doi:10.1111/j.1467-9892.2007.00529.x

MAXIMUM LIKELIHOOD ESTIMATION OF VARMA MODELS USING A STATE-SPACE EM ALGORITHM

By Konstantinos Metaxoglou and Aaron Smith

University of California, Davis

First Version received October 2005

Abstract. We introduce a state-space representation for vector autoregressive moving-average models that enables maximum likelihood estimation using the EM algorithm. We obtain closed-form expressions for both the E- and M-steps; the former requires the Kalman filter and a fixed-interval smoother, and the latter requires least squares-type regression. We show via simulations that our algorithm converges reliably to the maximum, whereas gradient-based methods often fail because of the highly nonlinear nature of the likelihood function. Moreover, our algorithm converges in a smaller number of function evaluations than commonly used direct-search routines. Overall, our approach achieves its largest performance gains when applied to models of high dimension. We illustrate our technique by estimating a high-dimensional vector moving-average model for an efficiency test of California's wholesale electricity market.

Keywords. Vector autoregressive moving average; Kalman filter; missing data; closed form.

1. INTRODUCTION

There exists no closed-form solution to the problem of maximizing the likelihood of a vector autoregressive moving-average (VARMA) model. Consequently, applied researchers employ iterative gradient-based or direct-search optimization routines. Gradient-based methods may fail to find the maximum because of the highly nonlinear nature of the likelihood function. Direct-search methods usually require a large number of function evaluations because they use no information about the steepness and curvature of the function. As the model dimension increases, the instability of quasi-Newton and the slow convergence of direct searches become more severe problems. In this article, we develop a state-space representation that enables maximum likelihood (ML) estimation using closed-form expressions via the EM algorithm.

Our state-space representation exploits two features of VARMA models. First, a vector moving-average (VMA) model possesses the same Wold representation as a VMA plus white noise. Thus, we add white noise to the MA component of the model, which preserves its time-series structure and permits a state-space representation with nontrivial noise in the observation equation. In turn, this observation noise allows implementation of the EM algorithm as in Shumway

0143-9782/09/05 666–685 JOURNAL OF TIME SERIES ANALYSIS Vol. 28, No. 5
© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd., 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

and Stoffer (1982) and Engle and Watson (1983). Second, a stationary and invertible finite-order VARMA process reverts quickly to its mean. Hence, to account for the initial values of the autoregressive component, we develop our EM algorithm using a complete sample that includes a series of missing data prior to the observed sample.

Much of the difficulty in ML estimation of Gaussian ARMA models stems from the covariance matrix of the observed data. The inverse and determinant of this covariance matrix are highly nonlinear expressions of the models' parameters. Most methods in the literature either approximate this inverse and determinant to enable estimation using analytical expressions (e.g. Whittle, 1951; Durbin, 1959), or they employ various transformations to reduce the computational burden of ML estimation (e.g. Newbold, 1974; Ansley, 1979).

Whittle (1951) and Durbin (1959) develop approximations for univariate MA models. Godolphin (1984) extends Whittle's method to univariate ARMA models and shows how to calculate parameter estimates directly from the sample autocorrelations of the observed data. Durbin's approximation generates an algorithm in which the MA coefficients are calculated from the coefficients of a long-order autoregression. Hannan and Rissanen (1982) and Koreisha and Pukkila (1989) extend Durbin's method to univariate and multivariate ARMA models, respectively. Tunnicliffe-Wilson (1973), Reinsel *et al.* (1992) and de Frutos and Serrano (1997) propose similar approximate ML estimation methods. These closed-form methods are asymptotically equivalent to ML because their approximation error becomes negligible in large samples. However, they may deviate substantially from ML in small samples or for models with parameters close to the stationarity or invertibility boundary.

Box and Jenkins (1970) evaluate the likelihood function for a univariate ARMA by writing it as a function of the observed data and 'backcasted' values of pre-sample innovations. They use the infinite MA representation of the model and therefore require a pre-sample of sufficient size to allow complete mean reversion. Newbold (1974) improves computational efficiency by proposing a method that requires a small pre-sample of size equal to the maximum order of the autoregressive (AR) and MA components. Osborn (1977) and Hillmer and Tiao (1979) extend Newbold's method to VMA processes, and Nicholls and Hall (1979) adapt it to VARMA models. However, the formulation suggested by Newbold still demands calculation of the inverse and determinant of a high-dimensional covariance matrix that is highly nonlinear in the parameters. By modifying the algorithm so that it uses operations on lower dimension matrices, Mauricio (1995, 1997) improves its computational efficiency.

Using a different approach, Ansley (1979) obtains a covariance band matrix for a univariate ARMA model, which enables a simple Cholesky decomposition and computationally efficient evaluation of the likelihood. Ansley's formulation, which builds on Phadke and Kedem's (1978) results for VMA models, is often referred to as the innovations form of the likelihood. Alternatively, Gardner *et al.* (1980) show that this innovations form can be calculated by applying the Kalman filter to a Markovian state-space representation of the model. However, Mauricio

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

(2002) adapts Ansley's algorithm to the multivariate case and shows that it is more efficient for likelihood evaluation than the Kalman filter and methods that follow Newbold's approach.

As pointed out by Mauricio (1995), the literature on likelihood methods focuses much more on efficient evaluation of the likelihood than on its maximization. Conversely, we focus on likelihood maximization. Therefore, we use the Kalman filter along with a fixed interval smoother because it provides a convenient vehicle for the E step in the EM algorithm. The Kalman filter also facilitates a backcasting approach similar to Box and Jenkins (1970) to account for the initial values of the AR component, which we treat as missing data. Our state-space representation is related to that in Ansley and Kohn (1983) (also Jones, 1980), which applies to a VARMA model with missing data and observational error. Incorporating observational error is akin to adding noise to the AR component of the model and, in general, it does not preserve the Wold representation of the VARMA process. In contrast, we add noise to the MA component of the model, which preserves the Wold representation and enables us to uncover the original VARMA parameters.

We introduce our estimation method in Section 2. In Section 3, we compare the performance of our state-space EM algorithm with seven other optimization routines for ML estimation of both mixed VARMA and pure VMA models. In Section 4, we illustrate our technique through an efficient market hypothesis test of California's wholesale electricity market using the high-dimensional VMA model implied by the market structure. We find pronounced inefficiencies in the market, especially between June and November 2000, when the state's utilities overpaid to purchase electricity in the spot market. Section 5 concludes the article.

2. ESTIMATION METHOD

We consider the following VARMA(p, q) model of dimension d:

$$\Phi(L)Y_t = \Theta(L)u_t, \quad t = 1, \ldots, n, \tag{1}$$

where

$$\Phi(L) = I - \Phi_1 L - \cdots - \Phi_p L^p, \quad \Theta(L) = I + \Theta_1 L + \cdots + \Theta_q L^q,$$

Φ_i and Θ_i are $d \times d$ matrices, L denotes the lag operator, and u_t is a Gaussian vector white-noise process with zero mean and covariance matrix Σ_u. We assume stationarity and invertibility, which require the roots of $|\Phi(L)| = 0$ and $|\Theta(L)| = 0$ to be outside the unit circle. In addition, we make two assumptions to ensure identifiability of the parameters (Φ, Θ): (i) $\Theta(L)$ and $\Phi(L)$ have no common left factors other than unimodular ones, and (ii) with q as small as possible, and p as small as possible given q, $\text{rank}([\Phi_p, \Theta_q]) = d$ (Reinsel, 1993, Sect. 2.3.4). The innovations form of the log-likelihood for eqn (1) is:

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

$$l(\theta|\mathbf{Y}_n) = l_p - \left(\frac{n-p}{2}\right)\ln|\Sigma_u| - \frac{1}{2}\text{trace}\sum_{t=p+1}^{n}\left((\Sigma_u)^{-1}u_t u_t^{\mathrm{T}}\right), \tag{2}$$

where the term $l_p = l(Y_1,\ Y_2,\ldots,Y_p,\ u_{p-q+1},\ u_{p-q+2},\ldots,u_p)$ captures the contribution to the log-likelihood of the starting values of Y and u. The vector $\theta = \text{vec}(\Phi, \Theta, \Sigma_u)$ contains the unknown parameters. We use 'T' and '$|\cdot|$' as the transpose and the determinant symbols, respectively.

Two features complicate maximization of eqn (2) with respect to θ. First, lags of u_t appear in the model as predictors of Y_t but are not observable. Second, l_p is a highly nonlinear function of θ. To remove these two complications, we write eqn (1) using a particular state-space representation that allows the EM algorithm to produce closed-form expressions for the likelihood equations. First, we make use of the result that a VMA(q) plus white noise remains a VMA(q) (Pieris, 1988, Thm 2), and we write $\Theta(L)u_t \equiv \Gamma(L)v_t + \varepsilon_t$, where v_t and ε_t denote white-noise processes. This setup allows us to treat v_t and its lags as observable in the complete-data log-likelihood that underlies the EM algorithm. Second, we account for the nonlinearity of l_p by expanding the sample $\mathbf{Y}_n$ to $\mathbf{Y} = (\mathbf{Y}_m, \mathbf{Y}_n)$, such that $\mathbf{Y}_m = (Y_0,\ Y_{-1},\ldots,Y_{-m-p+1})$ are initial *unobserved* values. We backcast $\mathbf{Y}_m$ conditional on the *observed* $\mathbf{Y}_n$ and choose m as the point beyond which our backcasts equal the observed series' unconditional mean, which implies that the values preceding Y_{-m} do not affect $l(\theta|\mathbf{Y}_n)$. Thus, we write eqn (1) in state-space form as:

$$\begin{aligned} Y_t &= \Phi X_t + Z a_t + \varepsilon_t \quad \varepsilon_t \sim N(0, \Sigma_\varepsilon) \\ a_t &= T a_{t-1} + \eta_t \quad \eta_t \sim N(0, \Sigma_\eta) \end{aligned} \tag{3}$$

where $t = -m+1,\ldots,n$ and

$$X_t^{\mathrm{T}} = [\,Y_{t-1}^{\mathrm{T}},\ \ Y_{t-2}^{\mathrm{T}},\ \ \ldots,\ \ Y_{t-p}^{\mathrm{T}}\,]^{\mathrm{T}},\quad \Phi = [\,I\ \ \Phi_1\ \ \cdots\ \ \Phi_p\,],\quad Z = [\,I\ \ \Gamma_1\ \ \cdots\ \ \Gamma_q\,]$$

$$a_t^{\mathrm{T}} = [v_t^{\mathrm{T}},\ldots,v_{t-q}^{\mathrm{T}}]^{\mathrm{T}},\quad T = \begin{bmatrix} 0 & 0 \\ I_{dq} & 0 \end{bmatrix},\quad \Sigma_\eta = \begin{bmatrix} \Sigma_v & 0 \\ 0 & 0 \end{bmatrix},\quad \eta_t^{\mathrm{T}} = [v_t^{\mathrm{T}}, 0, \ldots, 0]^{\mathrm{T}}$$

and I_{dq} denotes an identity matrix of dimension $(dq \times dq)$. We label the equation for Y_t the observation equation and we refer to the equation for a_t as the state equation. The Gaussian independently distributed disturbance vectors ε_t and η_t are mutually uncorrelated and independent of a_{-m}.

There exists a unique mapping from the parameters $(Z, \Sigma_\varepsilon, \Sigma_v)$ in eqn (3) to the parameters (Θ, Σ_u) in eqn (1) by matching the MA parameters Θ_j to the infinite MA representation of eqn (3):

$$Y_t - \Phi X_t = (I + Z(I - TL)^{-1}KL)u_t = (I + Z(I + TL + T^2L^2 + \cdots)KL)u_t, \tag{4}$$

where the matrix K denotes the steady-state value of the Kalman gain (Hamilton, 1994, Sect. 13.5). Therefore, we obtain the original MA coefficients from the expression $\Theta_j = ZT^{j-1}K$. The structure of the transition matrix T implies

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

$ZT^jK = 0$ for $j \geq q$. Moreover, Σ_u equals the steady-state value of the covariance of the one-step prediction error from the Kalman filter applied to eqn (3). The Wold representation in eqn (4) illustrates the unique mapping from $(Z, \Sigma_\varepsilon, \Sigma_v)$ to (Θ, Σ_u). However, the reverse mapping is not unique, which implies that the parameters $(Z, \Sigma_\varepsilon, \Sigma_v)$ are not separately identified. Therefore, to identify the parameters in eqn (3), we set Σ_ε to a *fixed diagonal* matrix, which allows us to obtain ML estimates of (Z, Σ_v) and then to calculate from eqn (4) the ML estimates of the parameters of interest (Θ, Σ_u).

Omitting constants, the *complete-data log-likelihood* for eqn (3) is:

$$l(\theta|\mathbf{Y}, a) = l_{-m} - \left(\frac{n+m}{2}\right)(\ln|\Sigma_\varepsilon| + \ln|\Sigma_v|) - \frac{1}{2}\text{trace} \sum_{t=-m+1}^{n} \left((\Sigma_\varepsilon)^{-1}\varepsilon_t\varepsilon_t^{\mathrm{T}} + (\Sigma_v)^{-1}v_t v_t^{\mathrm{T}}\right), \tag{5}$$

where $\varepsilon_t = Y_t - \Phi X_t - Za_t$, $v_t = a_t - Ta_{t-1}$ and $l_{-m} = l(Y_{-m}, Y_{-m-1}, \ldots, Y_{-m-p+1}, a_{-m})$ capture the contribution to the log-likelihood of the data prior to $-m+1$. Because we set m large enough such that the values preceding Y_{-m} do not affect $l(\theta|\mathbf{Y}_n)$, we can treat l_{-m} as a constant. To maximize the *incomplete (observed)-data* log-likelihood $l(\theta|\mathbf{Y}_n)$, we apply the EM algorithm of Dempster *et al.* (1977) using the following decomposition of the complete data log-likelihood:

$$l(\theta|\mathbf{Y}, a) = l(\theta|\mathbf{Y}_n) + \log f(a, \mathbf{Y}_m|\mathbf{Y}_n, \theta). \tag{6}$$

The second term in eqn (6) refers to the logarithm of the density of the *unobserved (missing)* data, given the observed data (Krishnan and McLachlan, 1997, and references therein). The EM algorithm involves an iterative point-to-set map, $M(\theta)$, from the parameter space to itself that finds the zeros for the score of the *expected complete-data* log-likelihood, conditional on $\mathbf{Y}_n$.

From eqn (6), we write $l(\theta|\mathbf{Y}_n) = l(\theta|\mathbf{Y},a) - \log f(a,\mathbf{Y}_m|\mathbf{Y}_n, \theta)$, and in the E step we take expectations over the distribution of $(a, \mathbf{Y}_m)$, given $\mathbf{Y}_n$ and a current estimate of θ, say $\theta^{(i)}$:

$$l(\theta|\mathbf{Y}_n) = Q(\theta|\theta^{(i)}) - H(\theta|\theta^{(i)}) = \int\int l(\theta|\mathbf{Y}, a) f(a, \mathbf{Y}_m|\mathbf{Y}_n, \theta^{(i)}) \mathrm{d}a\, \mathrm{d}\mathbf{Y}_m - \int\int \log f(a, \mathbf{Y}_m|\mathbf{Y}_n, \theta) f(a, \mathbf{Y}_m|\mathbf{Y}_n, \theta^{(i)}) \mathrm{d}a\, \mathrm{d}\mathbf{Y}_m. \tag{7}$$

The M step solves for $\theta^{(i+1)}$ by maximizing $Q(\theta|\theta^{(i)})$ with respect to θ. Hence, for a sequence of iterates $\theta^{(0)}, \theta^{(1)}, \ldots$, with $\theta^{(i+1)} = M(\theta^{(i)})$, the difference in $l(\theta|\mathbf{Y}_n)$ at successive iterates is:

$$l(\theta^{(i+1)}|\mathbf{Y}_n) - l(\theta^{(i)}|\mathbf{Y}_n) = Q(\theta^{(i+1)}|\theta^{(i)}) - Q(\theta^{(i)}|\theta^{(i)}) - (H(\theta^{(i+1)}|\theta^{(i)}) - H(\theta^{(i)}|\theta^{(i)})). \tag{8}$$

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

The difference of Q functions is positive by construction and the difference in the H functions is negative by the concavity of the logarithmic function and Jensen's inequality (Dempster *et al.*, 1977). Therefore, each iteration of the algorithm increases $l(\theta|\mathbf{Y}_n)$ and, if $l(\theta|\mathbf{Y}_n)$ is bounded from above, $l(\theta^{(i)}|\mathbf{Y}_n)$ converges to a stationary value of $l(\theta|\mathbf{Y}_n)$ (conditions of Wu, 1983, summarized in Krishnan and McLachlan, 1997, Sect. 3.4.2).

The E step requires passes of the Kalman filter and a fixed-interval smoother to obtain the conditional mean and variance of the state and disturbance vectors. The Kalman filter recursions applied to eqn (3) provide a convenient prediction error decomposition for eqn (3) through the mean and the covariance of the state vector a_{t+1}, conditional on the observations $\mathbf{Y}_t = (Y_1, Y_2, \ldots Y_t)$ i.e. $a_{t+1|t} = E[a_{t+1}|\mathbf{Y}_t]$ and $P_{t+1|t} = \mathrm{var}(a_{t+1}|\mathbf{Y}_t)$. The backward recursions of a fixed-interval smoother then give the estimates of the state and disturbance vectors along with their covariances conditional on $\mathbf{Y}_n$, i.e. $a_{t|n} = E[a_t|\mathbf{Y}_n]$, $\varepsilon_{t|n} = E[\varepsilon_t|\mathbf{Y}_n]$, $v_{t|n} = E[v_t|\mathbf{Y}_n]$, and $P_{t|n} = \mathrm{var}(a_t|\mathbf{Y}_n)$, $\mathrm{var}(\varepsilon_t|\mathbf{Y}_n)$ $\mathrm{var}(v_t|\mathbf{Y}_n)$, respectively. However, because we do not observe the pre-sample data $\mathbf{Y}_m$ we also require the mean and variance of $\mathbf{Y}_m$ conditional on the observed data $\mathbf{Y}_n$. Therefore, we obtain the smoothed values $a_{t|n}$, $Y_{t|n}$, $X_{t|n}$ and $v_{t|n}$, as well as the associated covariances, by applying the Kalman filter and fixed-interval smoother to the following modified state-space representation of eqn (3):

$$\begin{aligned} Y_t &= \Lambda \xi_t \\ \xi_t &= F\xi_{t-1} + \zeta_t \end{aligned} \tag{9}$$

$$F = \begin{bmatrix} \Phi^1 & \Phi_p & \Gamma^1 & \Gamma_q \\ I_{dp-d} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & I_{dq-d} & 0 \end{bmatrix},$$

$$\begin{aligned} \xi_t &= [Y_t^{\mathrm{T}}, Y_{t-1}^{\mathrm{T}}, \ldots, Y_{t-p+1}^{\mathrm{T}}, v_t^{\mathrm{T}}, v_{t-1}^{\mathrm{T}}, \ldots, v_{t-q+1}^{\mathrm{T}}]^{\mathrm{T}} \\ \zeta_t &= [(v_t + \varepsilon_t)^{\mathrm{T}}, 0, \ldots, 0, v_t^{\mathrm{T}}, 0, \ldots, 0]^{\mathrm{T}} \\ \Lambda &= [I_d \quad 0 \quad \cdots \quad 0] \end{aligned},$$

where

$$\Phi^1 \equiv [\Phi_1 \quad \cdots \quad \Phi_{p-1}], \quad \Gamma^1 \equiv [\Gamma_1 \quad \cdots \quad \Gamma_{q-1}]$$

and I_r denotes an identity matrix of dimension $(r \times r)$. We treat the initial unobserved data $\mathbf{Y}_m$ as missing values of Y, which the Kalman filter and the fixed-interval smoother incorporate by setting the Kalman gain to zero for missing observations (Durbin and Koopman, 2001, Sect. 4.8). Furthermore, the algorithm can handle missing values in $\mathbf{Y}_n$ by applying the same adjustment.

Omitting constants, the *expected complete-data log-likelihood* is:

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

$$Q(\theta|\theta^{(i)}) = -\left(\frac{n+m}{2}\right)(\ln|\Sigma_\varepsilon|) - \frac{1}{2}\text{trace}\sum_{t=-m+1}^{n}\left((\Sigma_\varepsilon)^{-1}\left(\varepsilon_{t|n}\varepsilon_{t|n}^{\mathrm{T}} + \text{var}(\varepsilon_t|\mathbf{Y}_n)\right)\right)$$
$$- \left(\frac{n+m}{2}\right)(\ln|\Sigma_v|) - \frac{1}{2}\text{trace}\sum_{t=-m+1}^{n}((\Sigma_v)^{-1}(v_{t|n}v_{t|n}^{\mathrm{T}} + \text{var}(v_t|\mathbf{Y}_n))). \tag{10}$$

Using standard matrix calculus results, the M step, which solves for $\theta^{(i+1)}$ by maximizing $Q(\theta|\theta^{(i)})$ with respect to θ, implies the following analytical expressions:

$$\begin{bmatrix} Z^{(i+1)} \\ \Phi^{(i+1)} \end{bmatrix} = \begin{bmatrix} a_{|n}^{\mathrm{T}}a_{|n} + P_{|n}^{aa} & a_{|n}^{\mathrm{T}}\mathbf{X}_{|n} + P_{|n}^{aX} \\ \mathbf{X}_{|n}^{\mathrm{T}}a_{|n} + P_{|n}^{Xa} & \mathbf{X}_{|n}^{\mathrm{T}}\mathbf{X}_{|n} + P_{|n}^{XX} \end{bmatrix}^{-1} \begin{bmatrix} \left(\mathbf{Y}_{|n} - v_{|n}\right)^{\mathrm{T}}a_{|n} + P_{|n}^{Ya} - P_{|n}^{va} \\ \mathbf{X}_{|n}^{\mathrm{T}}\mathbf{Y}_{|n} + P_{|n}^{XY} \end{bmatrix} \tag{11}$$

$$\Sigma_v^{(i+1)} = \left(\frac{1}{n+m}\right)(v_{|n}^{\mathrm{T}}v_{|n} + P_{|n}^{vv}), \tag{12}$$

where

$$a_{|n}^{\mathrm{T}} = (a_{-m+1|n}^{\mathrm{T}}, a_{-m+2|n}^{\mathrm{T}}, \ldots, a_{n|n}^{\mathrm{T}}),$$

and $\mathbf{X}_{|n}^{\mathrm{T}}$ and $\mathbf{Y}_{|n}$ are defined similarly. In addition, we define

$$P_{|n}^{aX} \equiv \sum_{t=-m+1}^{n}\text{cov}(a_t, X_t|\mathbf{Y}_n)$$

and follow analogous notation for the other covariance matrices $P_{|n}^{vv}$, $P_{|n}^{aa}$, $P_{|n}^{Xa}$, $P_{|n}^{XX}$, $P_{|n}^{Ya}$, $P_{|n}^{va}$ and $P_{|n}^{XY}$. Finally, because Y_t is of dimension $d \times 1$ and Σ_ε is diagonal, eqn (11) is equivalent to d univariate regressions. We conclude this section with several remarks about our estimation method:

REMARK 1. The conditions for parameter identification in VARMA(p, q) models often require linear constraints on the elements of Φ and Θ, which we incorporate by amending the M step in eqn (11) to its restricted least squares analogue as in Shumway and Stoffer (1982). Specifically, for a matrix R and vectors β and q satisfying $R\beta = q$, with $\beta = \text{vec}([Z^{\mathrm{T}}, \Phi^{\mathrm{T}}]^{\mathrm{T}})$, the M step becomes:

$$\beta_R^{(i+1)} = \beta^{(i+1)} - (I \otimes G^{-1})R'(R(I \otimes G^{-1})R')^{-1}(R\beta^{(i+1)} - q),$$

where

$$\beta^{(i+1)} = \text{vec}([Z^{(i+1)\mathrm{T}}, \Phi^{(i+1)\mathrm{T}}]^{\mathrm{T}}) = (I \otimes G^{-1})\text{vec}(D)$$

is the unrestricted estimate in eqn (11) for appropriately defined G and D matrices.

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

REMARK 2. Applied researchers have discretion over the magnitude of the approximation error induced by backcasting. To make this error negligible, m should be chosen such that the backcasted value $Y_{-m|n}$ approximates $E(Y_{-m}) = 0$ to a high degree of precision. Thus, the required value of m may be large when the roots of $|\Phi(L)| = 0$ lie close to the unit circle (see Newbold, 1974).

REMARK 3. In a pure VMA(q) model, our EM algorithm requires no backcasting because lags of Y do not enter the model and the initial value a_0 is multivariate Gaussian with mean zero and variance Σ_v. In this case, omitting constants, the complete-data log-likelihood is:

$$l(\theta|\mathbf{Y}, a) = -\frac{n}{2}\ln|\Sigma_\eta| - \left(\frac{n+1}{2}\right)\ln|\Sigma_v| - \frac{1}{2}\text{trace}\left(\sum_{t=1}^{n}(\Sigma_\varepsilon)^{-1}\varepsilon_t\varepsilon_t^{\mathrm{T}} + \sum_{t=0}^{n}(\Sigma_v)^{-1}v_t v_t^{\mathrm{T}}\right). \tag{13}$$

Thus, the E step requires the Kalman filter and smoother applied directly to eqn (3) and the M step produces

$$Z^{(i+1)} = (a_{|n}^{\mathrm{T}}a_{|n} + P_{|n}^{aa})^{-1}\left((\mathbf{Y}_{|n} - v_{|n})^{\mathrm{T}}a_{|n} + P_{|n}^{Ya} - P_{|n}^{va}\right).$$

REMARK 4. The VARMA(p, q) process in eqn (1) has mean zero. We incorporate deterministic explanatory variables such as constants or time trends by adding these variables to the matrix X_t and their coefficients to Φ in eqn (3). However, generalizing to VARMAX models with stochastic explanatory variables would require a model to impute $\mathbf{X}_m$, the pre-sample values of the X variables.

REMARK 5. In the E step, we apply the univariate filtering and smoothing algorithm of Koopman and Durbin (2000). In the smoother pass, we use the iterations summarized in Durbin and Koopman (2001, Sect. 4.3.1), which avoid the inversion of the contemporaneous covariance matrix $P_{t|t}$. This univariate algorithm saves computation time and simplifies the calculation of the log-likelihood because it requires no matrix inversion in either the filter or the smoother pass.

REMARK 6. To assess the impact of noise in the observation equation on the global convergence rate of the EM algorithm, we performed simulations for a wide range of values of Σ_ε. Following Meng and Rubin (1994) we measured the convergence rate using

$$r = \max r_j, \quad r_j = \lim_{i\to\infty}\left(\frac{\left|\theta_j^{(i+1)} - \theta_j^{(i)}\right|}{\left|\theta_j^{(i)} - \theta_j^{(i-1)}\right|}\right),$$

where $\theta_j^{(i)}$ is the jth element of θ in the ith iteration. Our simulations showed that r is decreasing in the diagonal elements of Σ_ε, which implies a faster rate of

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

convergence for larger Σ_ε. Moreover, because a larger value of Σ_ε implies a lower value of Σ_v, these simulation findings are consistent with the theoretical result that the convergence rate is decreasing in the fraction of missing information (Krishnan and McLachlan, 1977, Sect. 3.9.3).

REMARK 7. Several methods exist for estimating the covariance of ML estimates obtained from the EM algorithm. In the Appendix, we present a closed-form expression for the asymptotic covariance of the ML estimates by means of a VAR(1) representation of a VARMA(p, q). This method is particularly useful for high-dimensional models, where numerical differentiation is computationally demanding. For low-dimensional models, Meilijson (1989) shows that numerical computation of the empirical observed information matrix consistently estimates the observed information matrix. An alternative procedure for a numerically stable asymptotic covariance estimation is the supplemented EM of Meng and Rubin (1991). Another option for consistent standard error estimation is to use the bootstrapping method for state-space models of Stoffer and Wall (1991).

3. SIMULATIONS

In this section, we compare our state-space EM algorithm with seven alternative optimization techniques, including one derivative-based method, three deterministic direct-search algorithms and three stochastic direct-search algorithms. The derivative-based method is quasi-Newton with BFGS Hessian updating, which is available as the 'fminunc' command in the Matlab optimization toolbox. The three deterministic search routines are the Nelder-Mead simplex (Lagarias *et al.*, 1998), the Generalized Pattern Search (Torczon, 1997), and the Mesh Adaptive Direct Search (Audet and Dennis, 2003) methods. The first of these three routines is available as the Matlab 'fminsearch' command and the other two are available via the 'patternsearch' command in Matlab's Genetic Algorithm and Direct Search (GADS) toolbox. For the first stochastic search routine, we use simulated annealing, for which we wrote a Matlab module based on Gauss code by Tsionas.[1] (See Corana *et al.*, 1987, for additional discussion of this version of simulated annealing; see also Goffe *et al.*, 1994.) Finally, we apply two genetic algorithms, the one available as the 'ga' command in Matlab's GADS toolbox, as well as Matlab code written by Gordy[2] for the genetic algorithm described by Dorsey and Mayer (1995). We use Matlab 7.1.0.124 (R14) under Microsoft Windows XP 5.1 on a Dell Dimension 4700 desktop with 1GB of RAM and an Intel Pentium 4 processor at 2.8 GHz. For each optimization routine available as a Matlab command, we use the default parameter settings. For the remainder, we use the parameter settings suggested by the author of each routine.

We generate data from VMA(1) and VARMA(1,1) models of dimensions d = 2, 5 and 15. To ensure identification, we specify the AR coefficients as diagonal matrices and set each diagonal element equal to 0.6. We draw the MA coefficients

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

from a normal distribution, obtaining smallest roots of 29.9, 9.2 and 1.69 for $N = 2$, 5 and 15, respectively. Because the smallest roots of both the AR and MA components exceed 1, the model is stationary and invertible. In all cases, we generate a sample size $T = 200$ and use a backcasting sample of 40 observations to estimate the VARMA models. In the E-step, we apply the univariate filtering and smoothing algorithm of Koopman and Durbin (2000) to the pure VMA models and the standard multivariate Kalman filter and smoother for the VARMA models.

The convergence criterion for our EM algorithm is $\min(\|\theta_{k+1} - \theta_k\|_\infty, \|l_{k+1} - l_k\|_\infty)$, where θ_k and l_k denote the values of the parameter vector and the log likelihood at the kth iteration. We use tolerance levels of 10^{-4} for d = 2, 5 and 10^{-2} for d = 15. In addition, we restrict the number of function evaluations for our EM algorithm to 500, with every pass of the Kalman filter constituting a function evaluation. We terminate the alternative routines based on one of three criteria: (i) if the algorithm achieves the EM convergence criteria, (ii) if the number of function evaluations exceeds the number required by the EM algorithm to meet its convergence criteria, and (iii) if the raw computation time in seconds exceeds the corresponding time required by the EM algorithm to meet its convergence criteria. For Matlab's genetic algorithm, we always use termination condition (iii) because its interface provides no way to directly control the number of function evaluations.

To assess stability, we begin each optimization routine with 10 different sets of initial values. We start the MA and AR parameters at their true values and vary the starting error covariance matrix Σ_u. We fix the diagonal covariance matrices Σ_ε at 0.2 times the starting value of the corresponding elements in the error covariance matrix Σ_v. We calculate four statistics to measure the performance of each optimization routine. The first statistic records the number of times out of 10 that each optimization routine fails. Failures are most commonly caused by non-invertible covariance matrices for the one-step prediction error in the Kalman filter. This measure reveals information about the stability of the routine. Next, conditional on the routine not failing, we record the mean and standard deviation of the log-likelihood function at termination. These statistics capture both the speed and stability of the algorithms. Slow routines may still be far from the maximum when they are terminated, causing them to exhibit a low average likelihood. Unstable algorithms may have a high standard deviation because they terminate at different points, depending on starting values. Finally, we record the computation time in seconds for each optimization routine, conditional on not failing. Slower algorithms may not exhibit a longer computation time than our algorithm because we do not allow any routine to use more function evaluations than our algorithm. As a point of reference, we also report the value of the log-likelihood for each model when evaluated at the true parameters $l(\theta^*)$.

Tables I and II show that in all cases our EM algorithm gets very close to the maximum. This result is consistent with the well-documented convergence properties of the EM algorithm in other contexts (e.g. Redner and Walker, 1984). The standard deviation of its log-likelihood value at termination is always

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

TABLE I
VARMA SIMULATION RESULTS

	No. of failures	LLF at termination		Time in seconds (mean)
		Mean	SD	
d = 2, $l(\theta^*) = -587.9$				
State-space EM	0	−571.3	0.0	21.7
Quasi-Newton	4	−571.2	0.1	26.8
Nelder-Mead Simplex	1	-573.7	1.1	24.1
Generalized Pattern Search	0	−591.7	21.1	13.2
Mesh Adaptive Direct Search	8	260,133.4	368,718.8	11.5
Simulated Annealing	10	–	–	–
Genetic Algorithm (Matlab)	7	4.9	4.8	21.0
Genetic Algorithm (Dorsey-Mayer)	10	–	–	–
d = 5, $l(\theta^*) = -1876.8$				
State-Space EM	0	−1857.3	0.1	82.8
Quasi-Newton	9	−2372.7	–	100.8
Nelder-Mead Simplex	0	−3241.4	937.3	57.3
Generalized Pattern Search	0	−2821.5	605.8	58.3
Mesh Adaptive Direct Search	4	45,358.2	87,504.5	49.0
Simulated Annealing	9	−367.5	–	32.0
Genetic Algorithm (Matlab)	6	38.6	73.4	88.6
Genetic Algorithm (Dorsey-Mayer)	10	–	–	–
d = 15, $l(\theta^*) = -7855.5$				
State-Space EM	0	−7683.6	3.8	1552.3
Quasi-Newton	5	−9499.5	1268.8	1232.8
Nelder-Mead Simplex	0	−11,843.7	2838.1	1379.2
Generalized Pattern Search	0	−11,655.6	2721.1	1248.8
Mesh Adaptive Direct Search	0	−10,708.6	2052.9	1247.2
Simulated Annealing	0	−4,695,418.6	9,519,968.9	1298.5
Genetic Algorithm (Matlab)	3	−269.1	615.5	1492.2
Genetic Algorithm (Dorsey-Mayer)	10	–	–	–

small and is often orders of magnitude smaller than the corresponding standard deviations for the other methods. For the VARMA models, the quasi-Newton method fails 4, 9 and 5 times out of 10 for d = 2, 5 and 15, respectively. When it does converge, it locates the maximum every time for the d = 2 case, but is far from the global maximum the one time it converges for d = 5. For d = 15, the quasi-Newton method returns an average log-likelihood of −9499.5, which is far below the corresponding value of −7683.6 for our state-space EM algorithm. Moreover, the standard deviation of the quasi-Newton log-likelihood equals 1268.8, compared with 3.8 for our algorithm. Coupled with the high failure rate, this large variation shows that the quasi-Newton method is unstable when applied to VARMA models. In the pure VMA models, the quasi-Newton method had zero failures but it proved to be particularly sensitive to its starting values with standard deviations of 172.2 and 3274.4 for the d = 5 and 15 cases. Thus, although the quasi-Newton method works well for VMA models of dimension 2, it is not robust to starting values for models of higher dimension.

In most cases, the direct-search methods perform worse than our state-space EM algorithm. Either they are too slow and therefore do not get close to the maximum

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

TABLE II
VMA Simulation Results

	No. of failures	LLF at termination		Time in seconds (mean)
		Mean	SD	
d = 2, $l(\theta^*) = -577.7$				
State-Space EM	0	−573.5	0.0	18.3
Quasi-Newton	0	−573.7	0.8	6.9
Nelder-Mead Simplex	0	−574.6	0.2	7.6
Generalized Pattern Search	0	−587.7	18.6	7.3
Mesh Adaptive Direct Search	0	−574.5	1.3	7.3
Simulated Annealing	10	–	–	–
Genetic Algorithm (Matlab)	7	12,398,543.5	21,277,906.4	18.1
Genetic Algorithm (Dorsey-Mayer)	0	−577.8	0	3.4
d = 5, $l(\theta^*) = -1876.8$				
State-Space EM	0	−1871.9	0.0	45.3
Quasi-Newton	0	−2069.7	172.2	24.1
Nelder-Mead Simplex	0	−3368.7	1014.9	19.8
Generalized Pattern Search	0	−3054.7	772.1	19.9
Mesh Adaptive Direct Search	0	−2135.8	195.9	20.4
Simulated Annealing	10	–	–	–
Genetic Algorithm (Matlab)	8	6,439,422.4	9,061,032.6	49.8
Genetic Algorithm (Dorsey-Mayer)	0	−2068.9	0.48	46.4
d = 15, $l(\theta^*) = -7855.5$				
State-Space EM	0	−8184.2	0.0	367.7
Quasi-Newton	0	−12,992.8	3274.4	456.1
Nelder-Mead Simplex	0	−12,969.2	3261.5	450.7
Generalized Pattern Search	0	−12,806.6	3167.3	151.2
Mesh Adaptive Direct Search	0	−9878.5	1263.6	151.4
Simulated Annealing	0	−8210.2	0.1	175.9
Genetic Algorithm (Matlab)	10	–	–	–
Genetic Algorithm (Dorsey-Mayer)	0	−8210.2	0.1	1357.2

in the allotted number of function evaluations, or they fail. The Nelder-Mead simplex and the Generalized Pattern Search algorithms perform well for d = 2, but poorly for $N = 5$ and 15. In these higher-dimensional cases, the zero failure rate suggests that these routines would eventually converge if given enough time. The other deterministic search algorithm, Mesh Adaptive Direct Search, performs poorly in all cases. Of the stochastic search methods, simulated annealing was the worst performer, failing almost every time for the VMA models and yielding nonsense results for the VARMA models. This algorithm requires the user to choose many inputs, and moving away from the recommended choices could improve the algorithm's performance (Corana *et al.*, 1987). The genetic algorithms also performed poorly in many cases. The exception is the version presented by Dorsey and Mayer (1995) which, despite being slow in terms of computation time, always finds the global maximum for the VMA models. However, it fails every time for the VARMA models.

Overall, our state-space EM algorithm proves robust and relatively fast. The only case in which its average computation time exceeds a reliable alternative is the VMA with d = 2. For this model, many of the methods achieve convergence

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

in less than 8 seconds, compared to 18 seconds for our algorithm. However, if we enlarge the model by adding autoregressive terms or extra variables, then our algorithm maintains its stability and becomes much faster than the alternatives. Moreover, to apply our algorithm the researcher has only to specify the covariance matrix of the additional noise and the number of additional observations for the purpose of the backcasting in the VARMA models. In contrast, numerous tuning parameters are required to initialize the other gradient-free methods.

4. PRICE SPREADS IN CALIFORNIA ELECTRICITY MARKETS

In this section, we test the efficiency of the California electricity markets using the high-dimensional VMA model implied by the market structure. On 1 April, 1998, the restructured California wholesale electricity sector commenced operation. The legislation assigned leading roles to two institutions, the Power Exchange (PX) and the Independent System Operator (ISO). The PX operated a day-ahead forward energy market, and the ISO operated the state's real-time or spot energy market. Bohn *et al*. (1998) provide additional details for the ISO spot and the PX market operations. The efficient market hypothesis implies that the suppliers of electricity try to sell in the market with the highest price whereas the buyers of electricity try to buy in the lower-priced market until they reach an equilibrium in the sense that the day-ahead price equals the expected spot price. We test the hypothesis that the difference (spread) between the spot and day-ahead prices has zero mean. (See Longstaff and Wang (2004), Saravia (2003), and Borenstein *et al*. (2004) for similar tests in the Pennsylvania/New Jersey/Maryland, New York, and California markets, respectively.)

In the day-ahead market, PX participants, such as generators, utilities, marketers and retailers submitted bids to sell and buy energy for the 24 hours of the following day, starting with the midnight to 1:00 AM interval. PX participants had to submit their bids by 7:00 AM the day before the production day for all of its 24 hours. At the 7:00 AM deadline, the intersection of 24 aggregate supply-and-demand curves produced a separate market-clearing price for each hour. To maintain system-wide balance between demand and supply, the ISO used bids in its real-time market to increment and decrement supply or to decrement demand. For the purpose of real-time pricing, the ISO calculated the net incremental or decremental energy usage over 10-minute intervals for every hour. A MWh-weighted average of the six incremental and/or decremental prices produced the real-time *ex post* price. We refer to this real-time *ex post* price as the spot price.

Because PX participants submitted their bids up to 7:00 AM, there was a window of between 18 and 41 hours from the submission of the day-ahead bids to the time that the ISO cleared the corresponding spot market. Thus, the difference between the day-ahead and the spot prices depends on accumulated information in the intervening hours, which implies an MA structure for the

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

price spreads (Borenstein *et al.*, 2004). We denote the information arriving in hour ending h and day d by u_{hd} and we express the difference between the day-ahead and the spot prices as MA processes of between 18 and 41 u_{hd} terms. With an observation in our sample being $y_{hd} = \mathrm{ISO}_{hd} - \mathrm{PX}_{hd}$, where ISO_{hd} and PX_{hd} are the spot and the day-ahead prices for hour ending $h = 1,\ldots,24$ and day d, we write:

$$\begin{array}{lllllllll}
y_{1d} = \beta_1 & & & & + & u_{1d} & + & \theta_{1,1}u_{24,d-1} & +\cdots+ \quad \theta_{1,18}u_{7,d-1} \\
y_{2d} = \beta_2 & & + & u_{2d} & + & \theta_{2,1}u_{1d} & + & \theta_{2,2}u_{24,d-1} & +\cdots+ \quad \theta_{2,19}u_{7,d-1} \\
\vdots & & & \vdots & & \vdots & & \vdots & \vdots \\
y_{24d} = \beta_{24} & + \quad u_{24d} & +\cdots+ & \theta_{24,22}u_{2d} & + & \theta_{24,23}u_{1d} & + & \theta_{24,24}u_{24,d-1} & +\cdots+ \quad \theta_{24,41}u_{7,d-1}
\end{array}$$

with an equivalent VMA(1) representation of dimension 24:

$$Y_d = B + \Theta_0 U_d + \Theta_1 U_{d-1}, \tag{14}$$

where $Y_d = \mathrm{ISO}_d - \mathrm{PX}_d$, for appropriately defined matrices Θ_0 and Θ_1. The serially uncorrelated zero-mean Gaussian error vector U_d has diagonal covariance matrix Σ_u and $B = (\beta_1,\ldots,\beta_{24})'$. Because our hypothesis is that the day-ahead prices are unbiased estimates of the spot prices, we test whether B is statistically significantly different from zero. Taking into account the MA structure is important for correct inference about B.

The model in eqn (14) contains 708 MA parameters, so numerical differentiation by gradient-based methods is computationally infeasible. We estimate the model by applying our EM algorithm to the case of a pure VMA model (Remark 3). We iterate the filtering and smoothing recursions of Koopman and Durbin (2000) in the E-step and 24 univariate regressions in the M-step until convergence. We calculate the standard errors for B using the empirical observed information matrix of Meilijson (1989). Following Remark 7, we derive the standard errors associated with the MA parameters in Θ as in the Appendix, $\mathrm{var}(\mathrm{vec}(\Theta)) = (1/n)\Sigma_U \otimes \Sigma_Y^{-1}$.

The California electricity market entered an almost year-long period of crisis beginning in May 2000. Because of the crisis, PX ceased its operations at the end of January 2001, although it was already malfunctioning in December 2000 (Borenstein *et al.*, 2004). Hence, we perform our hypothesis test separately for the pre-crisis period, 1 April 1998 to 31 May 2000, and the crisis period, 1 June 2000 to 30 November 2000. We have 792 and 183 observations for each hour in the two periods, respectively. The ISO spot prices are publicly available from the Open Access Same Time Information System of the California ISO. The PX prices are also publicly available, through the University of California Energy Institute website. In both the day-ahead and spot markets, binding transmission constraints along Path 15, the state's major transmission line gave rise to different prices north of Path 15 (NP15) and south of Path 15 (SP15). We estimate the 24-hourly price spreads (spot minus day-ahead) for both NP15 and SP15.

To illustrate the MA structure, we plot in Figure 1 the estimated MA coefficients for the peak hour 16:00 for SP15 in the pre-crisis period. MA

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

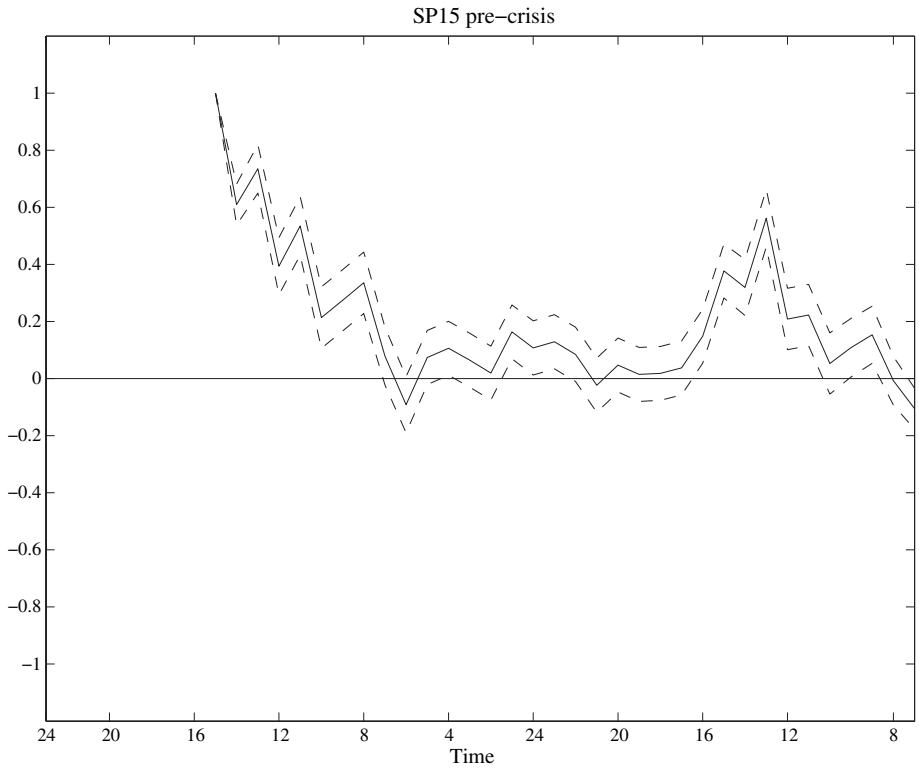

FIGURE 1. MA coefficient estimates for hour 16:00

parameters for hours around 16:00 exhibit similar patterns in terms of their values and signs. This pattern implies that most information arrives in the market during daytime hours, beginning with the hour that market participants submit their day-ahead bids. Other daily events that provide information include two rounds of bidding in the ISO reserves market (10:00 and 12:00 AM), the ISO's announcement of its requirements for reliability must run units (around 14:00 PM), and the ISO's 2-day-ahead forecasts of system conditions (17:00 PM). Similar to Figure 1, early morning and late evening hours generally exhibit large and positive MA coefficients at short lags. However, at longer lags, these hours exhibit some idiosyncrasies, most probably due to thinner trading outside regular business hours. There are many more significant MA coefficients for all hours in both SP15 and NP15 during the pre-crisis period than the crisis period. The reduction in significance in the crisis period probably reflects the unpredictable conditions of the entire wholesale energy market in the state during that time.

For the pre-crisis period, the NP15 and SP15 spreads are different from zero at 5% significance level for 14 and 12 out of the 24 hours, respectively, as shown in Figure 2. For the crisis period, the number of hours of significant NP15 spreads

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

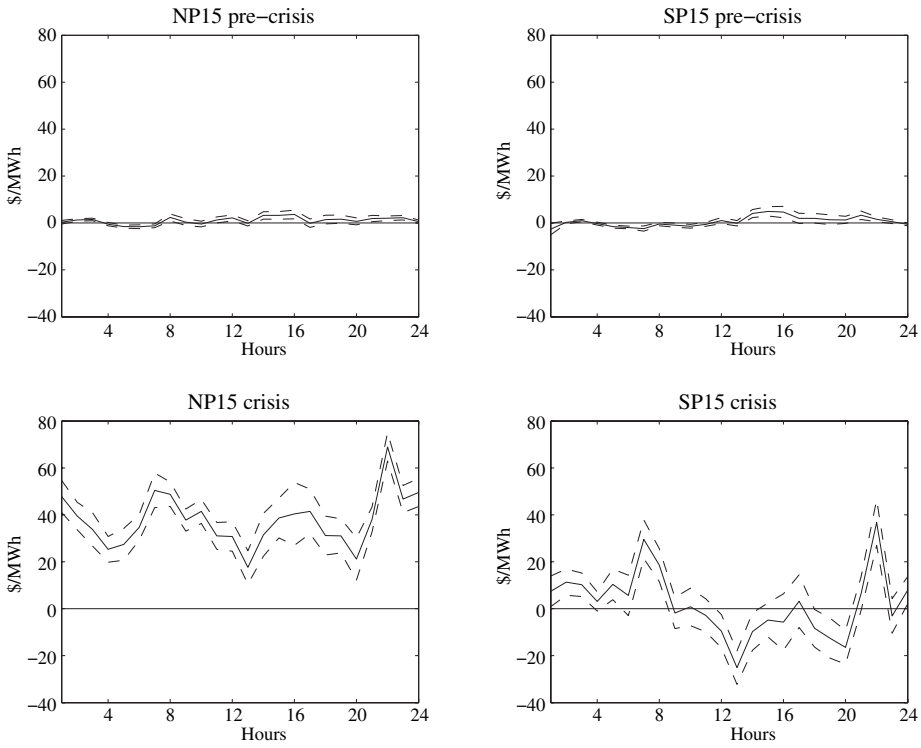

FIGURE 2. Estimated mean hourly spreads (B)

rises to 24, and the number of significant SP15 spreads increases slightly to 14 hours. The price spreads are substantially greater during the crisis period. Averaging over the significant NP15 spreads, we get spreads of \$1.5/MWh (pre-crisis) and \$37.7/MWh (crisis). When evaluated at the mean day-ahead price, these spreads imply day-ahead premia of 5% and 30.8%, respectively. In the SP15, the same calculations imply average spreads of \$0.7/MWh (pre-crisis) and \$3.6/MWh (crisis), which correspond to day-ahead premia of 2.6% and 3.1%, respectively. Thus, utilities paid significant premia to buy electricity in the spot market, especially in NP15 during the crisis period. To account for the effect of outliers, a common feature of the electricity prices, we repeated our estimation treating spreads below the 1% and above the 99% percentiles as missing. The number of statistically significant spreads and their values did not change substantially.

Why did electricity suppliers not take advantage of the systematically higher prices by supplying more electricity in the spot market, especially in NP15 during the crisis? Borenstein *et al.* (2004) conclude that uncertainty regarding regulatory penalties for spot trading led the majority of the market participants to avoid arbitrage between the spot and day-ahead markets. Moreover, those participants

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

who undertook such risky arbitrage did not find it profit-maximizing to eliminate the price differences, and so they limited their trading volume. On the demand side, the restructured market had left utilities with little incentive to respond to price differences because they were collecting a Competition Transition Charge, the difference between their fixed retail revenue and their wholesale costs, up to May 2000. In the crisis period after May 2000, Borenstein *et al.* (2004) argue that Pacific Gas and Electric, the largest utility operating in NP15, may have exercised monopsony power to affect prices to its advantage.

5. CONCLUSION

We provide a new method to alleviate the computational burden associated with ML estimation of VARMA models. Specifically, we propose a state-space representation that allows the EM algorithm to produce analytical expressions for the log-likelihood equations. The E step of our algorithm involves a pass of the Kalman filter along with a fixed-interval smoother. The M step collapses to least-squares-type regression. We show using simulations that our method is robust to starting values and converges quickly to the maximum. We illustrate our technique by estimating a high-dimensional VMA for an efficiency test of the restructured wholesale California electricity market.

The appealing properties of our algorithm mirror those of the EM algorithm in many other contexts. As Redner and Walker (1984, Sect. 2.4) note, the EM algorithm 'has been found in most instances to have the advantage of reliable global convergence, low cost per iteration, economy of storage, and ease of programming, as well as a certain heuristic appeal.' However, compared with the quadratic convergence of Newton–Raphson and the super-linear convergence of quasi-Newton methods, the convergence of the EM algorithm is linear at a rate determined by the proportion of missing information. Jamshidian and Jennrich (1997) and Meng and Van Dyk (1997) discuss acceleration techniques to improve the convergence rate. Nevertheless, the simplicity of our closed-form EM algorithm is a redeeming feature that may trump a more complicated accelerated algorithm (see Lange, 1995).

APPENDIX

We write the d-dimensional VARMA(p, q) in eqn (1) as a VAR(1) $\bar{Y}_t = A\bar{Y}_{t-1} + N_t$, with N_t being zero mean independently and identically distributed with covariance matrix Σ_N (Lütkepohl, 1993, Sect. 6.3), where:

$$\bar{Y}_t = [\, Y_t^{\mathrm{T}} \quad Y_{t-1}^{\mathrm{T}} \quad \cdots \quad Y_{t-p+1}^{\mathrm{T}} \quad u_t^{\mathrm{T}} \quad u_{t-1}^{\mathrm{T}} \quad \cdots \quad u_{t-q+1}^{\mathrm{T}} \,]^{\mathrm{T}}$$

$$N_t = [\, u_t^{\mathrm{T}} \quad 0 \quad \cdots \quad 0 \quad u_t^{\mathrm{T}} \quad 0 \quad \cdots \quad 0 \,]^{\mathrm{T}}$$

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad \Sigma_N = \begin{bmatrix} \Sigma_N & 0 & \cdots & 0 & \Sigma_N & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \Sigma_N & 0 & \cdots & 0 & \Sigma_N & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$A_{11} = \underset{[dp\times dp]}{\begin{bmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_{p-1} & \Phi_p \\ I_d & 0 & \cdots & 0 & 0 \\ 0 & I_d & & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_d & 0 \end{bmatrix}}, \quad A_{12} = \underset{[dq\times dq]}{\begin{bmatrix} \Theta_1 & \Theta_2 & \cdots & \Theta_{q-1} & \Theta_q \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}}$$

$$A_{21} = \underset{[dq\times dp]}{\begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}}, \quad A_{22} = \underset{[dq\times dq]}{\begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ I_d & 0 & \cdots & 0 & 0 \\ 0 & I_d & & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_d & 0 \end{bmatrix}}$$

As a result of the asymptotic equivalence between the exact and conditional ML estimate A, say $\hat{A}$, we use the covariance matrix of the conditional ML estimate (Reinsel, 1993, Sect. 4.3.1):

$$\begin{aligned} \operatorname{var}(\operatorname{vec}(\hat{A})) &= \left(\frac{1}{n}\right)\Sigma_N \otimes \Sigma_Y^{-1}, \\ \operatorname{vec}(\Sigma_Y) &= (I_{d^2(p+q)^2} - (A \otimes A))^{-1}\operatorname{vec}(\Sigma_N). \end{aligned} \tag{15}$$

ACKNOWLEDGEMENTS

We are grateful to an associate editor and an anonymous referee for their suggestions on an earlier draft of the paper. We would also like to thank Colin Cameron, Timothy Cogley, Oscar Jorda, Prasad Naik, Robert Shumway and the participants of the spring 2005 econometrics seminar at UC Davis. All remaining errors are ours.

NOTES

1. See the Gauss Archive of the American Univeristy.

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

2. Gordy's Matlab code is available at http://mgordy.tripod.com/research.html#software.

Corresponding author: Aaron Smith, Department of Ag & Resource Economics, University of California, One Shields Avenue, Davis, CA 95616, USA. E-mail: adsmith@ucdavis.edu

REFERENCES

Ansley, C. F. (1979) An algorithm for the exact likelihood of mixed autoregressive moving average process. *Biometrika* 66, 59–65.

Ansley, C. F. and Kohn, R. (1983) Exact likelihood of vector autoregressive-moving average processes with missing or aggregated data. *Biometrika* 70, 275–78.

Audet, C. and Dennis J. E. Jr. (2003) Analysis of generalized pattern searches. *SIAM Journal on Optimization* 13, 889–903.

Bohn, R. E., Klevorick, A. K. and Stalon, C. G. (1998) *Report on Market Issues in the California Power Exchange Energy Markets*. Prepared for the Federal Energy Regulatory Commission by the Market Monitoring Committee of the California Power Exchange. http://www.ucei.berkeley.edu/restructuring.html

Borenstein, S., Bushnell, J., Knittel, C. R. and Wolfram, C. (2004) Inefficiencies and market power in financial arbitrage: a study of California's electricity markets, POWER Working Paper PWP-138, University of California Energy Institute.

Box, G. E. P. and Jenkins, G. M. (1970) *Time Series Analysis: Forecasting and Control*. Oakland, CA: Holden-Day.

Corana, A., Marchesi, M., Martini, C. and Ridella, S. (1987) Minimizing multimodal functions of continuous variables with the simulated annealing algorithm. *ACM Transactions on Mathematical Software* 13, 262–80.

Dempster, A., Laird, N. and Rubin, D. B. (1977) Maximum Likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B)* 39, 1–38.

Dorsey, R. E. and Mayer, W. J. (1995) Genetic algorithms for estimation problems with multiple optima, nondifferentiability and other irregular features. *Journal of Business and Economics Statistics* 13, 53–66.

Durbin, J. (1959) Efficient estimation of parameters in moving-average models. *Biometrika* 46, 306–16.

Durbin, J. and Koopman, S. J. (2001) *Time Series Analysis by State Space Models*. Oxford: Oxford University Press.

Engle, R. F. and Watson, M. (1983) Alternative Algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics* 23, 385–400.

de Frutos, R. F. and Serrano, G. R. (1997) A generalized least squares estimation method for invertible vector moving average models. *Economics Letters* 57, 149–56.

Gardner, G., Harvey, A. C. and Phillips, G. D. A. (1980) An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of kalman filtering. *Applied Statistics* 29, 311–22.

Godolphin, E. J. (1984) A direct representation for the large sample maximum likelihood estimator of a Gaussian autoregressive moving average process. *Biometrika* 71, 281–89.

Goffe, W., Ferrier, G. D. and Rogers, J. (1994) Global optimization of statistical functions with simulated annealing. *Journal of Econometrics* 60, 65–99.

Hamilton, J. (1994) *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hannan, E. J. and Rissanen J. (1982) Recursive estimation of mixed autoregressive-moving average order. *Biometrika* 69, 81–94.

Hillmer, S. C. and Tiao, G. C. (1979) Likelihood function of stationary multiple autoregressive moving average models. *Journal of the American Statistical Association* 74, 652–60.

Jones, R. H. (1980) Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics* 22, 389–95.

Jamshidian, M. and Jennrich, R. I. (1997) Acceleration of the EM Algorithm using quasi-newton methods. *Journal of the Royal Statistical Society (Series B)* 59, 569–87.

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.

KOOPMAN, S. J. and DURBIN, J. (2000) Fast filtering and smoothing for multivariate state space models. *Journal of Time Series Analysis* 21, 281–96.
KOREISHA, S. and PUKKILA, T. (1989) Fast linear estimation methods for vector moving average models. *Journal of Time Series Analysis* 10, 325–39.
KRISHNAN, T. and MCLACHLAN (1997) *The EM Algorithm and Extensions*. New York: Willey.
LAGARIAS, J. C., REEDS, J. A., WRIGHT, M. H. and WRIGHT, P. E. (1998) Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal on Optimization* 9, 112–47.
LANGE, K. (1995) A quasi-Newton acceleration of the EM Algorithm. *Statistica Sinica* 5, 1–18.
LONGSTAFF, F. A. and WANG, A. W. (2004) Electricity forward prices: a high-frequency empirical analysis. *Journal of Finance* 59, 1877–900.
LÜTKEPOHL, H. (1993) *Introduction to Multiple Time Series Analysis*. New York: Springer-Verlag.
MAURICIO, J. A. (1995) Exact maximum likelihood estimation of stationary vector ARMA models. *Journal of the American Statistical Association* 90, 282–91.
MAURICIO, J. A. (1997) The exact likelihood of a vector autoregressive moving average model. *Applied Statistics* 46, 157–71.
MAURICIO, J. A. (2002) An algorithm for the exact likelihood of a vector autoregressive moving average model. *Journal of Time Series Analysis* 23, 473–86.
MENG, X. L. and VAN DYK, D. (1997) The EM algorithm – an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society (Series B)* 59, 511–67.
MENG, X. L. and RUBIN, D. B. (1991) Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association* 86, 899–909.
MENG, X. L. and RUBIN, D. B. (1994) On the global and component-wise rates of convergence of the EM algorithm. *Linear Algebra and Its Applications* 199, 413–25.
MEILIJSON, I. (1989) A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society (Series B)* 51, 127–38.
NEWBOLD, P. (1974) The exact likelihood function for a mixed autoregressive moving average process. *Biometrika* 61, 423–6.
NICHOLLS, D. F. and HALL, A. D. (1979) The exact likelihood function of multivariate autoregressive-moving average models. *Biometrika* 66, 259–64.
OSBORN, D. R. (1977) Exact and approximate maximum likelihood estimators for vector moving average processes. *Journal of the Royal Statistical Society (Series B)* 39, 114–8.
PHADKE, M. S. and KEDEM, G. (1978) Computation of the exact likelihood function of multivariate moving average models. *Biometrika* 65, 511–19.
PIERIS, S. (1988) On the study of some functions of multivariate ARMA processes. *Journal of Multivariate Analysis* 25, 146–51.
REDNER, R. A. and WALKER, H. F. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* 26, 195–239.
REINSEL, G. C. (1993) *Elements of Multivariate Time Series Analysis*. New York: Springer-Verlag.
REINSEL, G. C., BASU, S. and YAP, S. F. (1992) Maximum likelihood estimators in the multivariate autoregressive moving average model from a generalized least squares viewpoint. *Journal of Time Series Analysis* 13, 133–45.
SARAVIA, C. (2003) Speculative trading and market performance: the effect of arbitrageurs on efficiency and market power in the New York electricity market. CSEM Working Paper CSEM WP 121, University of California Energy Institute.
SHUMWAY, R. H. and STOFFER, D. S. (1982) An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis* 3, 253–64.
STOFFER, D. S. and WALL, K. D. (1991) Bootstrapping state space models: Ggaussian maximum likelihood estimation and the Kalman filter. *Journal of the American Statistical Association* 86, 1024–33.
TORCZON, V. (1997) On the convergence of pattern search algorithms. *SIAM Journal on Optimization* 7, 1–25.
TUNNICLIFFE-WILSON, G. (1973) The estimation of parameters in multivariate time series models. *Journal of the Royal Statistical Society (Series B)* 35, 76–85.
WHITTLE, P. (1951) *Hypothesis testing in time series Analysis*. Upsala: Almquist and Wiksell.
WU, C. F. J. (1983) On the convergence properties of the EM algorithm. *The Annals of Statistics* 11, 95–103.

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd.