Statistica Sinica **15**(2005), 427-446

BEST MEAN SQUARE PREDICTION FOR MOVING AVERAGES

F. Jay Breidt and Nan-Jung Hsu

Colorado State University and National Tsing-Hua University

Abstract: Best mean square prediction for moving average time series models is generally non-linear prediction, even in the invertible case. Gaussian processes are an exception, since best linear prediction is always best mean square prediction. Stable numerical recursions are proposed for computation of residuals and evaluation of unnormalized conditional distributions in invertible or non-invertible moving average models, including those with distinct unit roots. The conditional distributions allow evaluation of the best mean square predictor via computation of a low-dimensional integral. For finite, discrete innovations, the method yields best mean square predictors exactly. For continuous innovations, an importance sampling scheme is proposed for numerical approximation of the best mean square predictor and its prediction mean square error. In numerical experiments, the method accurately computes best mean square predictors for cases with known solutions. The approximate best mean square predictor dominates the best linear predictor for out-of-sample forecasts of monthly US unemployment rates.

Key words and phrases: Discrete time series, importance sampling, 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 (i.i.d.) sequence of random variables with zero mean and finite variance, B is the backshift operator ($B^kY_t = Y_{t-k}$ for $k = 0, \pm1, \pm2, \ldots$), $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$, 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) and Scargle (1981)), design of communication systems (Benveniste, Goursat and Roget (1980)), processing of blurry images (Donoho (1981) and Chien, Yang and Chi (1997)), and modeling of vocal tract filters (Rabiner and Schafer (1978) and 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) and 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 maximum likelihood estimation (Lii and Rosenblatt (1992)).

In contrast to estimation, the problem of prediction for general MA processes has received far less attention. Kanter (1979) provides lower bounds for the best mean square prediction error in general MA processes, but does not provide the predictors. Shepp, Slepian and Wyner (1980) extend Kanter's result and provide explicit formulas for one-step-ahead prediction of MA(1) processes driven by exponential, uniform, or binary noise. Rosenblatt (2000, Chap.5) summarizes these and other results for prediction in minimum and non-minimum phase systems, but notes (p.217) that "little is known about the form of best predictors in mean square except for very special examples".

In Section 2, we consider numerical evaluation of the best mean square predictor for invertible or non-invertible moving average models, including those with distinct unit roots. We first derive stable recursions for computing residuals from a realization of a moving average process, given r initial and s final conditions, where $q = r + s$. (Stability of these recursions is established in a technical appendix.) These residuals are used in computation of unnormalized conditional distributions, which in turn allows evaluation of the best mean square predictor via a q-dimensional integration. For finite, discrete innovations, the integration method yields best mean square predictors exactly. For continuous innovations, an importance sampling scheme is proposed for numerical approximation of the integral. This leads to approximate best mean square predictors and prediction mean square errors.

In Section 3, we conduct numerical experiments using the proposed algorithm. In the discrete case, we compare the best mean square predictor to the best linear predictor for MA(1) and MA(2) with binary innovations. In the continuous case, we evaluate the numerical approximations to the best mean square predictors in a simulation study, and compare to the best linear predictors. For cases with closed-form best mean square predictors (MA(1) with exponential or uniform innovations; Gaussian MA(q)) the importance sampling method accurately reproduces the predictors. The importance sampling method can be tailored to specific innovations distributions for improved numerical performance.

In Section 4, we apply the importance sampling methodology to the computation of best mean square predictors for a non-invertible MA(29) model of monthly changes in seasonally-adjusted US unemployment rates. The approximate best mean square predictors dominate the best linear predictors in out-of-sample forecasts.

2. Main Results

Rewrite (1) as

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

with $\theta^*(z) = 1+\theta_1^* z+\cdots+\theta_s^* z^s \neq 0$ for $|z| > 1$, and $\theta^{\dagger}(z) = 1+\theta_1^{\dagger} z+\cdots+\theta_r^{\dagger} z^r \neq 0$ for $|z| \leq 1$, 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 refer to the case in which $r = 0$ and $s > 0$ as purely non-invertible.

Define $\theta_0 = 1$. We seek the best mean square predictor of X_{n+k} given $X_1, \ldots, X_n$,

$$\mathrm{E}\,[X_{n+k}\,|\,X_1, \ldots, X_n] = \sum_{j=0}^{q} \theta_j \mathrm{E}\,[Z_{n+k-j}\,|\,X_1, \ldots, X_n]. \tag{3}$$

Clearly this predictor is zero for $k > q$. In the case of Gaussian noise, the best mean square predictor is well-known to be a linear function of $X_1, \ldots, X_n$. If $\theta(z)$ is invertible, then the best mean square predictor based on the infinite past $\{\ldots, X_{-1}, X_0, X_1, \ldots, X_n\}$ is a linear predictor (Rosenblatt (2000, pp.83-84)). In general, however, the best mean square predictor is a non-linear function of $X_1, \ldots, X_n$.

We use g to denote a generic probability density function (or probability mass function in the discrete case) for one or more random variables, which can be inferred from the argument(s) of g. We assume $Z_t \sim f(z_t)$, where f is known. Since

$$Z_{n-q} = \frac{X_n - Z_n - \theta_1 Z_{n-1} - \cdots - \theta_{q-1} Z_{n-q+1}}{\theta_q},$$

it suffices for the computation of (3) to consider

$$g(z_{n-q+1}, \ldots, z_n \,|\, x_1, \ldots, x_n) \propto g(z_{n-q+1}, \ldots, z_n, x_1, \ldots, x_n).$$

In general, this distribution is not known, so that the integrals in (3) are intractable.

As the referee has pointed out, the standard change-of-variable approach to this problem is to write

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \\ Z_{n+1-q} \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} A & C \\ 0 & I_{q\times q} \end{pmatrix} \begin{pmatrix} Z_{1-q} \\ \vdots \\ Z_{n-q} \\ Z_{n+1-q} \\ \vdots \\ Z_n \end{pmatrix}, \tag{4}$$

where A consists of the first n columns of the $n \times (n+q)$ matrix

$$\begin{pmatrix} \theta_q & \theta_{q-1} & \cdots & \theta_1 & 1 & 0 & 0 & 0 & \cdots 0 \\ 0 & \theta_q & \theta_{q-1} & \cdots & \theta_1 & 1 & 0 & 0 & \cdots 0 \\ 0 & 0 & \theta_q & \theta_{q-1} & \cdots \theta_1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ \vdots & \vdots & & & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \cdots & \cdots & 0 & \theta_q & \theta_{q-1} & \cdots \theta_1 & 1 \end{pmatrix},$$

C consists of the remaining columns, and $I_{q\times q}$ is the $q \times q$ identity matrix. Then

$$g(z_{n-q+1}, \ldots, z_n, x_1, \ldots, x_n) \propto \prod_{t=-q+1}^{n} f(z_t),$$

where

$$\begin{pmatrix} z_{1-q} \\ \vdots \\ z_{n-q} \end{pmatrix} = A^{-1} \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} - C \begin{pmatrix} z_{n+1-q} \\ \vdots \\ z_n \end{pmatrix} \right\}.$$

Since A is an upper triangular matrix, its inverse has the form

$$A^{-1} = \begin{pmatrix} a_1 & a_2 & \cdots & \cdots & a_n \\ 0 & a_1 & a_2 & \cdots & a_{n-1} \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & a_2 \\ 0 & 0 & \cdots & 0 & a_1 \end{pmatrix},$$

where $\{a_k : k = 1, \ldots, n\}$ can be solved recursively by $AA^{-1} = I_{n\times n}$; that is,

$$\begin{aligned} & a_1\theta_q = 1, \\ & a_2\theta_q + a_1\theta_{q-1} = 0, \\ & a_3\theta_q + a_2\theta_{q-1} + a_1\theta_{q-2} = 0, \\ & \vdots \\ & a_q\theta_q + a_{q-1}\theta_{q-1} + \cdots + a_1\theta_1 = 0, \\ & \sum_{j=0}^{q} \theta_{q-j} a_{k-j} = 0, \quad k = q+1, q+2, \ldots, n, \end{aligned}$$

where $\theta_0 = 1$.

The last equation is equivalent to the following difference equation:

$$\Big(1 + \frac{\theta_{q-1}}{\theta_q} B + \frac{\theta_{q-2}}{\theta_q} B^2 + \cdots + \frac{1}{\theta_q} B^q\Big) a_k = 0.$$

As n increases, the general solution of this difference equation remains bounded for all k if and only if all roots of the polynomial

$$1 + \frac{\theta_{q-1}}{\theta_q} z + \frac{\theta_{q-2}}{\theta_q} z^2 + \cdots + \frac{1}{\theta_q} z^q$$

are outside the unit circle or on the unit circle but distinct. Equivalently, A^{-1} is numerically stable with increasing n if and only if all roots of the polynomial $\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q$ are inside the unit circle or on the unit circle but distinct, which corresponds to the purely non-invertible case.

For illustration, consider the MA(1) process, for which the elements in A^{-1} satisfy

$$a_k = \left(\frac{1}{\theta_1}\right)^k (-1)^{k-1}, \quad k = 1, 2, \ldots, n.$$

In the invertible case, $|\theta_1| < 1$, so that A^{-1} is numerically unstable for large n.

We work around the numerical difficulty of the standard approach by developing a set of forward-backward recursions that exploit the structure of the time series to yield numerically stable computations. Effectively, we split the unstable linear transformation in (4) into a pair of stable linear transformations, expressed in time series operator notation. The first transformation is defined as

$$W_t = \theta^\dagger(B)Z_t, \tag{5}$$

so that $Z_t = W_t - (\theta^\dagger(B) - 1)Z_t$. The second transformation is defined as

$$\begin{aligned} X_t &= \theta^*(B)W_t \\ &= (1 + \theta_1^* B + \cdots + \theta_s^* B^s)W_t \\ &= \theta_s^* \tilde{\theta}(B^{-1})W_{t-s}, \end{aligned} \tag{6}$$

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,

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

(Alternatively, we could have written (5) as a lower-triangular linear transformation from $(Z_{-q+1},\ldots,Z_n)$ to $(Z_{-q+1},\ldots,Z_{-q+r},W_{-s+1},\ldots,W_n)$, and (6) as an upper-triangular linear transformation from the latter vector to $(Z_{-q+1},\ldots, Z_{-q+r},X_1,\ldots,X_n,W_{n-s+1},\ldots,W_n)$.)

The corresponding numerical recursions, given a realization $\boldsymbol{x}_n = (x_1,\ldots,x_n)'$ and arbitrary initial conditions $\boldsymbol{z}_r^{(i)} = (z_{-q+1}^{(i)},\ldots,z_{-q+r}^{(i)})'$ and $\boldsymbol{w}_s^{(i)} = (w_{n-s+1}^{(i)},\ldots, w_n^{(i)})'$, are as follows:

- Starting from $w_n^{(i)},\ldots,w_{n+1-s}^{(i)}$, compute

$$w_{t-s}^{(i)} = \frac{x_t}{\theta_s^*} - \left(\tilde{\theta}(B^{-1}) - 1\right) w_{t-s}^{(i)}, \qquad \text{for } t = n, n-1,\ldots,1. \tag{7}$$

- Starting from $z_{-q+1}^{(i)},\ldots,z_{-q+r}^{(i)}$, compute

$$z_{-s+t}^{(i)} = w_{-s+t}^{(i)} - \left(\theta^\dagger(B) - 1\right) z_{-s+t}^{(i)}, \qquad \text{for } t = 1,2,\ldots,n+s. \tag{8}$$

Basically, the backward recursion (7) is stable since the roots of $\tilde{\theta}(z)$ are either outside the unit circle, or on the unit circle but distinct. The forward recursion (8) is stable since all the roots of $\theta^\dagger(z)$ are outside the unit circle. See the Appendix for proofs of these assertions and for details on the solutions of the systems of difference equations implied by the backward and forward recursions.

From the linear transformations in (5)−(6), we have that

$$\begin{aligned} g(z_{n-q+1},\ldots,z_n,x_1,\ldots,x_n) &\propto g(z_{-q+1},\ldots,z_{-q+r},x_1,\ldots,x_n,w_{n+1-s},\ldots,w_n) \\ &\propto \prod_{t=-q+1}^{n} f(z_t). \end{aligned} \tag{9}$$

Then, for $t = n - q, \ldots, n$, we evaluate $\mathrm{E}\,[Z_t \,|\, x_1, \ldots, x_n]$ via

$$\begin{aligned}
\mathrm{E}\,[Z_t \,|\, x_1, \ldots, x_n] &= \int z_t g(z_{n-q+1}, \ldots, z_n \,|\, x_1, \ldots, x_n)\, d\boldsymbol{z} \\
&= \frac{\int z_t g(z_{n-q+1}, \ldots, z_n, x_1, \ldots, x_n)\, d\boldsymbol{z}}{\int g(z_{n-q+1}, \ldots, z_n, x_1, \ldots, x_n)\, d\boldsymbol{z}} \\
&= \frac{\int z_t g(z_{-q+1}, \ldots, z_{-q+r}, x_1, \ldots, x_n, w_{n-s+1}, \ldots, w_n)\, d\boldsymbol{z}_r\, d\boldsymbol{w}_s}{\int g(z_{-q+1}, \ldots, z_{-q+r}, x_1, \ldots, x_n, w_{n-s+1}, \ldots, w_n)\, d\boldsymbol{z}_r\, d\boldsymbol{w}_s} \\
&= \frac{\int z_t g(\boldsymbol{z}_r, \boldsymbol{x}_n, \boldsymbol{w}_s)\, d\boldsymbol{z}_r\, d\boldsymbol{w}_s}{\int g(\boldsymbol{z}_r, \boldsymbol{x}_n, \boldsymbol{w}_s)\, d\boldsymbol{z}_r\, d\boldsymbol{w}_s} \\
&= \frac{\int z_t \prod_{t=-q+1}^{n} f(z_t)\, d\boldsymbol{z}_r\, d\boldsymbol{w}_s}{\int \prod_{t=-q+1}^{n} f(z_t)\, d\boldsymbol{z}_r\, d\boldsymbol{w}_s},
\end{aligned} \tag{10}$$

where $\boldsymbol{z}_r = (z_{-q+1}, \ldots, z_{-q+r})'$, $\boldsymbol{w}_s = (w_{n-s+1}, \ldots, w_n)'$, and $\{z_t\}_{t=-q+r+1}^{n}$ are computed from (7)−(8).

2.1. Exact best mean square prediction for discrete innovations

If Z_t is a discrete random variable with mass at $k < \infty$ distinct points, then the integrals in (10) can be evaluated exactly. First, use the k^{r+s} possible values of $(z_{n-r-s+1}, \ldots, z_n)'$ to enumerate all possible values of the vector $\boldsymbol{w}_s$. There are at most k^{r+s} possible values. Starting from each possible value, run the backward recursion (7). Next, enumerate the k^r possible values of $\boldsymbol{z}_r$. Using each possible value of $\boldsymbol{z}_r$ together with each possible backward recursion, run the forward recursion (8) to get at most k^{2r+s} possible residual sequences. Evaluate (10) by summing over all possible residual sequences, noting that some of these sequences may have probability zero. Plugging in to (3) then yields the exact best mean square predictor.

2.2. Importance sampling for continuous innovations

In the case of continuous innovations, the q-dimensional integrals in (10) can be evaluated numerically via importance sampling (see, for example Chap.6 of Evans and Swartz (2000), and the references therein) as follows. Let $h(\boldsymbol{z}_r, \boldsymbol{w}_s)$ be a q-dimensional joint density, the importance sampler. Assume the support of h is $\mathrm{supp}(h) \supset \mathrm{supp}\left(\int g(\boldsymbol{z}_r, \boldsymbol{x}_n, \boldsymbol{w}_s)\, d\boldsymbol{x}_n\right)$. For any $(\boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)}) \in \mathrm{supp}(h)$, define the importance weight

$$A(\boldsymbol{z}_r^{(i)}, \boldsymbol{x}_n, \boldsymbol{w}_s^{(i)}) = \frac{\prod_{t=-q+1}^{n} f(z_t^{(i)})}{h(\boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)})},$$

where the $\{z_t^{(i)}\}_{t=-q+r+1,\ldots,n}$ are computed recursively from (7)−(8).

Note that (10) can be written as

$$\mathrm{E}\left[Z_t \,|\, \boldsymbol{x}_n\right] = \frac{\displaystyle\int z_t \frac{\prod_{t=-q+1}^{n} f(z_t)}{h(\boldsymbol{z}_r, \boldsymbol{w}_s)} h(\boldsymbol{z}_r, \boldsymbol{w}_s) d\boldsymbol{z}_r d\boldsymbol{w}_s}{\displaystyle\int \frac{\prod_{t=-q+1}^{n} f(z_t)}{h(\boldsymbol{z}_r, \boldsymbol{w}_s)} h(\boldsymbol{z}_r, \boldsymbol{w}_s)\, d\boldsymbol{z}_r d\boldsymbol{w}_s} \equiv \frac{\mathrm{E}_h\left[Z_t\, A(\boldsymbol{Z}_r, \boldsymbol{x}_n, \boldsymbol{W}_s)\right]}{\mathrm{E}_h\left[A(\boldsymbol{Z}_r, \boldsymbol{x}_n, \boldsymbol{W}_s)\right]}, \quad (11)$$

where the expectation in (11) is taken with respect to $h(\boldsymbol{z}_r, \boldsymbol{w}_s)$. The ratio in (11) can be approximated via Monte Carlo (MC) as

$$\hat{\mathrm{E}}\left[Z_t \,|\, \boldsymbol{x}_n\right] = \frac{\sum_{i=1}^{m} z_t^{(i)} A\left(\boldsymbol{z}_r^{(i)}, \boldsymbol{x}_n, \boldsymbol{w}_s^{(i)}\right)}{\sum_{i=1}^{m} A\left(\boldsymbol{z}_r^{(i)}, \boldsymbol{x}_n, \boldsymbol{w}_s^{(i)}\right)}, \quad (12)$$

where m is the number of draws in the importance sampling, $\{(\boldsymbol{z}_r^{(i)}, \boldsymbol{w}_s^{(i)}); i = 1, \ldots, m\}$ are q-dimensional random vectors drawn from the importance sampler $h(\boldsymbol{z}_r, \boldsymbol{w}_s)$, and $\{z_t^{(i)}\}$ is the corresponding residual sequence computed from $(\boldsymbol{z}_r^{(i)}, \boldsymbol{x}_n, \boldsymbol{w}_s^{(i)})$ according to the backward and forward recursions (7)−(8). By the Strong Law of Large Numbers, this MC estimator converges to $\mathrm{E}\left[Z_t \,|\, \boldsymbol{x}_n\right]$ almost surely as $m \to \infty$. By the Central Limit Theorem, the magnitude of the approximation error has order $m^{-1/2}$ if $\mathrm{Var}_h(Z_t A(\boldsymbol{Z}_r, \boldsymbol{x}_n, \boldsymbol{W}_s))$ and $\mathrm{Var}_h(A(\boldsymbol{Z}_r, \boldsymbol{x}_n, \boldsymbol{W}_s))$ are both finite, where Var_h is taken with respect to $h(\boldsymbol{z}_r, \boldsymbol{w}_s)$.

Consequently, by substitution in (3), the MC estimator of $\mathrm{E}\left[X_{n+k} \,|\, \boldsymbol{x}_n\right]$ is

$$\widehat{\mathrm{BP}} = \hat{\mathrm{E}}\left[X_{n+k} \,|\, \boldsymbol{x}_n\right] = \sum_{j=0}^{q} \theta_j \hat{\mathrm{E}}\left[Z_{n+k-j} \,|\, \boldsymbol{x}_n\right]. \quad (13)$$

Theoretically, the MC estimator is valid for any $h(\boldsymbol{z}_r, \boldsymbol{w}_s)$ satisfying the condition on its support. However, the performance of the MC estimator, i.e., the variability of $\hat{\mathrm{E}}\left[Z_t \,|\, \boldsymbol{x}_n\right]$, depends on the choice of $h(\boldsymbol{z}_r, \boldsymbol{w}_s)$. A bad choice of importance sampler produces a lot of small importance weights and a few extremely large and influential weights so that the variability of the MC estimator is large. A good choice of importance sampler for the particular innovations distribution at hand leads to improved numerical performance. Even with a good choice of importance sampler, variance reduction methods can be employed to make the importance sampling algorithm still more accurate. Methods for choosing and evaluating importance samplers and for reducing variance in importance sampling are thoroughly reviewed in Evans and Swartz (2000, Chap.6).

The prediction MSE of $\mathrm{E}\left[X_{n+k} \,|\, \boldsymbol{x}_n\right]$ can also be approximated using the importance sampler. First, for $1 \le k \le q$, X_{n+k} can be represented as

$$X_{n+k} = \theta(B) Z_{n+k} = \sum_{j=0}^{q} \theta_j Z_{n+k-j} = \boldsymbol{\theta}_{k1}' \boldsymbol{Z}_{n1} + \boldsymbol{\theta}_{k2}' \boldsymbol{Z}_{n2},$$

where $\boldsymbol{\theta}_{k1} = (\theta_k, \theta_{k+1}, \ldots, \theta_q, 0, \ldots, 0)'$, $\boldsymbol{\theta}_{k2} = (0, \ldots, 0, \theta_0, \ldots, \theta_{k-1})'$, $\boldsymbol{Z}_{n1} = (Z_n, Z_{n-1}, \ldots, Z_{n-q+1})'$, $\boldsymbol{Z}_{n2} = (Z_{n+q}, Z_{n+q-1}, \ldots, Z_{n+1})'$. Therefore, the conditional prediction MSE of $\mathrm{E}\,[X_{n+k}\,|\,\boldsymbol{x}_n]$ satisfies

$$\begin{aligned}
&\mathrm{E}\left[(X_{n+k} - E(X_{n+k}|\boldsymbol{x}_n))^2\,|\boldsymbol{x}_n\right] \\
&= \mathrm{E}\left[\left(\boldsymbol{\theta}'_{k1}\left(\boldsymbol{Z}_{n1} - \mathrm{E}(\boldsymbol{Z}_{n1}|\boldsymbol{x}_n)\right) + \boldsymbol{\theta}'_{k2}\boldsymbol{Z}_{n2}\right)^2\,|\boldsymbol{x}_n\right] \\
&= \boldsymbol{\theta}'_{k1}\mathrm{E}\left[(\boldsymbol{Z}_{n1} - E(\boldsymbol{Z}_{n1}|\boldsymbol{x}_n))^2\right]\boldsymbol{\theta}_{k1} + \boldsymbol{\theta}'_{k2}\left[\mathrm{Var}\left(\boldsymbol{Z}_{n2}\right)\right]\boldsymbol{\theta}_{k2} \\
&= \boldsymbol{\theta}'_{k1}\left\{\mathrm{E}\left[\boldsymbol{Z}_{n1}\boldsymbol{Z}'_{n1}|\boldsymbol{x}_n\right] - \mathrm{E}\left[\boldsymbol{Z}_{n1}|\boldsymbol{x}_n\right]\mathrm{E}\left[\boldsymbol{Z}'_{n1}|\boldsymbol{x}_n\right]\right\}\boldsymbol{\theta}_{k1} + \left(\boldsymbol{\theta}'_{k2}\boldsymbol{\theta}_{k2}\right)\,\mathrm{Var}(Z_t).
\end{aligned}$$

This quantity can be approximated by a MC estimator using the same importance sampling draws used for $\widehat{\mathrm{BP}}$:

$$\begin{aligned}
&\hat{\mathrm{E}}\left[(X_{n+k} - \mathrm{E}(X_{n+k}|\boldsymbol{x}_n))^2\,|\boldsymbol{x}_n\right] \\
&= \boldsymbol{\theta}'_{k1}\left\{\hat{\mathrm{E}}\left[\boldsymbol{Z}_{n1}\boldsymbol{Z}'_{n1}|\boldsymbol{x}_n\right] - \hat{\mathrm{E}}\left[\boldsymbol{Z}_{n1}|\boldsymbol{x}_n\right]\hat{\mathrm{E}}\left[\boldsymbol{Z}'_{n1}|\boldsymbol{x}_n\right]\right\}\boldsymbol{\theta}_{k1} + \left(\boldsymbol{\theta}'_{k2}\boldsymbol{\theta}_{k2}\right)\,\mathrm{Var}(Z_t),
\end{aligned}$$

where $\hat{\mathrm{E}}(\boldsymbol{Z}_{n1}|\boldsymbol{x}_n)$ is the vector version of (12) and

$$\hat{\mathrm{E}}\left[\boldsymbol{Z}_{n1}\boldsymbol{Z}'_{n1}\,|\,\boldsymbol{x}_n\right] = \frac{\sum_{i=1}^m \boldsymbol{z}_{n1}^{(i)}\left(\boldsymbol{z}_{n1}^{(i)}\right)' A\left(\boldsymbol{z}_r^{(i)}, \boldsymbol{x}_n, \boldsymbol{w}_s^{(i)}\right)}{\sum_{i=1}^m A\left(\boldsymbol{z}_r^{(i)}, \boldsymbol{x}_n, \boldsymbol{w}_s^{(i)}\right)},$$

in which $\boldsymbol{z}_{n1}^{(i)} = (z_n^{(i)}, z_{n-1}^{(i)}, \ldots, z_{n-q+1}^{(i)})'$.

3. Numerical Experiments

The performance of our methodology for computing or approximating the best mean square predictor (BP) is investigated numerically for several MA processes in the following subsections. We consider discrete innovations in Section 3.1 and continuous (Gaussian or non-Gaussian) innovations in Section 3.2. Both invertible and non-invertible cases are included. In what follows, the best linear predictor (BLP) has prediction MSE σ^2_{BLP}, while the best predictor (BP) has prediction MSE σ^2_{BP}. The prediction MSE of the BLP is evaluated analytically for all innovations distributions. The prediction MSE of the BP is evaluated analytically for discrete innovations, and via simulation for continuous innovations.

3.1. Discrete innovations

We consider the binary innovations distribution with $P(Z_t = 1) = P(Z_t = -1) = 0.5$. For this case, the best MS predictor can be evaluated exactly using our backward-forward algorithm as noted in Section 2.1. The performance of the BP is judged in terms of its efficiency relative to the BLP. We consider both invertible and non-invertible MA(1), as well as non-invertible MA(2) processes.

For an invertible MA(1), the relative efficiency of BP to BLP converges to one as $n \to \infty$, since the BP based on the infinite past is linear. This convergence is evident even for $n = 10$ in the binary case. For $\theta = 0.9$, the relative efficiency of BP to BLP decreases smoothly from 1.362 at $n = 1$ to 1.021 at $n = 10$. For $\theta = 0.5$ the decay is even faster, from 1.050 at $n = 1$ to 1.000 at $n = 10$. This is not surprising, since the coefficients in the corresponding AR(∞) representation decay much faster for the case with smaller $|\theta|$.

For non-invertible MA(1) processes with $\theta^\dagger(z) = 1$ and $\theta^*(z) = (1 + \theta z)$, BP can be much more efficient than BLP as θ^{-1} approaches zero. This is shown in Table 1 both at $n = 1$ and $n = 10$. Indeed, it can be shown that the relative efficiency approaches θ^2 as $n \to \infty$.

Table 1. Prediction mean square errors of the best mean square predictor (BP) and the best linear predictor (BLP) and relative efficiency of BP to BLP in non-invertible MA(1) processes with binary innovations.

	$n = 1$			$n = 10$		
θ^{-1}	σ^2_{BLP}	σ^2_{BP}	$\sigma^2_{\text{BLP}}\sigma^{-2}_{\text{BP}}$	σ^2_{BLP}	σ^2_{BP}	$\sigma^2_{\text{BLP}}\sigma^{-2}_{\text{BP}}$
1.0	1.500	1.500	1.000	1.091	1.001	1.090
0.9	1.682	1.000	1.682	1.260	1.000	1.260
0.7	2.370	1.000	2.370	2.041	1.000	2.041
0.5	4.200	1.000	4.200	4.000	1.000	4.000
0.3	11.194	1.000	11.194	11.111	1.000	11.111
0.1	100.010	1.000	100.010	100.000	1.000	100.000

We consider two classes of non-invertible MA(2) processes:

$$X_t = (1 + \theta^{-1}B)(1 + (1 - \theta)^{-1}B)Z_t, \tag{14}$$

$$X_t = (1 + \theta^{-1}B)(1 + \theta B)Z_t, \tag{15}$$

where $\theta = 0.1, 0.3, \ldots, 0.9$. The former is purely non-invertible and the latter is not. Note that, in (14), the cases with $\theta = a$ and $\theta = 1 - a$ are identical, therefore the results for $\theta = 0.1, 0.3$ are omitted below.

In Table 2, the BP is much more efficient than the BLP for one-step-ahead prediction in nearly all cases considered, with greater efficiency gains at the larger sample size. Large efficiency gains are also evident for two-step-ahead prediction in the purely non-invertible models. For the non-purely non-invertible model, on the other hand, the BLP is fairly competitive with the BP for two-step-ahead prediction.

3.2. Continuous innovations

For the continuous innovations case, the performance of the MC estimator (denoted by $\widehat{\text{BP}}$) for approximating BP is judged in terms of its bias, root mean square error and efficiency relative to both BP and the BLP:

$$\begin{aligned}
\text{BIAS} &= \mathrm{E}\left[\hat{\mathrm{E}}[X_{n+k} \,|\, \boldsymbol{X}_n] - \mathrm{E}[X_{n+k} \,|\, \boldsymbol{X}_n]\right], \\
\text{MSE} &= \mathrm{E}\left\{\hat{\mathrm{E}}[X_{n+k} \,|\, \boldsymbol{X}_n] - \mathrm{E}[X_{n+k} \,|\, \boldsymbol{X}_n]\right\}^2, \\
\text{RMSE} &= \{\text{MSE}\}^{1/2}, \\
\text{RE}(\widehat{\text{BP}}, \text{BP}) &= \frac{\mathrm{E}\left\{X_{n+k} - \mathrm{E}[X_{n+k} \,|\, \boldsymbol{X}_n]\right\}^2}{\mathrm{E}\left\{X_{n+k} - \hat{\mathrm{E}}[X_{n+k} \,|\, \boldsymbol{X}_n]\right\}^2} \equiv \frac{\sigma^2_{\text{BP}}}{\sigma^2_{\text{BP}} + \text{MSE}},
\end{aligned}$$

Table 2. Prediction mean square errors of the k-step BP and BLP ($k = 1, 2$), and relative efficiency of BP to BLP in non-invertible MA(2) processes with binary innovations.

purely non-invertible MA(2): $\theta(z) = (1 + \theta^{-1}z)(1 + (1 - \theta)^{-1}z)$							
		$k = 1$			$k = 2$		
	θ	σ^2_{BLP}	σ^2_{BP}	$\sigma^2_{\text{BLP}}\sigma^{-2}_{\text{BP}}$	σ^2_{BLP}	σ^2_{BP}	$\sigma^2_{\text{BLP}}\sigma^{-2}_{\text{BP}}$
$n = 1$	0.9	174.870	62.728	2.788	247.416	124.457	1.988
	0.7	30.110	12.338	2.440	45.862	23.676	1.937
	0.5	20.879	9.000	2.320	32.515	17.000	1.913
$n = 10$	0.9	126.156	1.000	126.156	246.941	124.457	1.984
	0.7	22.684	1.000	22.684	45.352	23.676	1.916
	0.5	16.000	1.000	16.000	32.000	17.000	1.882

non-purely non-invertible MA(2): $\theta(z) = (1 + \theta^{-1}z)(1 + \theta z)$							
		$k = 1$			$k = 2$		
	θ	σ^2_{BLP}	σ^2_{BP}	$\sigma^2_{\text{BLP}}\sigma^{-2}_{\text{BP}}$	σ^2_{BLP}	σ^2_{BP}	$\sigma^2_{\text{BLP}}\sigma^{-2}_{\text{BP}}$
$n = 1$	0.9	3.368	3.022	1.114	5.879	5.545	1.060
	0.7	3.756	3.265	1.150	6.378	6.031	1.058
	0.5	5.220	4.125	1.265	8.129	7.750	1.049
	0.3	11.727	7.601	1.543	15.135	14.701	1.030
	0.1	100.087	52.005	1.925	104.000	103.510	1.005
$n = 10$	0.9	1.431	1.000	1.431	5.363	5.045	1.063
	0.7	2.054	1.000	2.054	6.046	5.531	1.093
	0.5	4.000	1.000	4.000	8.000	7.250	1.103
	0.3	11.111	1.000	11.111	15.111	14.201	1.064
	0.1	100.000	1.000	100.000	104.000	103.010	1.010

$$\text{RE}(\widehat{\text{BP}}, \text{BLP}) = \frac{\mathrm{E}\left\{X_{n+k} - \text{BLP}\right\}^2}{\mathrm{E}\left\{X_{n+k} - \hat{\mathrm{E}}[X_{n+k} \,|\, \boldsymbol{X}_n]\right\}^2} \equiv \frac{\sigma^2_{\text{BLP}}}{\sigma^2_{\text{BP}} + \text{MSE}}.$$

These quantities can be evaluated numerically by

$$\text{bias} = \frac{1}{R}\sum_{i=1}^{R}\left(\hat{\mathrm{E}}\left[X_{n+k}\,|\,\boldsymbol{x}_n^{(i)}\right] - \mathrm{E}\left[X_{n+k}\,|\,\boldsymbol{x}_n^{(i)}\right]\right), \tag{16}$$

$$\text{mse} = \frac{1}{R}\sum_{i=1}^{R}\left(\hat{\mathrm{E}}\left[X_{n+k}\,|\,\boldsymbol{x}_n^{(i)}\right] - \mathrm{E}\left[X_{n+k}\,|\,\boldsymbol{x}_n^{(i)}\right]\right)^2,$$

$$\text{rmse} = \{\text{mse}\}^{1/2}, \tag{17}$$

$$\text{re}(\widehat{\text{BP}}, \text{BP}) = \frac{\sigma^2_{\text{BP}}}{\sigma^2_{\text{BP}} + \text{mse}}, \tag{18}$$

$$\text{re}(\widehat{\text{BP}}, \text{BLP}) = \frac{\sigma^2_{\text{BLP}}}{\sigma^2_{\text{BP}} + \text{mse}}, \tag{19}$$

where R is the number of replications and $\boldsymbol{x}_n^{(i)}$ is the realization of $\boldsymbol{X}_n$ in the ith replication.

In our importance sampling experiments, the importance sampler is set to be $h(\boldsymbol{z}_r, \boldsymbol{w}_s) = \prod h_z(z_t) \prod h_w(w_t)$, in which $h_z(z)$ and $h_w(w)$ are uniform densities when the domain of $f(z)$ is bounded and Gaussian densities (or truncated Gaussian densities) otherwise. These are fairly naive importance samplers. They have not been tailored for optimality with a particular innovations distribution. The performance of $\widehat{\text{BP}}$ could be improved through a better choice of importance sampler, a larger value of m, or both.

Section 3.2.1 considers invertible MA processes with Gaussian and non-Gaussian innovations, checking equivalence of the approximated BP and the BLP in the Gaussian case, and the convergence of the approximated BP to the infinite-past BLP in the non-Gaussian case. Section 3.2.2 compares the approximated BP to the known BP for certain non-invertible non-Gaussian MA(1) processes. Finally, Section 3.2.3 compares the approximated BP to the known BP=BLP for Gaussian MA(2) processes.

3.2.1. Invertible MA(1) with Gaussian and non-Gaussian innovations

We consider invertible MA(1) processes with $\theta^\dagger(z) = 1+\theta z$ for both Gaussian (GAUSS) and non-Gaussian innovations, all centered and scaled to zero mean and unit variance. Two continuous, non-Gaussian innovations distributions are considered: the centered exponential (EXP) with probability density function $e^{-(z+1)}1_{\{z>-1\}}$ and the uniform (UNIF) on $[-1/\sqrt{3}, 1/\sqrt{3}]$.

Table 3 presents the empirical values of relative efficiency for $n = 1, \ldots, 10$ under each innovations distribution evaluated when R, the number of replications, is set to be 100 for each process and m, the number of draws in the importance sampling, is set to be 4,000 for each replication.

Table 3. The relative efficiency of $\widehat{\text{BP}}$ to BLP, $\text{re}(\widehat{\text{BP}}, \text{BLP})$, for invertible MA(1) processes with continuous innovations.

$\theta = 0.5$			
n	GAUSS	UNIF	EXP
1	1.000	1.003	1.006
2	1.000	1.001	1.007
3	1.000	1.001	1.002
4	1.000	1.000	1.000
5	1.000	1.000	1.000
6	1.000	1.000	1.000
7	1.000	1.000	1.000
8	1.000	1.000	1.000
9	1.000	1.000	1.000
10	1.000	1.000	1.000

$\theta = 0.9$			
n	GAUSS	UNIF	EXP
1	1.000	1.001	1.009
2	1.000	1.019	1.088
3	1.000	1.026	1.060
4	1.000	1.027	1.055
5	1.000	1.026	1.053
6	1.000	1.021	1.036
7	1.000	1.020	1.033
8	1.000	1.017	1.025
9	1.000	1.013	1.021
10	1.000	1.011	1.016

The MC estimator correctly reproduces the BLP in the Gaussian case, as $\text{re}(\widehat{\text{BP}}, \text{BLP}) = 1$ to three decimal places. As in the binary case, the relative efficiency of BP to BLP converges to one as $n \to \infty$, since the BP based on the infinite past is linear. The convergence is slower for larger $|\theta|$.

3.2.2. Non-invertible MA(1) with non-Gaussian innovations

We next consider non-invertible MA(1) processes satisfying (2) with $\theta^{\dagger}(z) = 1$ and $\theta^{*}(z) = 1 + \theta z$ where $\theta^{-1} = \pm 0.1, \pm 0.3, \ldots, \pm 0.9$. As above, we consider EXP and UNIF as the non-Gaussian innovations distributions. Results on bias, rmse and relative efficiency of $\widehat{\text{BP}}$ are reported in Table 4 and 5 for $n = 1$ and $n = 10$, based on $R = 1,000$ and $m = 4,000$. The true values of BP and the corresponding prediction mean square error are derived in Shepp, Slepian and Wyner ((1980); noting that the right-hand expression in (17) of that paper should be negated, and that each δ in an exponent in (28) of that paper should be replaced by δ^{-1}).

Our results show that the MC predictors do deviate from the true BP's, with greater empirical bias and rmse as $|\theta^{-1}|$ approaches zero for both innovations distributions. However, the values of $\text{re}(\widehat{\text{BP}}, \text{BP})$ exceed 0.998 in Table 4 and equal 1.000 in Table 5, indicating our algorithm works extremely well for finding BP for these non-Gaussian, non-invertible MA processes. Moreover, the MC predictor under $n = 10$ performs better than that under $n = 1$ in terms of smaller bias and rmse. The small efficiency losses are due to sampling errors; better performance could be achieved by increasing the number of draws in the importance sampling or choosing a better importance sampler.

Table 4. Bias, rmse and relative efficiency of MC predictor $\widehat{\mathrm{BP}} = \hat{E}[X_{n+1}|\boldsymbol{X}_n]$ for approximating the best one-step MS predictor in a non-invertible MA(1) with exponential innovations, centered to zero mean and scaled to unit variance.

		$n = 1$		
θ^{-1}	bias	rmse	re($\widehat{\mathrm{BP}}$,BP)	re($\widehat{\mathrm{BP}}$,BLP)
0.9	-0.012	0.045	0.999	0.999
0.7	-0.022	0.065	0.998	1.009
0.5	-0.028	0.069	0.999	1.037
0.3	-0.052	0.102	0.999	1.051
0.1	-0.162	0.282	0.999	1.036
-0.1	0.139	0.309	0.999	1.158
-0.3	0.019	0.090	0.999	1.515
-0.5	0.011	0.051	0.999	1.512
-0.7	0.006	0.034	0.999	1.409
-0.9	0.002	0.024	1.000	1.288

		$n = 10$		
θ^{-1}	bias	rmse	re($\widehat{\mathrm{BP}}$,BP)	re($\widehat{\mathrm{BP}}$,BLP)
0.9	-0.005	0.004	1.000	1.170
0.7	-0.003	0.015	1.000	1.499
0.5	-0.011	0.034	1.000	1.431
0.3	-0.037	0.081	0.999	1.221
0.1	-0.160	0.287	0.999	1.057
-0.1	0.112	0.307	0.999	1.288
-0.3	0.020	0.087	0.999	1.822
-0.5	0.002	0.040	0.999	2.104
-0.7	-0.001	0.023	1.000	1.596
-0.9	-0.001	0.012	1.000	1.228

As a benchmark, the relative efficiency of $\widehat{\mathrm{BP}}$ to BLP is evaluated and reported in the tables. The efficiency gain is large for the exponential innovations, especially for the cases with negative θ. Under the uniform innovations, the efficiency gain from using BP instead of BLP is up to 20%. These empirical results are consistent with those presented in Figures 2, 5, 6 of Shepp, Slepian and Wyner (1980).

3.2.3. Non-invertible MA(2) with Gaussian innovations

In this section, we study the same non-invertible MA(2) processes ((14)–(15)) used in the case of binary innovations. All innovations $\{Z_t\}$ are Gaussian with mean zero and variance one. We consider the Gaussian case because the form of the best mean square predictors for MA(2) is known, which is not the case for other continuous distributions.

Table 5. Bias, rmse and relative efficiency of MC predictor $\widehat{\mathrm{BP}} = \hat{E}[X_{n+1}|\boldsymbol{X}_n]$ for approximating the best one-step MS predictor in a non-invertible MA(1) with uniform innovations, centered to zero mean and scaled to unit variance.

$n = 1$				
θ^{-1}	bias	rmse	re($\widehat{\mathrm{BP}}$,BP)	re($\widehat{\mathrm{BP}}$,BLP)
0.9	0.000	0.013	1.000	1.002
0.7	-0.001	0.019	1.000	1.019
0.5	0.000	0.027	1.000	1.049
0.3	0.000	0.047	1.000	1.071
0.1	-0.005	0.153	1.000	1.039
-0.1	0.006	0.154	1.000	1.043
-0.3	0.001	0.048	1.000	1.074
-0.5	-0.001	0.028	1.000	1.053
-0.7	0.000	0.018	1.000	1.018
-0.9	0.000	0.013	1.000	1.002

$n = 10$				
θ^{-1}	bias	rmse	re($\widehat{\mathrm{BP}}$,BP)	re($\widehat{\mathrm{BP}}$,BLP)
0.9	0.000	0.006	1.000	1.128
0.7	0.000	0.013	1.000	1.234
0.5	0.000	0.025	1.000	1.189
0.3	0.001	0.046	1.000	1.137
0.1	0.006	0.156	1.000	1.043
-0.1	-0.005	0.158	1.000	1.050
-0.3	-0.001	0.047	1.000	1.134
-0.5	0.000	0.024	1.000	1.214
-0.7	-0.001	0.013	1.000	1.229
-0.9	0.000	0.005	1.000	1.132

The performance of the MC estimator for approximating the k-step ahead BP for MA(2) ($k = 1, 2$) is summarized in Table 6. For all of the cases, our importance sampling algorithm accurately reproduces the BP=BLP, since re($\widehat{\mathrm{BP}}$,BLP) ≈ 1.

4. Application

Huang and Pawitan (2000) have fitted the following non-invertible SARIMA $(0, 1, 5) \times (0, 0, 2)_{12}$ to seasonally-adjusted, monthly US unemployment rates $\{Y_t\}_{t=0}^{598}$ for the $n + 1 = 598$ months from January 1948 to October 1997. These data plus additional months through the present are available online at http://data.bls.gov/cgi-bin/surveymost?bls. The fitted model is

$$X_t = (1 - B)Y_t = \left(1 + \sum_{i=1}^{5} \theta_i^{\dagger} B^i\right)(1 + \theta_{12}^{*} B^{12} + \theta_{24}^{*} B^{24})Z_t,$$

Table 6. Bias, rmse and relative efficiency of MC predictor $\widehat{\mathrm{BP}} = \hat{E}[X_{n+k}|\boldsymbol{X}_n]$ for approximating the best k-step MS predictor ($k = 1, 2$) in a non-invertible MA(2) with Gaussian innovations.

purely non-invertible MA(2): $\theta(z) = (1 + \theta^{-1}z)(1 + (1 - \theta)^{-1}z)$

		$k = 1$			$k = 2$		
	θ	bias	rmse	re($\widehat{\mathrm{BP}}$,BP)	bias	rmse	re($\widehat{\mathrm{BP}}$,BP)
$n = 1$	0.9	0.066	0.289	1.000	0.047	0.246	1.000
	0.7	0.020	0.110	1.000	0.012	0.101	1.000
	0.5	0.000	0.097	1.000	-0.000	0.091	1.000
$n = 10$	0.9	-0.020	0.291	0.999	-0.005	0.307	1.000
	0.7	0.007	0.120	0.999	0.006	0.120	1.000
	0.5	-0.006	0.102	0.999	-0.011	0.099	1.000

non-purely non-invertible MA(2): $\theta(z) = (1 + \theta^{-1}z)(1 + \theta z)$

		$k = 1$			$k = 2$		
	θ	bias	rmse	re($\widehat{\mathrm{BP}}$,BP)	bias	rmse	re($\widehat{\mathrm{BP}}$,BP)
$n = 1$	0.9	-0.013	0.100	0.997	-0.009	0.062	0.999
	0.7	-0.007	0.096	0.998	-0.005	0.055	1.000
	0.5	-0.006	0.092	0.998	-0.003	0.050	1.000
	0.3	0.002	0.106	0.999	-0.001	0.048	1.000
	0.1	0.001	0.161	1.000	-0.001	0.045	1.000
$n = 10$	0.9	0.006	0.056	0.998	0.003	0.050	1.000
	0.7	-0.001	0.087	0.996	-0.001	0.062	0.999
	0.5	0.010	0.113	0.997	0.005	0.065	0.999
	0.3	0.008	0.114	0.999	0.004	0.057	1.000
	0.1	0.010	0.188	1.000	0.003	0.052	1.000

where $(\theta_1^\dagger, \ldots, \theta_5^\dagger) = (-0.0163, 0.1844, 0.1329, 0.1235, 0.1834)$, $(\theta_{12}^*, \theta_{24}^*) = (1.1832, -4.415)$ and $\mathrm{Var}(Z_t) = (0.0483661)^2$. The non-seasonal MA(5) is purely invertible and the seasonal MA is purely non-invertible, so $r = 5$, $s = 24$, and $q = 29$.

We compute residuals from this model and assume that the normalized residuals $\{Z_t/\sqrt{\mathrm{Var}(Z_t)}\}$ are independently and identically distributed as $t_\nu\sqrt{(\nu-2)/\nu}$, which is a t-distribution with ν degrees of freedom, scaled to unit variance. Fitting ν via maximum likelihood, we get $\hat{\nu}_{\text{mle}} = 4.63$.

Treating the parameter estimates as fixed, we then generate multi-step out-of-sample forecasts for the monthly differences $X_{598+k} = Y_{598+k} - Y_{598+k-1}$ ($k = 1, 2, \ldots, 29$) and compare to the corresponding observed values. (Out-of-sample forecasts for $k > 29$ are identically zero.) The approximate best mean square predictors $\{\widehat{\mathrm{BP}}_k\}_{k=1,2,\ldots,29}$ are computed with our importance sampling algorithm, using a rescaled $t_{4.63}$ importance sampler. One million draws are obtained from this distribution, and then a subsample of these draws of size $m = 100,000$ is selected with probabilities proportional to their importance weights. This importance resampling step is included to improve the importance sampling approximation (Gelman, Carlin, Stern and Rubin, (1995, Section 10.5)

In addition, the best linear predictors $\{\text{BLP}_k\}_{k=1,\ldots,29}$ are computed, and both sets of predictors are compared to the actual data $\{X_{598+k}\}_{k=1,\ldots,29}$. Summary results show that the MC predictor dominates the BLP in out-of-sample forecasts over these 29 months:

$$\frac{\sum_{k=1}^{29}(X_{598+k}-\text{BLP}_k)^2}{\sum_{k=1}^{29}\left(X_{598+k}-\widehat{\text{BP}}_k\right)^2}=1.08 \text{ and } \frac{\sum_{k=1}^{29}|X_{598+k}-\text{BLP}_k|}{\sum_{k=1}^{29}|X_{598+k}-\widehat{\text{BP}}_k|}=1.06.$$

These small advantages of the estimated BP over the BLP for forecasting monthly differences cumulate into larger advantages in forecasting monthly unemployment rates:

$$\frac{\sum_{k=1}^{29}(Y_{598+k}-\text{BLP}_k)^2}{\sum_{k=1}^{29}\left(Y_{598+k}-\widehat{\text{BP}}_k\right)^2}=1.33 \text{ and } \frac{\sum_{k=1}^{29}|Y_{598+k}-\text{BLP}_k|}{\sum_{k=1}^{29}|Y_{598+k}-\widehat{\text{BP}}_k|}=1.14.$$

Again, the MC predictor dominates the BLP in out-of-sample prediction over these 29 months.

The predicted values are quite similar using other importance samplers, both with heavier tails (rescaled $t_{2.5}$) and lighter tails (Gaussian). This gives us some confidence that the choice of the importance sampler is not too critical in this particular case.

Appendix. Stability of Recursions

Write

$$\tilde\theta(z)=(1+z)^{u_1}(1-z)^{u_2}\prod_{j=1}^{\ell}(1-2\cos\lambda_j z+z^2)\prod_{j=1}^{s-u_1-u_2-2\ell}(1-\xi_j^{-1}z),$$

where $u_1,u_2\in\{0,1\}$, $\lambda_1<\lambda_2<\cdots<\lambda_\ell$, and $|\xi_j|>1$. Then, starting from $w_n^{(i)},\ldots,w_{n+1-s}^{(i)}$, we have the backward recursion

$$\begin{aligned} w_{t-s}^{(i)}&=\frac{x_t}{\theta_s^*}-(\tilde\theta(B^{-1})-1)w_{t-s}^{(i)}\\ &=\frac{\theta_s^*\tilde\theta(B^{-1})w_{t-s}}{\theta_s^*}-(\tilde\theta(B^{-1})-1)w_{t-s}^{(i)}, \end{aligned}$$

for $t=n,n-1,\ldots,$ which can be rearranged as

$$\begin{aligned} 0&=\tilde\theta(B^{-1})\left\{w_{t-s}-w_{t-s}^{(i)}\right\}\equiv\tilde\theta(B^{-1})\delta_{t-s},\quad t=n,n-1,\ldots\\ &=\tilde\theta(B)\delta_{n-t},\quad t=s,s+1,\ldots, \end{aligned}$$

subject to initial conditions $\delta_n, \ldots, \delta_{n+1-s}$. These initial conditions are unknown in practice, but in theory the general solution of this homogeneous linear difference equation with constant coefficients is

$$\delta_{n-t} = u_1 a_1 + u_2 a_2 (-1)^t + \sum_{j=1}^{\ell} b_j \cos(\lambda_j t + c_j) + \sum_{j=1}^{s'} \sum_{k=1}^{r_j} d_{jk} t^{k-1} \xi_j^{-t},$$

where r_j is the multiplicity of root j, $\sum_{j=1}^{s'} r_j = s - u_1 - u_2 - 2\ell$, and the constants a_1, a_2, $\{b_j\}$, $\{c_j\}$, $\{d_{jk}\}$ are determined from the initial conditions (e.g., Brockwell and Davis, (1991, Section 3.6)). This solution remains bounded for all $t = 0, 1, \ldots$.

Now, starting from $z_{-q+1}^{(i)}, \ldots, z_{-q+r}^{(i)}$, run the forward recursions

$$\begin{aligned} z_{t-s}^{(i)} &= w_{t-s}^{(i)} - (\theta^\dagger(B) - 1) z_{t-s}^{(i)} \quad t = 1, 2, \ldots, n+s \\ &= w_{t-s} - \delta_{t-s} - (\theta^\dagger(B) - 1) z_{t-s}^{(i)} \\ &= z_{t-s} + (\theta^\dagger(B) - 1) z_{t-s} - \delta_{t-s} - (\theta^\dagger(B) - 1) z_{t-s}^{(i)}. \end{aligned}$$

Rearranging this expression yields

$$\delta_{t-s} = \theta^\dagger(B)(z_{t-s} - z_{t-s}^{(i)}) \equiv \theta^\dagger(B) \Delta_{t-s}, \tag{20}$$

a nonhomogeneous linear difference equation with constant coefficients. The general solution of the associated homogeneous equation is

$$\Delta_{t-s}^H = \sum_{j=s'+1}^{s'+r'} \sum_{k=1}^{r_j} d_{jk} t^{k-1} \xi_j^{-t},$$

where $\sum_{j=s'+1}^{s'+r'} r_j = r$. Clearly $\{\Delta_{t-s}^H\}$ remains bounded for all $t = 1, 2, \ldots$. A particular solution of (20) is

$$\Delta_{t-s}^P = \frac{\delta_{t-s}}{\theta^\dagger(B)} = \sum_{j=0}^{\infty} \psi_j \delta_{t-s-j},$$

where the ψ_j are determined from $1 = \theta^\dagger(z)(1 + \psi_1 z + \psi_2 z^2 + \cdots)$. This particular solution is convergent by the boundedness of $\{\delta_{t-s}\}$ and absolute summability of $\{\psi_j\}$. Thus, the general solution of (20) is $\Delta_{t-s} = \Delta_{t-s}^H + \Delta_{t-s}^P$ (see, for example, Hildebrand (1968, p.31)), which remains bounded for all $t = 1, 2, \ldots$. In other words, the recursion is numerically stable because its error remains bounded.

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 Trans. Automat. 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. *Ann. Statist.* **29**, 919-946.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods.* 2nd edition. 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 Trans. Signal Processing* **45**, 1742-1762.

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.

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

Giannakis, G. B. and Swami, A. (1990). On estimating noncausal nonminimum phase ARMA models of non-Gaussian processes. *IEEE Trans. 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 Trans. Geoscience and Remote Sensing* **23**, 797-808.

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

Kanter, M. (1979). Lower bounds for nonlinear prediction error in moving average processes. *Ann. Probab.* **7**, 128-138.

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.

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

Shepp, L. A., Slepian, D. and Wyner, A. D. (1980). On prediction of moving average processes. *The Bell System Tech. J.* **59**, 367-415.

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

Department of Statistics, Colorado State University, 201 Statistics Building, Ft. Collins, CO80523, U.S.A.

E-mail: jbreidt@stat.colostate.edu

Institute of Statistics, National Tsing-Hua University, Hsin-chu, Taiwan 30043.

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

(Received June 2003; accepted April 2004)