot)=\frac{f'}{f}(\cdot)$, A_2 satisfies

$$\begin{aligned}
A_2 &= \sum_{t=-1}^{n}|z_t(\boldsymbol{\theta})-\hat{z}_t(\boldsymbol{\theta})|\frac{1}{\sigma}\left|\frac{f'}{f}\left(\frac{\hat{z}_t^*(\boldsymbol{\theta})}{\sigma}\right)-\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta}_0)}{\sigma}\right)\right| \\
&\le \frac{1}{\sigma}\sum_{t=-1}^{n}|z_t(\boldsymbol{\theta})-\hat{z}_t(\boldsymbol{\theta})|\,A\left\{\left(1+|z_t(\boldsymbol{\theta}_0)/\sigma|^k\right)\sigma^{-1}\,|\hat{z}_t^*(\boldsymbol{\theta})-z_t(\boldsymbol{\theta}_0)|+\sigma^{-\ell}\,|\hat{z}_t^*(\boldsymbol{\theta})-z_t(\boldsymbol{\theta}_0)|^{\ell}\right\} \\
&\le \frac{A}{\sigma}\sum_{t=-1}^{n}|z_t(\boldsymbol{\theta})-\hat{z}_t(\boldsymbol{\theta})|\left\{\left(1+|z_t(\boldsymbol{\theta}_0)/\sigma|^k\right)\sigma^{-1}\left(|z_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta}_0)|+|\hat{z}_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta})|\right)\right. \\
&\qquad\left.+\,\sigma^{-\ell}\left(|z_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta}_0)|+|\hat{z}_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta})|\right)^{\ell}\right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sup_{\boldsymbol{\psi}\in Q_\epsilon} A_2 \;&\le\; \frac{A}{(\sigma_0-\epsilon)^2}\sum_{t=-1}^{n}\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right)\sup_{\boldsymbol{\psi}\in Q_\epsilon}|z_t(\boldsymbol{\theta})-\hat{z}_t(\boldsymbol{\theta})|\sup_{\boldsymbol{\psi}\in Q_\epsilon}|z_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta}_0)| \\
&\quad+\frac{A}{(\sigma_0-\epsilon)^2}\sum_{t=-1}^{n}\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right)\sup_{\psi\in Q_\epsilon}|z_t(\boldsymbol{\theta})-\hat{z}_t(\boldsymbol{\theta})|^2 \\
&\quad+\frac{A}{(\sigma_0-\epsilon)^{\ell+1}}\sum_{t=-1}^{n}\sup_{\boldsymbol{\psi}\in Q_\epsilon}|z_t(\boldsymbol{\theta})-\hat{z}_t(\boldsymbol{\theta})|\left(2^{\ell}\sup_{\boldsymbol{\psi}\in Q_\epsilon}|z_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta}_0)|^{\ell}+2^{\ell}\sup_{\boldsymbol{\psi}\in Q_\epsilon}|\hat{z}_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta})|^{\ell}\right) \\
&\equiv A\left(A_{21}+A_{22}+A_{23}\right),
\end{aligned}$$

where A_{21}, A_{22} and A_{23} are further derived below. According to Equation (3.3') in Lii and Rosenblatt (1992),

$$\begin{aligned}
A_{21} &\le \frac{C\,\epsilon^{1/2}}{(\sigma_0-\epsilon)^2}\sum_{t=-1}^{n}\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right)\left(\sum_{j=-\infty}^{\infty}d^{|j|}|X_{t-j}|\right)\left\{\frac{2}{1-d^2}d^{n-t}|\hat{w}_n-w_n|+d^{t+1}|\hat{z}_{-1}-z_{-1}|\right\} \\
&= \frac{2C\,\epsilon^{1/2}}{(1-d^2)(\sigma_0-\epsilon)^2}|\hat{w}_n-w_n|\sum_{t=-1}^{n}d^{n-t}\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right)\left(\sum_{j=-\infty}^{\infty}d^{|j|}|X_{t-j}|\right) \\
&\quad+\frac{C\,\epsilon^{1/2}\,d}{(\sigma_0-\epsilon)^2}|\hat{z}_{-1}-z_{-1}|\sum_{t=-1}^{n}d^{t}\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right)\left(\sum_{j=-\infty}^{\infty}d^{|j|}|X_{t-j}|\right),
\end{aligned}$$

where the summations in the last equation can be shown to converge almost surely as follows.

$$\begin{aligned}
&\sum_{t=-1}^{n}d^{t}\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right)\left(\sum_{j=-\infty}^{\infty}d^{|j|}|X_{t-j}|\right) \\
&\quad\le \left\{\sum_{t=-1}^{n}d^{t}\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right)^2\right\}^{1/2}\left\{\sum_{t=-1}^{n}d^{t}\left(\sum_{j=-\infty}^{\infty}d^{|j|}|X_{t-j}|\right)^2\right\}^{1/2}, \qquad (23)
\end{aligned}$$

where the first series in (23) converges almost surely by Lemma 1 with

$$
\begin{aligned}
E\log^+\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right)^2 &= 2E\log\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right) \\
\leq\ & 2E\left[\log\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right)1_{\{|z_t(\boldsymbol{\theta}_0)|\leq\sigma_0-\epsilon\}}\right]+2E\left[\log\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right)1_{\{|z_t(\boldsymbol{\theta}_0)|>\sigma_0-\epsilon\}}\right] \\
\leq\ & 2(\log 2)P(|z_t(\boldsymbol{\theta}_0)|\leq\sigma_0-\epsilon)+2E\left[\log\left(2\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right)1_{\{|z_t(\boldsymbol{\theta}_0)|>\sigma_0-\epsilon\}}\right] \\
\leq\ & 2(\log 2)+2k\frac{E|z_t(\boldsymbol{\theta}_0)|}{\sigma_0-\epsilon} \\
<\ & \infty
\end{aligned}
$$

and the second series in (23) also converges almost surely by Lemma 2 with

$$
E\left(\sum_{j=-\infty}^{\infty}d^{|j|}|X_{t-j}|\right)^2\leq EX_t^2\left(\sum_{j=-\infty}^{\infty}d^{|j|}\right)^2<\infty.
$$

Therefore, A_{21} converges almost surely. Similarly, for A_{22}, we have

$$
\begin{aligned}
A_{22} &= \frac{1}{(\sigma_0-\epsilon)^2}\sum_{t=-1}^{n}\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right)\sup_{\boldsymbol{\psi}\in Q_\epsilon}|z_t(\boldsymbol{\theta})-\hat{z}_t(\boldsymbol{\theta})|^2 \\
&\leq \frac{1}{(\sigma_0-\epsilon)^2}\sum_{t=-1}^{n}\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right)\left\{\frac{2}{1-d^2}d^{n-t}|\hat{w}_n-w_n|+d^{t+1}|\hat{z}_{-1}-z_{-1}|\right\}^2 \\
&\leq \frac{4}{(\sigma_0-\epsilon)^2(1-d^2)^2}|\hat{w}_n-w_n|^2\sum_{t=-1}^{n}d^{2(n-t)}\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right) \\
&\quad+\frac{d^2}{(\sigma_0-\epsilon)^2}|\hat{z}_{-1}-z_{-1}|^2\sum_{t=-1}^{n}d^{2t}\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right) \\
&\quad+\frac{2d}{(\sigma_0-\epsilon)^2(1-d^2)}|\hat{w}_n-w_n||\hat{z}_{-1}-z_{-1}|\sum_{t=-1}^{n}d^{t}\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{(\sigma_0-\epsilon)^k}\right).
\end{aligned}
$$

Similar to the derivation for A_{21}, the three series in the last equation all converge almost surely by Lemma 1 and consequently, A_{22} converges almost surely. For A_{23}, we have

$$
\begin{aligned}
A_{23} &\leq \frac{2^\ell C^\ell\epsilon^{\ell/2}}{(\sigma_0-\epsilon)^{\ell+1}}\sum_{t=-1}^{n}\left\{\frac{2}{1-d^2}d^{n-t}|\hat{w}_n-w_n|+d^{t+1}|\hat{z}_{-1}-z_{-1}|\right\}\left(\sum_{j=-\infty}^{\infty}d^{|j|}|X_{t-j}|\right)^\ell \\
&\quad+\frac{2^\ell}{(\sigma_0-\epsilon)^{\ell+1}}\sum_{t=-1}^{n}\left\{\frac{2}{1-d^2}d^{n-t}|\hat{w}_n-w_n|+d^{t+1}|\hat{z}_{-1}-z_{-1}|\right\}^{\ell+1},
\end{aligned}
$$

where the second series converges obviously and, by Lemma 2, the first series converges almost surely if

$$
E\left(\sum_{j=-\infty}^{\infty}d^{|j|}|X_{t-j}|\right)^\ell<\infty. \tag{24}
$$

Consequently, A_{23} converges almost surely. The proof for (19) is complete since

$$\sup_{\boldsymbol{\psi}\in Q_\epsilon} \frac{1}{n}|\ell_n(\boldsymbol{\psi}) - \hat{\ell}_n(\boldsymbol{\psi})| \le n^{-1} \sup_{\boldsymbol{\psi}\in Q_\epsilon} (A_1 + A_2) \to 0, \quad \text{a.s.}$$

We begin to show (20) as follows. First, we consider the partial derivative with respect to $\boldsymbol{\theta}$:

$$\begin{aligned}
\left|\frac{\partial}{\partial\boldsymbol{\theta}}\log f_\sigma(\hat{z}_t(\boldsymbol{\theta})) - \frac{\partial}{\partial\boldsymbol{\theta}}\log f_\sigma(z_t(\boldsymbol{\theta}))\right| &= \left|\frac{f'_\sigma}{f_\sigma}(\hat{z}_t(\boldsymbol{\theta}))\frac{\partial}{\partial\boldsymbol{\theta}}\hat{z}_t(\boldsymbol{\theta}) - \frac{f'_\sigma}{f_\sigma}(z_t(\boldsymbol{\theta}))\frac{\partial}{\partial\boldsymbol{\theta}}z_t(\boldsymbol{\theta})\right| \\
&\le \frac{1}{\sigma}\left|\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta})}{\sigma}\right)\right|\left|\frac{\partial}{\partial\boldsymbol{\theta}}(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}))\right| + \frac{1}{\sigma}\left|\frac{\partial}{\partial\boldsymbol{\theta}}z_t(\boldsymbol{\theta})\right|\left|\frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta})}{\sigma}\right) - \frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta})}{\sigma}\right)\right| \\
&\quad + \frac{1}{\sigma}\left|\frac{\partial}{\partial\boldsymbol{\theta}}(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}))\right|\left|\frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta})}{\sigma}\right) - \frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta})}{\sigma}\right)\right| \\
&\equiv B_{1t} + B_{2t} + B_{3t}. \qquad (25)
\end{aligned}$$

Accordingly, the derivative with respect to $\boldsymbol{\theta}$ in Equation (20) can be bounded by the sum of three series subject to the decomposition in (25). By (16), the first term of (20) subject to B_{1t} in (25) satisfies

$$\begin{aligned}
n^{-1/2}\left[\sum_{t=-1}^{n} B_{1t}\right]_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} &= n^{-1/2}\sigma_0^{-1}\sum_{t=-1}^{n}\left|\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta}_0)}{\sigma_0}\right)\right|\left|\frac{\partial}{\partial\boldsymbol{\theta}}(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}))\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
\le\ & n^{-1/2}\sigma_0^{-1}\sum_{t=-1}^{n}\left|\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta}_0)}{\sigma_0}\right)\right|\left\{\frac{2(n-t+1)d^{n-t+1} + 2td^{n+t+1}}{(1-d^2)^2}|\hat{w}_n - w_n| + (t+1)d^t|\hat{z}_{-1} - z_{-1}|\right\},
\end{aligned}$$

in which, by Lemma 2, all of the series converge almost surely since

$$\sum_{t=-1}^{n} td^t < \infty \quad \text{and} \quad E\left|\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta}_0)}{\sigma_0}\right)\right| < \infty,$$

and therefore, $n^{-1/2}[\sum_{t=-1}^{n} B_{1t}]_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}$ converges to zero almost surely. According to (18), the second term of (20) subject to B_{2t} in (25) satisfies

$$\begin{aligned}
n^{-1/2}\left[\sum_{t=-1}^{n} B_{2t}\right]_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} &= n^{-1/2}\sigma_0^{-1}\sum_{t=-1}^{n}\left|\frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta}_0)}{\sigma_0}\right) - \frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta}_0)}{\sigma_0}\right)\right|\left|\frac{\partial}{\partial\boldsymbol{\theta}}z_t(\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
\le\ & n^{-1/2}\sigma_0^{-1}\sum_{t=-1}^{n} A\left[\left(1 + \frac{|z_t(\boldsymbol{\theta}_0)|^k}{\sigma_0^k}\right)\sigma_0^{-1}|\hat{z}_t^*(\boldsymbol{\theta}_0) - z_t(\boldsymbol{\theta}_0)| + \sigma_0^{-\ell}|\hat{z}_t^*(\boldsymbol{\theta}_0) - z_t(\boldsymbol{\theta}_0)|^\ell\right]\left(\sum_{j=-\infty}^{\infty}\gamma_j|Z_{t-j}|\right).
\end{aligned}$$

Following the similar arguments as those in calculating A_{21}, $n^{-1/2}[\sum_{t=-1}^{n} B_{2t}]_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}$ and $n^{-1/2}[\sum_{t=-1}^{n} B_{3t}]_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}$ both converge to zero almost surely by Lemma 2 since

$$E\left(\sum_{j=-\infty}^{\infty}\gamma_j|Z_{t-j}|\right)^2 \le EZ_t^2\left(\sum_{j=-\infty}^{\infty}\gamma_j\right)^2 < \infty,$$

where $|\gamma_j|$ defined in (18) decay exponentially. Finally, the third term of (20) subject to B_{3t} in (25) satisfies

$$\begin{aligned}
n^{-1/2}\left[\sum_{t=-1}^{n} B_{3t}\right]_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} &= n^{-1/2}\sigma_0^{-1}\sum_{t=-1}^{n}\left|\frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta}_0)}{\sigma_0}\right)-\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta}_0)}{\sigma_0}\right)\right|\left|\frac{\partial}{\partial\boldsymbol{\theta}}\left(\hat{z}_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta})\right)\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
\leq\ & A\,n^{-1/2}\sigma_0^{-1}\sum_{t=-1}^{n}\left\{\left(1+\frac{|z_t(\boldsymbol{\theta}_0)|^k}{\sigma_0^k}\right)\sigma_0^{-1}\left|\hat{z}_t^*(\boldsymbol{\theta}_0)-z_t(\boldsymbol{\theta}_0)\right|+\sigma_0^{-\ell}\left|\hat{z}_t^*(\boldsymbol{\theta}_0)-z_t(\boldsymbol{\theta}_0)\right|^{\ell}\right\} \\
&\times\left\{\frac{2(n-t+1)d^{n-t+1}+2td^{n+t+1}}{(1-d^2)^2}|\hat{w}_n-w_n|+(t+1)d^t|\hat{z}_{-1}-z_{-1}|\right\} \\
\to\ & \mathbf{0}, \quad a.s.,
\end{aligned}$$

under similar arguments as those in calculating A_{22}.

To complete (20), we then consider the partial derivative with respect to σ:

$$\begin{aligned}
&\left|\frac{\partial}{\partial\sigma}\log f_\sigma(\hat{z}_t(\boldsymbol{\theta}))-\frac{\partial}{\partial\sigma}\log f_\sigma(z_t(\boldsymbol{\theta}))\right| = \frac{1}{\sigma^2}\left|\frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta})}{\sigma}\right)\hat{z}_t(\boldsymbol{\theta})-\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta})}{\sigma}\right)z_t(\boldsymbol{\theta})\right| \\
\leq\ & \frac{1}{\sigma^2}\left|\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta})}{\sigma}\right)\right|\left|\hat{z}_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta})\right|+\frac{1}{\sigma^2}|z_t(\boldsymbol{\theta})|\left|\frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta})}{\sigma}\right)-\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta})}{\sigma}\right)\right| \\
&+\frac{1}{\sigma^2}\left|\hat{z}_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta})\right|\left|\frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta})}{\sigma}\right)-\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta})}{\sigma}\right)\right| \\
\equiv\ & B_{4t}+B_{5t}+B_{6t}. \qquad (26)
\end{aligned}$$

Accordingly, the derivative with respect to σ in Equation (20) can be bounded by the sum of three series subject to the decomposition in (26). Similar to deriving the convergence for A_1, $\sum_t B_{4t}$ and $\sum_t B_{6t}$ at $\boldsymbol{\psi}=\boldsymbol{\psi}_0$ both converge almost surely. Also, since

$$z_t(\boldsymbol{\theta}_0)=\pi(B)X_t=\sum_{j=-\infty}^{\infty}\pi_j X_{t-j},$$

where $\pi(B)=[\theta^\dagger(B)]^{-1}\left[\theta_s^* B^s\tilde{\theta}(B^{-1})\right]^{-1}$ with $\sum_j \pi_j<\infty$, the upper bound for $\sum_t B_{5t}$ at $\boldsymbol{\psi}=\boldsymbol{\psi}_0$ converges almost surely under similar arguments as those for B_{2t}. Therefore, (20) holds from the above results.

It is left to show (21). First, let

$$n^{-1}(\hat{\boldsymbol{B}}(\boldsymbol{\psi}^*)-\boldsymbol{B}(\boldsymbol{\psi}_0))=n^{-1}(\boldsymbol{B}(\boldsymbol{\psi}^*)-\boldsymbol{B}(\boldsymbol{\psi}_0))+n^{-1}(\hat{\boldsymbol{B}}(\boldsymbol{\psi}^*)-\boldsymbol{B}(\boldsymbol{\psi}^*)).$$

In Lii and Rosenblatt (1992), it has been shown that

$$n^{-1}(\boldsymbol{B}(\boldsymbol{\psi}^*)-\boldsymbol{B}(\boldsymbol{\psi}_0))\to 0, \quad \text{in probability},$$

as $\boldsymbol{\psi}^*\to\boldsymbol{\psi}_0$. In the following, we only need to show

$$n^{-1}(\hat{\boldsymbol{B}}(\boldsymbol{\psi}^*)-\boldsymbol{B}(\boldsymbol{\psi}^*))\to 0, \quad \text{in probability}.$$

Since

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f_\sigma(z_t(\boldsymbol{\theta})) = \frac{1}{\sigma}\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta})}{\sigma}\right)\frac{\partial^2 z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \frac{1}{\sigma^2}\left(\frac{f'}{f}\right)'\left(\frac{z_t(\boldsymbol{\theta})}{\sigma}\right)\frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'},$$

we have

$$\begin{aligned}
\left|\hat{\boldsymbol{B}}_{\theta\theta}(\boldsymbol{\psi}^*) - \boldsymbol{B}_{\theta\theta}(\boldsymbol{\psi}^*)\right| &\le \sum_{t=-1}^{n}\left|\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f_\sigma(\hat{z}_t(\boldsymbol{\theta})) - \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f_\sigma(z_t(\boldsymbol{\theta}))\right|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*} \\
\le\ & \frac{1}{\sigma^*}\sum_{t=-1}^{n}\left|\frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\left[\frac{\partial^2 \hat{z}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} - \frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\left[\frac{\partial^2 z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right| \\
& + \frac{1}{\sigma^{*2}}\sum_{t=-1}^{n}\left|\left(\frac{f'}{f}\right)'\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\left[\frac{\partial \hat{z}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \hat{z}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} - \left(\frac{f'}{f}\right)'\left(\frac{z_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\left[\frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right| \\
\le\ & \frac{1}{\sigma^*}\sum_{t=-1}^{n}\left|\frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\right|\left|\frac{\partial^2(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} + \frac{1}{\sigma^*}\sum_{t=-1}^{n}\left|\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta}^*)}{\sigma^*}\right) - \frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\right|\left|\frac{\partial^2 z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \\
& + \frac{1}{\sigma^{*2}}\sum_{t=-1}^{n}\left|\left(\frac{f'}{f}\right)'\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\right|\left|\frac{\partial(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}}\frac{\partial(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}'} + \frac{\partial(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}}\frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right. \\
& \qquad\qquad \left. + \frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial(\hat{z}_t(\boldsymbol{\theta}) - z_t(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \\
& + \frac{1}{\sigma^{*2}}\sum_{t=-1}^{n}\left|\left(\frac{f'}{f}\right)'\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right) - \left(\frac{f'}{f}\right)'\left(\frac{z_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\right|\left|\frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}. \qquad (27)
\end{aligned}$$

Using similar techniques used in the previous proof, it can be shown that each series in the decomposition in (27) converges almost surely. Similarly, since

$$\frac{\partial^2}{\partial\sigma^2}\log f_\sigma(z_t(\boldsymbol{\theta})) = \frac{z_t^2(\boldsymbol{\theta})}{\sigma^4}\left(\frac{f'}{f}\right)'\left(\frac{z_t(\boldsymbol{\theta})}{\sigma}\right) + \frac{2\,z_t(\boldsymbol{\theta})}{\sigma^3}\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta})}{\sigma}\right),$$

we have

$$\begin{aligned}
|\hat{\boldsymbol{B}}_{\sigma\sigma}(\boldsymbol{\psi}^*) - \boldsymbol{B}_{\sigma\sigma}(\boldsymbol{\psi}^*)| \ \le\ & \frac{1}{\sigma^{*4}}\sum_{t=-1}^{n}\left|\left(\frac{f'}{f}\right)'\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\hat{z}_t^2(\boldsymbol{\theta}^*) - \left(\frac{f'}{f}\right)'\left(\frac{z_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)z_t^2(\boldsymbol{\theta}^*)\right| \\
& + \frac{2}{\sigma^{*3}}\sum_{t=-1}^{n}\left|\frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\hat{z}_t(\boldsymbol{\theta}^*) - \frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)z_t(\boldsymbol{\theta}^*)\right|. \qquad (28)
\end{aligned}$$

Ignoring the scalar $1/\sigma^{*4}$, the first term on the right hand side of (28) can be bounded by

$$\begin{aligned}
& \sum_{t=-1}^{n}\left|\left(\frac{f'}{f}\right)'\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right) - \left(\frac{f'}{f}\right)'\left(\frac{z_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\right|z_t^2(\boldsymbol{\theta}^*) + \left|\left(\frac{f'}{f}\right)'\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\right||\hat{z}_t^2(\boldsymbol{\theta}^*) - z_t^2(\boldsymbol{\theta}^*)| \\
\le\ & \sum_{t=-1}^{n}\left|\left(\frac{f'}{f}\right)'\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right) - \left(\frac{f'}{f}\right)'\left(\frac{z_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\right|z_t^2(\boldsymbol{\theta}^*) + \left\{(\hat{z}_t(\boldsymbol{\theta}^*) - z_t(\boldsymbol{\theta}^*))^2 + 2|z_t(\boldsymbol{\theta}^*)||\hat{z}_t(\boldsymbol{\theta}^*) - z_t(\boldsymbol{\theta}^*)|\right\} \\
& \times\left|\left(\frac{f'}{f}\right)'\left(\frac{z_t(\boldsymbol{\theta}_0)}{\sigma_0}\right) + \left\{\left(\frac{f'}{f}\right)'\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right) - \left(\frac{f'}{f}\right)'\left(\frac{z_t(\boldsymbol{\theta}_0)}{\sigma_0}\right)\right\}\right|.
\end{aligned}$$

Analogous to the previous proof, it can be shown that the series associated with each term in the above decomposition converges with probability one. The second term on the right hand side of (28) is identical to

$$\frac{2}{\sigma^*}\sum_{t=-1}^{n}\left|\frac{\partial}{\partial\sigma}\log f_\sigma(\hat{z}_t(\boldsymbol{\theta}))-\frac{\partial}{\partial\sigma}\log f_\sigma(z_t(\boldsymbol{\theta}))\right|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}$$

which also converges with probability one using the same decomposition given in (26).

Finally, we consider $\boldsymbol{B}_{\theta\sigma}$. Since

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\sigma}\log f_\sigma(z_t(\boldsymbol{\theta}))=-\frac{1}{\sigma^3}\left(\frac{f'}{f}\right)'\left(\frac{z_t(\boldsymbol{\theta})}{\sigma}\right)z_t(\boldsymbol{\theta})\frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}-\frac{1}{\sigma^2}\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta})}{\sigma}\right)\frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}},$$

we have

$$\begin{aligned}
|\hat{\boldsymbol{B}}_{\theta\sigma}(\boldsymbol{\psi}^*)-\boldsymbol{B}_{\theta\sigma}(\boldsymbol{\psi}^*)| &\le \sum_{t=-1}^{n}\left|\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\sigma}\log f_\sigma(\hat{z}_t(\boldsymbol{\theta}))-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\sigma}\log f_\sigma(z_t(\boldsymbol{\theta}))\right|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}\\
&\le \frac{1}{\sigma^{*3}}\sum_{t=-1}^{n}\left|\left(\frac{f'}{f}\right)'\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\hat{z}_t(\boldsymbol{\theta}^*)\left[\frac{\partial\hat{z}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}-\left(\frac{f'}{f}\right)'\left(\frac{z_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)z_t(\boldsymbol{\theta}^*)\left[\frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right|\\
&\quad+\frac{1}{\sigma^{*2}}\sum_{t=-1}^{n}\left|\frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\left[\frac{\partial\hat{z}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}-\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\left[\frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right|\\
&\le \frac{1}{\sigma^{*3}}\sum_{t=-1}^{n}\left|\left(\frac{f'}{f}\right)'\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)(\hat{z}_t(\boldsymbol{\theta}^*)-z_t(\boldsymbol{\theta}^*))\right|\left|\frac{\partial(\hat{z}_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\\
&\quad+\frac{1}{\sigma^{*3}}\sum_{t=-1}^{n}\left|\left(\frac{f'}{f}\right)'\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\right|\left|z_t(\boldsymbol{\theta}^*)\frac{\partial(\hat{z}_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}}+(\hat{z}_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta}))\frac{\partial z_t(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\\
&\quad+\frac{1}{\sigma^{*3}}\sum_{t=-1}^{n}\left|\left(\frac{f'}{f}\right)'\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)-\left(\frac{f'}{f}\right)'\left(\frac{z_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\right||z_t(\boldsymbol{\theta}^*)|\left|\frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\\
&\quad+\frac{1}{\sigma^{*2}}\sum_{t=-1}^{n}\left|\frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\right|\left|\frac{\partial(\hat{z}_t(\boldsymbol{\theta})-z_t(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\\
&\quad+\frac{1}{\sigma^{*2}}\sum_{t=-1}^{n}\left|\frac{f'}{f}\left(\frac{\hat{z}_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)-\frac{f'}{f}\left(\frac{z_t(\boldsymbol{\theta}^*)}{\sigma^*}\right)\right|\left|\frac{\partial z_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}. \qquad (29)
\end{aligned}$$

Similarly, it can be shown that the series associated with each term in the decomposition in (29) converges with probability one. Consequently, $\hat{\boldsymbol{B}}(\boldsymbol{\psi}^*)-\boldsymbol{B}(\boldsymbol{\psi}^*)$ converges with probability one and $n^{-1}(\hat{\boldsymbol{B}}(\boldsymbol{\psi}^*)-\boldsymbol{B}(\boldsymbol{\psi}^*))\to\mathbf{0}$.

From (19), as $n\to\infty$, there exists a consistent sequence of estimators $\{\hat{\boldsymbol{\psi}}_n\in Q_\epsilon\}$ satisfying the proposed likelihood equations

$$\frac{\partial}{\partial\boldsymbol{\psi}}\hat{\ell}_n(\boldsymbol{\psi})=\mathbf{0}.$$

By Taylor expansion, we have

$$\mathbf{0} \;=\; n^{-1/2}\left[\frac{\partial}{\partial\boldsymbol{\psi}}\hat{\ell}_n(\boldsymbol{\psi})\right]_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}_n}$$

$$
\begin{aligned}
&= n^{-1/2}\sum_{t=-1}^{n}\left[\frac{\partial}{\partial\boldsymbol{\psi}}\log f_\sigma(\hat{z}_t(\boldsymbol{\theta}))\right]_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} + n^{-1}\hat{\boldsymbol{B}}(\boldsymbol{\psi}^*)\,n^{1/2}(\hat{\boldsymbol{\psi}}_n-\boldsymbol{\psi}_0)\\
&= n^{-1/2}\sum_{t=-1}^{n}\left[\frac{\partial}{\partial\boldsymbol{\psi}}\log f_\sigma(z_t(\boldsymbol{\theta}))\right]_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} + n^{-1/2}\sum_{t=-1}^{n}\left[\frac{\partial}{\partial\boldsymbol{\psi}}\log f_\sigma(\hat{z}_t(\boldsymbol{\theta})) - \frac{\partial}{\partial\boldsymbol{\psi}}\log f_\sigma(z_t(\boldsymbol{\theta}))\right]_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}\\
&\quad + n^{-1}\boldsymbol{B}(\boldsymbol{\psi}_0)\,n^{1/2}(\hat{\boldsymbol{\psi}}_n-\boldsymbol{\psi}_0) + n^{-1}\left[\hat{\boldsymbol{B}}(\boldsymbol{\psi}^*)-\boldsymbol{B}(\boldsymbol{\psi}_0)\right]\,n^{1/2}(\hat{\boldsymbol{\psi}}_n-\boldsymbol{\psi}_0),
\end{aligned}
$$

where $\boldsymbol{\psi}^*$ is between $\hat{\boldsymbol{\psi}}_n$ and $\boldsymbol{\psi}_0$. According to (20) and (21),

$$
\mathbf{0} = n^{-1/2}\sum_{t=-1}^{n}\left[\frac{\partial}{\partial\boldsymbol{\psi}}\log f_\sigma(z_t(\boldsymbol{\theta}))\right]_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} + n^{-1}\boldsymbol{B}(\boldsymbol{\psi}_0)\,n^{1/2}(\hat{\boldsymbol{\psi}}_n-\boldsymbol{\psi}_0) + o_p(1),
$$

which implies

$$
\begin{aligned}
n^{1/2}(\hat{\boldsymbol{\psi}}_n-\boldsymbol{\psi}_0) &= \left[n^{-1}\boldsymbol{B}(\boldsymbol{\psi}_0)\right]^{-1}\left\{-n^{-1/2}\sum_{t=-1}^{n}\left[\frac{\partial}{\partial\boldsymbol{\psi}}\log f_\sigma(z_t(\boldsymbol{\theta}))\right]_{\boldsymbol{\psi}=\boldsymbol{\psi}_0}\right\} + o_p(1)\\
&\to N\left(0,\Sigma^{-1}\right),
\end{aligned}
$$

where Σ is defined in (1.7) and (1.8) in Lii and Rosenblatt (1992).

References

Benveniste, A., Goursat, M., and Roget, G. (1980). Robust identification of a nonminimum phase system: blind adjustment of linear equalizer in data communications. *IEEE Transactions on Automatic Control* **AC–25**, 385–398.

Blass, W.E. and Halsey, G.W. (1981). *Deconvolution of Absorption Spectra.* Academic Press, New York.

Breidt, F.J., Davis, R.A., and Trindade, A.A. (2001). Least absolute deviation estimation for all-pass time series models. *Annals of Statistics* **29**, 919–946.

Breidt, J.F. and Hsu, N.-J. (2004). Best mean square prediction for moving averages. forthcoming in *Statistical Sinica.*

Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods,* 2nd ed. Springer-Verlag, New York.

Chi, C.-Y., and Kung, J.-Y. (1995). A new identification algorithm for allpass systems by higher-order statistics. *Signal Processing* **41**, 239–256.

Chien, H.-M., Yang, H.-L., and Chi, C.-Y. (1997). Parametric cumulant base phase estimation of 1-d and 2-d nonminimum phase systems by allpass filtering. *IEEE Transactions on Signal Processing* **45**, 1742–1762.

Donoho, D. (1981). On minimum entropy deconvolution. In *Applied Time Series Analysis II*, (D.F. Findley, ed.), Academic Press, New York, pp. 565–608.

Evans, M. and Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods.* Oxford University Press, Oxford, UK.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis.* Chapman & Hall, London, UK.

Giannakis, G.B., and Swami, A. (1990). On estimating noncausal nonminimum phase ARMA models of non-Gaussian processes. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **38**, 478–495.

Godfrey, R. and Rocca, F. (1981). Zero memory nonlinear deconvolution. *Geophysical Prospecting* **29**, 189–228.

Hildebrand, F.B. (1968). *Finite-Difference Equations and Simulations.* Prentice-Hall, Englewood Cliffs, NJ.

Hsueh, A.-C. and Mendel, J.M. (1985). Minimum-variance and maximum-likelihood deconvolution for non-causal channel models. *IEEE Transactions on Geoscience and Remote Sensing* **23**, 797–808.

Huang, J. and Pawitan, Y. (2000). Quasi-likelihood estimation of non-invertible moving average processes. *Scandinavian Journal of Statistics* **27**, 689–702.

Lii, K.-S. and Rosenblatt, M. (1982). Deconvolution and estimation of transfer function phase and coefficients for nonGaussian linear processes. *Annals of Statistics* **10**, 1195–1208.

Lii, K.-S. and Rosenblatt, M. (1992). An approximate maximum likelihood estimation for non-Gaussian non-minimum phase moving average processes. *Journal of Multivariate Analysis* **43**, 272–299.

Lii, K.-S. and Rosenblatt, M. (1996). Maximum likelihood estimation for nonGaussian nonminimum phase ARMA sequences. *Statistica Sinica* **6**, 1–22.

Ooe, M. and Ulrych, T.J. (1979). Minimum entropy deconvolution with an exponential transformation. *Geophysical Prospecting* **27**, 458–473.

Rabiner, L.R. and Schafer, R.M. (1978). *Digital Processing of Speech Signals.* Prentice-Hall, Englewood Cliffs, New Jersey.

Rosenblatt, M. (2000). *Gaussian and Non-Gaussian Linear Time Series and Random Fields.* Springer-Verlag, New York.

Scargle, J.D. (1981). Phase-sensitive deconvolution to model random processes, with special reference to astronomical data. In *Applied Time Series Analysis II*, (D.F. Findley, ed.), Academic Press, New York, 549–564.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *Annals of Statistics* **22**, 1701-1762.

Wiggins, R.A. (1978). Minimum entropy deconvolution. *Geoexploration* **16**, 21–35.