Statistica Sinica **19** (2009), 545-560

EXACT MAXIMUM LIKELIHOOD ESTIMATION FOR NON-GAUSSIAN MOVING AVERAGES

Nan-Jung Hsu and F. Jay Breidt

National Tsing-Hua University and Colorado State University

Abstract: A procedure for computing exact maximum likelihood estimates (MLEs) is proposed for non-Gaussian moving average (MA) processes. By augmenting the data with appropriate latent variables, a joint likelihood can be explicitly expressed based on the observed data and the latent variables. The exact MLE can then be obtained numerically by the EM algorithm. Two alternative likelihood-based methods are also proposed using different treatments of the latent variables. These approximate MLEs are shown to be asymptotically equivalent to the exact MLE. In simulations, the exact MLE obtained by EM performs better than other likelihood-based estimators, including another approximate MLE due to Lii and Rosenblatt (1992). The exact MLE has a smaller root mean square error in small samples for various non-Gaussian MA processes, particularly for the non-invertible cases.

Key words and phrases: EM algorithm, Monte Carlo, non-invertible, non-minimum phase.

1. Introduction

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

$$X_t = \theta(B)Z_t, \tag{1.1}$$

where $\{Z_t\}$ is an independent and identically distributed (i.i.d.) sequence of random variables with zero mean and finite variance, B is the backshift operator (i.e., $B^kY_t = Y_{t-k}$),

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

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 otherwise (Brockwell and Davis (1991, Thm. 3.1.2))

If the process $\{X_t\}$ is Gaussian, the polynomial $\theta(z)$ is not identifiable unless the invertibility assumption, or the minimum phase condition, is further imposed. This is because the probability structure of $\{X_t\}$ is fully determined by the modulus of $\theta(e^{-i\omega})$ (or the second order spectrum) in the Gaussian case, and the phase information of $\theta(e^{-i\omega})$ is not available. However, for a non-Gaussian process, the phase information is available based on the likelihood function or the higher-order spectra (which are zero for Gaussian processes), and therefore the model (1.1) becomes identifiable. In this study, we are interested in estimating $\theta(z)$ for a non-Gaussian process without imposing the invertibility assumption. Such non-Gaussian MA processes are useful for the deconvolution problems that arise in such fields as seismology, signal processing, astronomy, and engineering (Wiggins (1978), Donoho (1981), Scargle (1981) and Mendel (1991)). For example in seismology, $\{X_t\}$ is the observed seismogram sequence and the question of interest is to determine the weights $\{\theta_j\}$ that represent the signature of a disturbance passing through a medium, and to recover the signal $\{Z_t\}$ as the seismic reflectivity of Earth, which is typically non-Gaussian distributed and usually has a spiky appearance (Lii and Rosenblatt (1982)). As another example in signal processing, a voiced speech signal can be modeled by (1.1) as the output of a non-Gaussian (quasi-periodic) impulse train $\{Z_t\}$ passing through the vocal tract filter $\theta(B)$ (Rabiner and Schafer (1978)). More applications can be found in the above-mentioned references, and in Breidt and Hsu (2005).

In the literature, there are two likelihood-based estimation methods for (1.1) that do not impose the invertibility assumption. One is the quasi-likelihood method that leads to the least absolute deviation (LAD) estimators (Huang and Pawitan (2000)). The other is the approximate maximum likelihood estimation (Lii and Rosenblatt (1992)), which was also generalized for the non-minimum phase ARMA models (Lii and Rosenblatt (1996)). The approach by Huang and Pawitan (2000) uses the Laplacian likelihood but does not require the assumption of Laplacian errors. Conditioning on some initial variables set to zero, the maximizer of Huang and Pawitan's conditional Laplace likelihood was shown to be consistent only for a process with heavy tailed errors. Alternatively, Lii and Rosenblatt (1992) considered a truncation in the representation of the innovations in terms of all available observations, and approximated the likelihood function based on these truncated innovations. The maximizer of the truncated likelihood was shown to be asymptotically equivalent to the exact MLE under mild conditions. This truncation scheme provides a feasible implementation for non-Gaussian processes, though it causes some information loss in estimation that is negligible asymptotically, but more serious in small samples. In this study, we adopted ideas from Breidt and Hsu (2005) to give an expression for the exact likelihood by introducing q latent variables. Exact MLEs of the parameters for general MA processes can then be obtained. In addition, we propose two alternative estimators through different treatments of the latent variables. One is called the conditional MLE, in which the latent variables are set to be zero. The other is called the joint MLE, in which the latent variables are estimated simultaneously with the model parameters. We show that these alternative estimators have the same asymptotic distribution as the exact 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 the residuals and the likelihood for a MA process. In Section 3, the procedures for solving the exact MLE by the EM algorithm and two alternative estimators are introduced. In Section 4, numerical simulations are conducted to evaluate the performance of different estimators in finite samples for various non-Gaussian MA processes. A brief discussion follows in Section 5.

2. MA Processes and the Likelihoods

Assume the order q is finite and known. Rewrite (1.1) as

$$X_t = \theta(B)Z_t = \theta^{\dagger}(B)\theta^{*}(B)Z_t, \tag{2.1}$$

with

$$\theta^{*}(z) = 1 + \theta_1^{*}z + \cdots + \theta_s^{*}z^s \neq 0 \quad \text{for } |z| \geq 1,$$
$$\theta^{\dagger}(z) = 1 + \theta_1^{\dagger}z + \cdots + \theta_r^{\dagger}z^r \neq 0 \quad \text{for } |z| \leq 1,$$

where $r + s = q$, $\theta_s^{*} \neq 0$ if $s \neq 0$, and $\theta_r^{\dagger} \neq 0$ if $r \neq 0$. The moving average polynomial $\theta(z)$ is invertible if $s = 0$, and non-invertible otherwise. Here we only consider the cases without unit roots (i.e., $\theta(B)$ has no roots on the unit circle) since the asymptotic theory for the unit root cases is completely different and more complicated. Some recent asymptotic results for a non-Gaussian MA(1) with unit root can be found in Breidt, Davis, Hsu and Rosenblatt (2006). (See the references in that paper for the Gaussian MA(1) with unit root.)

According to Breidt and Hsu (2005), define

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

so that

$$Z_t = W_t - (\theta^{\dagger}(B) - 1)Z_t. \tag{2.2}$$

Then,

$$\begin{aligned} X_t &= \theta^{*}(B)W_t \\ &= (1 + \theta_1^{*}B + \cdots + \theta_s^{*}B^s)W_t \\ &= \theta_s^{*}\tilde{\theta}(B^{-1})W_{t-s}, \end{aligned} \tag{2.3}$$

where

$$\tilde{\theta}(z) = 1 + \frac{\theta_{s-1}^{*}}{\theta_s^{*}}z + \cdots + \frac{\theta_1^{*}}{\theta_s^{*}}z^{s-1} + \frac{1}{\theta_s^{*}}z^s. \tag{2.4}$$

Consequently,

$$W_{t-s} = \frac{X_t}{\theta_s^*} - \left(\tilde{\theta}(B^{-1}) - 1\right) W_{t-s}. \tag{2.5}$$

By incorporating the latent variables $\boldsymbol{Z}_r = (Z_{-q+1}, \ldots, Z_{-q+r})'$ and $\boldsymbol{W}_s = (W_{n-s+1}, \ldots, W_n)'$, the random vector $(\boldsymbol{X}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s)$ is a linear transformation of the residuals $(Z_{-q+1}, \ldots, Z_{-1}, Z_0, Z_1, \ldots, Z_n)$, and consequently the joint distribution of $(\boldsymbol{X}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s)$ satisfies

$$p(\boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s; \boldsymbol{\theta}, \sigma) = |\theta_s^*|^{-n} \prod_{t=-q+1}^{n} f_\sigma(z_t(\boldsymbol{\theta})), \tag{2.6}$$

where $f_\sigma(z) = \sigma^{-1} f(z/\sigma)$ is the probability density function of Z_t with the scale parameter σ, $\boldsymbol{\theta} = (\boldsymbol{\theta}^\dagger, \boldsymbol{\theta}^*)$ consists of the MA parameters with $\boldsymbol{\theta}^\dagger \equiv (\theta_1^\dagger, \ldots, \theta_r^\dagger)'$ and $\boldsymbol{\theta}^* \equiv (\theta_1^*, \ldots, \theta_s^*)'$ representing the invertible and non-invertible parts, respectively. Note that $\{z_t(\boldsymbol{\theta}) : t = -q+1, \ldots, n\}$ in (2.6) 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, which can be solved recursively by (2.3) and (2.5) with the initial conditions $(\boldsymbol{Z}_r, \boldsymbol{W}_s)$.

Since the latent variables $\boldsymbol{Z}_r$ and $\boldsymbol{W}_s$ are unobserved in practice, we propose three estimators of $\boldsymbol{\theta}$ by maximizing the joint distribution (2.6) subject to different treatments of the latent variables. Details about these estimators are described in the next section.

3. Estimation Methods

Three estimators of $\boldsymbol{\psi} = (\boldsymbol{\theta}', \sigma)'$ are introduced for a non-Gaussian MA process based on the joint likelihood in (2.6). The first estimator is the exact MLE solved by the EM algorithm in which the latent variables $(\boldsymbol{Z}_r, \boldsymbol{W}_s)$ are treated as missing data and are integrated out in the log joint likelihood. The algorithm is given in the following two steps.

- E-step: Compute the conditional expectation of the log likelihood $\log p(\boldsymbol{X}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s; \boldsymbol{\theta}, \sigma)$ given $\boldsymbol{X}_n = \boldsymbol{x}_n$ which satisfies

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

  where the expectation E_{old} is taken on the latent variables $(\boldsymbol{Z}_r, \boldsymbol{W}_s)$ with respect to the conditional distribution given $\boldsymbol{X}_n = \boldsymbol{x}_n$ evaluated at the current parameter estimates $\boldsymbol{\psi}_{old} = (\boldsymbol{\theta}'_{old}, \sigma_{old})'$.
- M-step: Update $\boldsymbol{\psi}$ by $\boldsymbol{\psi}_{new} = \arg\max_{\boldsymbol{\psi}} Q(\boldsymbol{\psi}|\boldsymbol{\psi}_{old})$.

These two steps are repeated recursively until convergence is achieved. In general, the maximizer of the scale parameter σ in the M-step has a closed form given $\boldsymbol{\theta}$, simplifying the maximization in the M-step. However the expectation in the E-step is analytically intractable and therefore we adopt a Monte Carlo (MC) method to compute the expectation numerically, which is exactly the MCEM algorithm introduced by Wei and Tanner (1990). The same idea was used in Breidt and Hsu (2005) for obtaining the best mean square predictors for non-Gaussian MA processes.

More precisely, the expectation in the E-step can be expressed as

$$\begin{aligned} Q_t(\boldsymbol{\psi}|\boldsymbol{\psi}_{old}) &\equiv E_{old}\left[\log f_\sigma\left(z_t(\boldsymbol{\theta}, \boldsymbol{X}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s)\right)|\boldsymbol{X}_n = \boldsymbol{x}_n\right] \\ &= \frac{\int \log f_\sigma\left(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s)\right) p(\boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s; \boldsymbol{\psi}_{old}) d\boldsymbol{z}_r d\boldsymbol{w}_s}{\int p(\boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s; \boldsymbol{\psi}_{old}) d\boldsymbol{z}_r d\boldsymbol{w}_s} \\ &= \frac{\int \log f_\sigma\left(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s)\right)\left[\prod_{k=-q+1}^n f_{\sigma_{old}}(z_k(\boldsymbol{\theta}_{old}, \boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s))\right] d\boldsymbol{z}_r d\boldsymbol{w}_s}{\int \prod_{k=-q+1}^n f_{\sigma_{old}}(z_k(\boldsymbol{\theta}_{old}, \boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s)) d\boldsymbol{z}_r d\boldsymbol{w}_s}. \quad (3.1) \end{aligned}$$

Following Breidt and Hsu (2005), these q-dimensional integrals can be evaluated by a MC method with importance sampling as follows. Let $h(\boldsymbol{z}_r, \boldsymbol{w}_s)$ be a q-dimensional joint density (called the importance sampler) satisfying

$$\operatorname{supp}(h) \supset \operatorname{supp}\left(\int p(\boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s; \boldsymbol{\psi}_{old})\, d\boldsymbol{x}_n\right), \quad (3.2)$$

where supp(h) is the support of h. For any $\left(\boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)}\right) \in \operatorname{supp}(h)$, 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 (3.1) can be written as

$$\begin{aligned} Q_t(\boldsymbol{\psi}|\boldsymbol{\psi}_{old}) &= \frac{\int \log f_\sigma\left(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s)\right) A\left(\boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s; \boldsymbol{\psi}_{old}\right) h\left(\boldsymbol{z}_r, \boldsymbol{w}_s\right) d\boldsymbol{z}_r d\boldsymbol{w}_s}{\int A\left(\boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s; \boldsymbol{\psi}_{old}\right) h\left(\boldsymbol{z}_r, \boldsymbol{w}_s\right) d\boldsymbol{z}_r d\boldsymbol{w}_s} \\ &= \frac{E_h\left[\log f_\sigma\left(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s)\right) A\left(\boldsymbol{x}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s; \boldsymbol{\psi}_{old}\right)\right]}{E_h\left[A\left(\boldsymbol{x}_n, \boldsymbol{Z}_r, \boldsymbol{W}_s; \boldsymbol{\psi}_{old}\right)\right]}, \quad (3.3) \end{aligned}$$

where the expectation in (3.3) is taken with respect to $h(\boldsymbol{z}_r, \boldsymbol{w}_s)$. The Monte Carlo approximation for the ratio in (3.3) is

$$\hat{Q}_t(\boldsymbol{\psi}|\boldsymbol{\psi}_{old}) = \frac{\sum_{i=1}^M \log f_\sigma\left(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)})\right) A\left(\boldsymbol{x}_n, \boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)}; \boldsymbol{\psi}_{old}\right)}{\sum_{i=1}^M A\left(\boldsymbol{x}_n, \boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)}; \boldsymbol{\psi}_{old}\right)}, (3.4)$$

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

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

which is the one used in the implementation of the E-step. Issues about the convergence of the MCEM algorithm, the choice of the MC sample size M and the computational cost can be found in Levine and Casella (2001) and the references therein.

Theoretically, the choice of the importance sampler $h(\boldsymbol{z}_r, \boldsymbol{w}_s)$ is fairly arbitrary as long as the condition in (3.2) is satisfied. However, the performance of the MC estimator depends on the choice of $h(\boldsymbol{z}_r, \boldsymbol{w}_s)$. One should avoid using an importance sampler that produces a lot of small importance weights and a few extremely large weights, since this makes the variability of the MC estimator large. Moreover, variance reduction methods can be incorporated with the importance sampling technique to further improve the accuracy of the algorithm. Methods for choosing and evaluating importance samplers and for reducing variance in importance sampling are thoroughly reviewed in Evans and Swartz (2000). For simplicity, one may use the marginal distributions of $\boldsymbol{Z}_r$ and $\boldsymbol{W}_s$ or the conditional distribution of $(\boldsymbol{Z}_r, \boldsymbol{W}_s)$ given $\boldsymbol{X}_n = \boldsymbol{x}_n$ under the Gaussian assumption.

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

$$\hat{\boldsymbol{\psi}}_c = \arg\max_{\boldsymbol{\psi}} \Big\{ -n \log|\theta_s^*| + \sum_{t=-q+1}^{n} \log f_\sigma\,(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{0}, \boldsymbol{0})) \Big\}. \tag{3.5}$$

The third estimator is the joint MLE in which the latent variables are estimated simultaneously with the other parameters; that is

$$(\hat{\boldsymbol{\psi}}_J, \hat{\boldsymbol{z}}_r, \hat{\boldsymbol{w}}_s) = \arg \max_{(\boldsymbol{\psi}, \boldsymbol{z}_r, \boldsymbol{w}_s)} \Big\{ -n \log|\theta_s^*| + \sum_{t=-q+1}^{n} \log f_\sigma\,(z_t(\boldsymbol{\theta}, \boldsymbol{x}_n, \boldsymbol{z}_r, \boldsymbol{w}_s)) \Big\}. \tag{3.6}$$

Lii and Rosenblatt (1992) studied the same estimation problem and proposed an approximate likelihood approach. In their approach, the innovation Z_t is first represented as a linear combination of both past and future data $\{X_t : t = 0, \pm 1, \pm 2, \ldots\}$ and then this infinite series is further approximated by a truncated series only in terms of observed data $\{X_t : t = 1, \ldots, n\}$. Accordingly, the likelihood function is approximated by plugging in these truncated innovations. The corresponding maximizer, denoted as LR MLE, is shown to be asymptotically equivalent to the exact MLE under some mild conditions. Under similar conditions, we show that the conditional MLE (3.5) and the joint MLE (3.6) are both equivalent to the exact MLE asymptotically.

Theorem 1. *Assume that the density of the innovations* $\{Z_t\}$ *satisfies* $f_\sigma(z) = \sigma^{-1}f(z/\sigma)$, $EZ_t^4 < \infty$, *and that the following conditions hold.*

$A1 : f(x) > 0$ *for all* x;

$A2 : f \in C^2$;

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

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

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

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

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

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

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

Then, for the MA process in (2.1), *we have*

$$n^{\frac{1}{2}}\left(\hat{\boldsymbol{\psi}}_c - \boldsymbol{\psi}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Sigma}^{-1}\right),$$
$$n^{\frac{1}{2}}\left(\hat{\boldsymbol{\psi}}_J - \boldsymbol{\psi}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Sigma}^{-1}\right),$$

where $\hat{\boldsymbol{\psi}}_c$ *is the conditional MLE defined in* (3.5), $\hat{\boldsymbol{\psi}}_J$ *is the joint MLE in* (3.6), $\boldsymbol{\psi}_0$ *is the true parameter vector and* $\boldsymbol{\Sigma}$ *is the Fisher information matrix associated with the exact likelihood defined at* (1.7) *in* Lii and Rosenblatt (1992).

Assumptions A1−A8 are exactly those used in Breidt, Davis, Lii and Rosenblatt (1991) and Lii and Rosenblatt (1992, 1996). These assumptions guarantee that the information matrix associated with the likelihood has the usual properties. Assumption A9, originally used by Lii and Rosenblatt (1992), is similar to the Lipschitz condition for the smoothness of the first and second derivatives of the log likelihood function. These assumptions are fairly standard and hold for a variety of densities, including mixture of normals, the t distribution with degrees of freedom greater than four, and the exponential power family $f(x) \propto \exp(-c|x/\sigma|^\alpha)$ with $\alpha \geq 2$. We prove Theorem 1 by first showing the asymptotic equivalence between the conditional MLE, joint MLE and the quasi-MLE (defined by Lii and Rosenblatt (1992)) and then using the results in Lii and Rosenblatt (1992). Details about the proof are given in the on-line supplement.

Table 1. Model settings for simulations.

model	type	parameter settings		error distribution
		$\theta^{\dagger}(B)$	$\theta^{*}(B)$	$f_{\sigma}(z)$
MA(1)	invertible	$1+0.5B$		Laplace or $t(4)$
	invertible	$1+0.8B$		Laplace
	non-invertible		$1+2B$	Laplace or $t(4)$
	non-invertible		$1+1.25B$	Laplace
MA(2)	invertible	$(1-0.5B)(1-0.8B)$		Laplace
	non-invertible	$1-0.5B$	$1-1.25B$	Laplace
	non-invertible		$(1-2B)(1-1.25B)$	Laplace

4. Simulation Study

To investigate the performance of the proposed estimators in finite samples, a simulation study was conducted as follows. We considered MA(1) and MA(2) processes for both invertible and non-invertible cases. Two innovation distributions were studied:

$$\text{Laplace}: f(x) = (2\sigma)^{-1}\exp(-\frac{|x|}{\sigma}),$$
$$t(4): f(x) = \frac{\Gamma(\frac{5}{2})}{2\sqrt{\pi}\sigma}\left(1+\frac{1}{4}\left(\frac{x}{\sigma}\right)^2\right)^{-\frac{5}{2}}.$$

For MA(1), we considered two invertible cases ($s=0, r=1$) with parameter values $\theta_1^{\dagger} = 0.5, 0.8$ and two non-invertible cases ($s=1, r=0$) with $\theta_1^{*} = 1.25, 2$. For MA(2), we considered three situations including purely invertible ($s=0, r=2$), non-purely non-invertible ($s=r=1$) and purely non-invertible ($s=2, r=0$) cases. The parameter settings are given in Table 1.

For each process, 500 realizations with sample sizes $n=50$ and $n=100$ were generated. For each realization, four likelihood-based estimators were computed, including the exact MLE by EM, the conditional MLE, the joint MLE and the LR MLE. For implementing the LR MLE, we set $q_{trunc}=10$ as the truncation number for both sample sizes, which is the same as Lii and Rosenblatt (1992) in their simulations. Consequently, the effective data length ($n-2q_{trunc}$) becomes 30 for the cases with $n=50$, and 80 for the cases with $n=100$. As expected, such truncation gains computational convenience in the implementation but loses information in the inference. For comparison, we also report the estimation performance for the MA coefficients using a cumulant-based method which maximizes the absolute residual kurtosis (see Breidt, Davis and Trindade (2001) at (4.17), or Rosenblatt (2000, Sec. 8.7)).

The estimation performance of various estimators for MA(1) processes with Laplace innovations is summarized in Tables 2−3 in terms of biases, standard

Table 2. Biases, standard errors, root mean square errors, and the proportion of correctly-identified invertibility structure (PROP) for various estimators under an invertible MA(1) process $X_t = (1 + \theta_1^\dagger B)Z_t$ with Laplace innovations (500 replicates).

(a) $\theta_1^\dagger = 0.5$

		$n = 50$			$n = 100$		
Estimator		$\theta_1^\dagger = 0.5$	$\sigma = 1$	PROP	$\theta_1^\dagger = 0.5$	$\sigma = 1$	PROP
LR MLE	bias	0.0002	-0.0143	0.494	-0.0007	0.0000	0.910
	s.e.	0.1527	0.1809		0.0845	0.1059	
	rmse	0.1527	0.1814		0.0845	0.1059	
MLE by EM	bias	0.0050	-0.0125	0.922	0.0006	-0.0004	0.966
	s.e.	0.1074	0.1358		0.0722	0.0920	
	rmse	0.1075	0.1364		0.0722	0.0920	
Joint MLE	bias	0.0030	-0.0279	0.786	0.0007	-0.0082	0.936
	s.e.	0.1116	0.1335		0.0727	0.0913	
	rmse	0.1117	0.1364		0.0727	0.0916	
Cond. MLE	bias	0.0010	-0.0272	0.696	0.0004	-0.0079	0.928
	s.e.	0.1060	0.1341		0.0704	0.0914	
	rmse	0.1060	0.1368		0.0704	0.0917	
ASD		0.1225	0.1414		0.0866	0.1000	
cumulant-based	bias	0.2465		0.596	0.1377		0.742
method	s.e.	0.3057			0.2811		
	rmse	0.3927			0.3130		

(b) $\theta_1^\dagger = 0.8$

		$n = 50$			$n = 100$		
Estimator		$\theta_1^\dagger = 0.8$	$\sigma = 1$	PROP	$\theta_1^\dagger = 0.8$	$\sigma = 1$	PROP
LR MLE	bias	0.0096	-0.0226	0.728	0.0028	-0.0151	0.928
	s.e.	0.0911	0.1746		0.0634	0.1086	
	rmse	0.0916	0.1760		0.0635	0.1097	
MLE by EM	bias	0.0047	-0.0220	0.838	0.0057	-0.0167	0.960
	s.e.	0.0805	0.1406		0.0551	0.0972	
	rmse	0.0806	0.1423		0.0554	0.0986	
Joint MLE	bias	0.0214	-0.0316	0.824	0.0082	-0.0228	0.958
	s.e.	0.0844	0.1375		0.0562	0.0965	
	rmse	0.0871	0.1411		0.0568	0.0992	
Cond. MLE	bias	-0.0079	-0.0206	0.800	-0.0078	-0.0152	0.924
	s.e.	0.0807	0.1423		0.0556	0.0969	
	rmse	0.0811	0.1437		0.0562	0.0981	
ASD		0.0849	0.1414		0.0600	0.1000	
cumulant-based	bias	0.0263		0.556	0.0431		0.702
method	s.e.	0.3080			0.2241		
	rmse	0.3092			0.2282		

Table 3. Biases, standard errors, root mean square errors, and the proportion of correctly-identified invertibility structure (PROP) for various estimators under a non-invertible MA(1) process $X_t = (1 + \theta_1^* B)Z_t$ with Laplace innovations (500 replicates).

(a) $\theta_1^* = 2$

		$n = 50$			$n = 100$		
Estimator		$\theta_1^* = 2$	$\sigma = 1$	PROP	$\theta_1^* = 2$	$\sigma = 1$	PROP
LR MLE	bias	0.2696	-0.0450	0.814	0.0983	-0.0361	0.980
	s.e.	0.9960	0.3122		0.3979	0.1926	
	rmse	1.0319	0.3154		0.4098	0.1960	
MLE by EM	bias	0.0331	0.0429	0.720	-0.0163	0.0469	0.820
	s.e.	0.4919	0.2552		0.2517	0.1746	
	rmse	0.4930	0.2587		0.2522	0.1807	
Joint MLE	bias	0.2149	-0.0639	0.866	0.0772	-0.0334	0.988
	s.e.	0.7184	0.2602		0.3218	0.1696	
	rmse	0.7499	0.2679		0.3309	0.1729	
Cond. MLE	bias	0.2566	-0.0674	0.864	0.0785	-0.0338	0.990
	s.e.	0.8857	0.2652		0.3213	0.1706	
	rmse	0.9221	0.2737		0.3307	0.1739	
ASD		0.4899	0.2828		0.3464	0.2000	
cumulant-based method	bias	0.0294		0.604	0.1491		0.756
	s.e.	1.6005			1.3662		
	rmse	1.6008			1.3743		

(b) $\theta_1^* = 1.25$

		$n = 50$			$n = 100$		
Estimator		$\theta_1^* = 1.25$	$\sigma = 1$	PROP	$\theta_1^* = 1.25$	$\sigma = 1$	PROP
LR MLE	bias	0.0334	-0.0319	0.836	0.0001	-0.0024	0.946
	s.e.	0.2723	0.2133		0.0894	0.1273	
	rmse	0.2744	0.2157		0.0894	0.1273	
MLE by EM	bias	-0.0022	-0.0024	0.878	0.0005	0.0029	0.964
	s.e.	0.1505	0.1689		0.0889	0.1188	
	rmse	0.1505	0.1689		0.0889	0.1188	
Joint MLE	bias	0.0000	-0.0153	0.838	-0.0093	-0.0003	0.970
	s.e.	0.1959	0.1760		0.0855	0.1176	
	rmse	0.1959	0.1767		0.0860	0.1176	
Cond. MLE	bias	0.0419	-0.0332	0.888	0.0176	-0.0132	0.952
	s.e.	0.2103	0.1769		0.0929	0.1186	
	rmse	0.2145	0.1800		0.0946	0.1193	
ASD		0.1326	0.1768		0.0938	0.1250	
cumulant-based method	bias	-0.1044		0.596	-0.0656		0.642
	s.e.	1.0471			0.7376		
	rmse	1.0522			0.7405		

errors and root mean square errors. The asymptotic standard error (ASD) under the likelihood-based methods for each process is also given in the table for comparison. Moreover, the proportion of correctly-identified invertibility structures (i.e., r and s) is recorded for each estimator (denoted as PROP in the tables). For example, in Table 2(a), the underlying process was invertible and there were 247 realizations (out of 500 realizations) with the LR MLE in the invertible region, which leads to PROP $= 247/500 = 0.494$. Based on the results in Table 2 for the invertible MA(1), the MLE by EM, the joint MLE and the conditional MLE performed competitively, and all of them outperformed the LR MLE in terms of root mean square error. Also note that the difference was more significant for the cases with $n = 50$. For determining invertibility, the MLE by EM performed best among the four estimators with very high PROP values. In contrast, the LR MLE had the lowest PROP value, but improved as the sample size increased. As expected, the cumulant-based method was much less efficient than the likelihood-based methods.

Based on the results in Table 3 for the non-invertible MA(1), the MLE by EM outperformed the other estimators with a large difference in terms of the root mean squared errors. The second best were the joint MLE and the conditional MLE, which were fairly competitive with each other. The LR MLE still performed worst among the likelihood-based methods, in particular for small samples and when the root was away from the unit circle. As an example, the relative efficiency of LR MLE with respect to MLE by EM was about 0.48 for $n = 50$ and 0.62 for $n = 100$ with $\theta = 2$. But, in contrast to the invertible cases, the PROP values associated with the LR MLE were much closer to those for the joint MLE and the conditional MLE. The cumulant-based methods perform even worse for non-invertible cases. The estimation results for the MA(1) processes with $t(4)$ innovations are shown in Table 4 and are fairly similar to the previous cases.

The results for MA(2) are summarized in Table 5. For the invertible case, the performances of the four likelihood-based estimators were very similar. For the invertible parameters in the non-purely non-invertible case, the performances of the four estimators were again very similar. But, for the non-invertible parameters, MLE by EM performed best and the conditional MLE performed worst.

In our simulation study, we used the Laplace distribution since the corresponding estimators are analogous to the least absolute deviation estimator, and we used the t distribution to illustrate the situations for heavy-tailed innovations. Strictly speaking, the Laplace distribution does not satisfy the assumption (A2) and the $t(4)$ distribution does not have finite fourth moment. But our results demonstrate that the proposed estimators are still effective even in some cases for which the assumptions used in the asymptotic theory are not completely satisfied.

Table 4. Biases, standard errors, root mean square errors, and the proportion of correctly-identified invertibility structure (PROP) for various estimators under MA(1) process with $t(4)$ innovations (500 replicates).

(a) Invertible MA(1): $X_t = (1 + 0.5B)Z_t$

		$n = 50$			$n = 100$		
Estimator		$\theta_1^\dagger = 0.5$	$\sigma = 1$	PROP	$\theta_1^\dagger = 0.5$	$\sigma = 1$	PROP
LR MLE	bias	0.0668	-0.0235	0.446	0.0107	-0.0123	0.794
	s.e.	0.1170	0.1724		0.0851	0.1023	
	rmse	0.1347	0.1740		0.0857	0.1030	
MLE by EM	bias	0.0237	-0.0114	0.782	0.0070	-0.0063	0.902
	s.e.	0.1221	0.1383		0.0807	0.0939	
	rmse	0.1244	0.1388		0.0810	0.0941	
Joint MLE	bias	0.0431	-0.0166	0.650	0.0106	-0.0144	0.848
	s.e.	0.0993	0.1388		0.0773	0.0944	
	rmse	0.1083	0.1397		0.0780	0.0955	
Cond. MLE	bias	0.0350	-0.0153	0.634	0.0058	-0.0141	0.844
	s.e.	0.0927	0.1387		0.0755	0.0945	
	rmse	0.0991	0.1395		0.0757	0.0955	
ASD		0.1449	0.1323		0.1025	0.0935	
cumulant-based	bias	0.2398		0.572	0.1459		0.696
method	s.e.	0.2961			0.2686		
	rmse	0.3811			0.3056		

(b) Non-invertible MA(1): $X_t = (1 + 2B)Z_t$

		$n = 50$			$n = 100$		
Estimator		$\theta_1^* = 2$	$\sigma = 1$	PROP	$\theta_1^* = 2$	$\sigma = 1$	PROP
LR MLE	bias	0.1379	-0.0091	0.894	0.1259	-0.0224	0.918
	s.e.	0.7350	0.3021		0.5128	0.2112	
	rmse	0.7478	0.3022		0.5280	0.2124	
MLE by EM	bias	0.0015	0.0515	0.714	-0.0319	0.0447	0.738
	s.e.	0.6656	0.2661		0.2997	0.1752	
	rmse	0.6656	0.2710		0.3014	0.1808	
Joint MLE	bias	0.1902	-0.0499	0.898	0.0938	-0.0263	0.954
	s.e.	0.6663	0.2630		0.4158	0.1873	
	rmse	0.6930	0.2677		0.4262	0.1892	
Cond. MLE	bias	0.2218	-0.0629	0.908	0.1132	-0.0340	0.954
	s.e.	0.6575	0.2614		0.4270	0.1859	
	rmse	0.6939	0.2689		0.4417	0.1889	
ASD		0.5797	0.3186		0.4099	0.2253	
cumulant-based	bias	-0.0406		0.636	0.0849		0.822
method	s.e.	1.5431			1.4425		
	rmse	1.5436			1.4450		

Table 5. Biases, standard errors, root mean square errors, and the proportion of correctly-identified invertibility structure (PROP) for various estimators under a MA(2) process with Laplace innovations ($n = 100$, 500 replicates).

(a) invertible: $X_t = (1 - 0.5B)(1 - 0.8B)Z_t$

Estimator		$\theta_1^\dagger = -1.3$	$\theta_2^\dagger = 0.4$	$\sigma = 1$	PROP
LR MLE	bias	-0.0135	0.0128	-0.0176	0.776
	s.e.	0.0772	0.0660	0.1061	
	rmse	0.0784	0.0672	0.1075	
MLE by EM	bias	-0.0149	0.0057	-0.0301	0.798
	s.e.	0.0749	0.0684	0.1165	
	rmse	0.0764	0.0687	0.1203	
Joint MLE	bias	-0.0202	0.0102	-0.0307	0.800
	s.e.	0.0731	0.0657	0.0939	
	rmse	0.0759	0.0665	0.0988	
Cond. MLE	bias	0.0047	-0.0001	-0.0158	0.650
	s.e.	0.0748	0.0596	0.0988	
	rmse	0.0749	0.0596	0.1001	
ASD		0.0917	0.0917	0.1000	
cumulant-based method	bias	-0.0360	0.1439		0.224
	s.e.	0.4031	0.2717		
	rmse	0.4047	0.3075		

(b) non-purely non-invertible: $X_t = (1 - 0.5B)(1 - 1.25B)Z_t$

Estimator		$\theta_1^\dagger = -0.5$	$\theta_1^* = -1.25$	$\sigma = 1$	PROP
LR MLE	bias	-0.0352	-0.0626	-0.0452	0.864
	s.e.	0.1019	0.2378	0.1429	
	rmse	0.1078	0.2459	0.1499	
MLE by EM	bias	-0.0122	-0.0399	-0.0228	0.602
	s.e.	0.0901	0.1198	0.1160	
	rmse	0.0910	0.1263	0.1183	
Joint MLE	bias	-0.0312	-0.0333	-0.0381	0.870
	s.e.	0.0957	0.1594	0.1289	
	rmse	0.1006	0.1628	0.1345	
Cond. MLE	bias	-0.0187	-0.0917	-0.0521	0.750
	s.e.	0.0912	0.1782	0.1282	
	rmse	0.0931	0.2004	0.1384	
ASD		0.1732	0.1875	0.1803	
cumulant-based method	bias	-0.1490	-0.0474		0.468
	s.e.	0.2875	1.4177		
	rmse	0.3238	1.4185		

(c) purely non-invertible: $X_t = (1-2B)(1-1.25B)Z_t$

Estimator		$\theta_1^* = -3.25$	$\theta_2^* = 2.5$	$\sigma = 1$	PROP
LR MLE	bias	-0.2093	0.2276	-0.0289	0.856
	s.e.	0.9141	1.0066	0.2165	
	rmse	0.9378	1.0321	0.2184	
MLE by EM	bias	-0.2906	0.2957	-0.0689	0.948
	s.e.	0.8296	0.9489	0.2204	
	rmse	0.8790	0.9939	0.2309	
Joint MLE	bias	-0.3256	0.3240	-0.0643	0.928
	s.e.	1.0561	1.1363	0.2103	
	rmse	1.1051	1.1816	0.2200	
Cond. MLE	bias	-0.4942	0.5709	-0.0988	0.842
	s.e.	1.3598	1.4938	0.2241	
	rmse	1.4468	1.5992	0.2449	
ASD		0.5387	0.5728	0.2500	
cumulant-based	bias	-0.1124	0.2203		0.346
method	s.e.	1.8020	1.9365		
	rmse	1.8056	1.9490		

5. Discussion

We proposed an algorithm for computing the exact MLE for a non-Gaussian MA process, in which the data are augmented by suitable latent variables so that the likelihood can be expressed without loss of information. Two alternative estimators, the joint MLE and conditional MLE, are suggested that are much easier to compute but still asymptotically equivalent to the MLE. Simulation results suggest that the exact MLE solved by the EM algorithm performs better than other estimators under various non-Gaussian MA processes in finite samples. Moreover, the joint MLE and the conditional MLE outperformed Lii and Rosenblatt's approximate MLE for most of the cases.

In defining the conditional MLE, we set each latent variable equal to its unconditional mean, zero, as a natural and convenient choice. In fact, the latent variables can be set to arbitrary finite values for computing the conditional MLE, since the effect of the latent variables is negligible asymptotically. However, initial values may affect the performance of estimators in finite samples. Empirical evidence has been found in our simulation results that the joint MLE (with the estimated initial values) performs better than the conditional MLE, especially for the non-invertible cases.

Another practical issue is that the innovation density, required by the likelihood-based methods, is usually unknown. One can estimate the MA coefficient by maximizing the Laplace likelihood first to obtain an initial estimate (which corresponds to the least absolute deviation estimate) and then computing the residuals. Based on these residuals, a suitable parametric density can be suggested and the estimates for MA coefficients can be refined. Similar recommendations were also made by Lii and Rosenblatt (1996) for the non-minimum phase ARMA processes.

Acknowledgements

This research was supported by the National Science Council (NSC 93-2118-M-007-004), R.O.C. The authors wish to thank two anonymous referees for their helpful comments.

References

Breidt, F. J., Davis, R. A., Hsu, N.-J. and Rosenblatt, M. (2006). *IMS Lecture Notes–Monograph Series: Time Series and Related Topics* **52**, 1-19.

Breidt, F. J., Davis, R. A., Lii, K.-S. and Rosenblatt, M. (1991). Maximum likelihood estimation for noncausal autoregressive processes. *J. Multivariate Anal.* **36**, 175-198.

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

Breidt, F. J. and Hsu, N.-J. (2005). Best mean square prediction for moving averages. *Statist. Sinica* **15**, 427-446.

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

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

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

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

Levine, R. A. and Casella, G. (2001). Implementations of the Monte Carlo EM algorithm. *J. Comput. Graph. Statist.* **10**, 422-439.

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

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

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

Mendel, J. M. (1991). Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications. *Proc. IEEE* **79**, 278-305.

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*, (Edited by D.F. Findley), 549-564. Academic Press, New York.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. Amer. Statist. Assoc.* **85**, 699-704.

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

Institute of Statistics, National Tsing-Hua University, Hsinchu, Taiwan 30013, R.O.C.

E-mail: njhsu@stat.uthu.edu.tw

Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, U.S.A.

E-mail: jbreidt@stat.colostate.edu

(Received May 2007; accepted January 2008)