The relationship between the conditional sum of squares and the exact likelihood for autoregressive moving average models

Neil Shephard
Nuffield College, Oxford University, Oxford OX1 1NF, UK
neil.shephard@nuf.ox.ac.uk

September 21, 1997

Summary

In this note I will study the relationship between the conditional sum of squares (CSS) estimator of moving averages and the maximum likelihood (ML) estimator. I will show that the CSS estimator can be converted into the ML estimator via the use of the EM algorithm. A by-product of the EM algorithm is an expression for the likelihood function and the score. This argument generalizes to autoregressive moving average (ARMA) models.

Some key words: Kalman filter, autoregression, EM algorithm, moving average, score.

1. INTRODUCTION

1.·1 *Motivation*

In this note I will study the relationship between the conditional sum of squares (CSS) estimator (see Box & Jenkins (1976, p. 237) or, for example, Harvey (1993, pp.60-62)) of moving averages and the maximum likelihood (ML) estimator. I will show that the CSS estimator can be converted into the ML estimator via the use of the EM algorithm. A by-product of the EM algorithm is a simple expression for the likelihood function and the score. This argument generalizes to autoregressive moving average (ARMA) models.

1.·2 *First order moving average*

To focus ideas consider the first order moving average for a time series y_t, $t = 1, ..., n$, where $y_t = \varepsilon_t + \theta\varepsilon_{t-1}$. Here ε_t is Gaussian, zero mean, white noise with a variance of σ^2 which, for simplicity of exposition, I will assume is known. The likelihood function can be computed in a number of ways, the most common being the Kalman filter (see, for example, Harvey (1993, Ch. 4)) while the score is available through the use of the Kalman filter and smoother. See de Jong (1989) for a discussion of the Kalman filter smoother and Koopman & Shephard (1992) for its use in deriving the score for models which can be put in state space form. A special case of this is the score for ARMA models. Here we give much simpler expressions for the likelihood and score based on the CSS estimator.

The CSS estimator works off the conditional likelihood function, writing $y = (y_1, ..., y_n)'$,

$$\log f(y|\varepsilon_0 = 0; \theta) = -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{n} \varepsilon_t^{*2},$$

where the $\varepsilon_1^*, ..., \varepsilon_n^*$ are computed recursively using

$$\varepsilon_t^* = y_t - \theta\varepsilon_{t-1}^*, \qquad t = 1, ..., n, \qquad \text{with} \qquad \varepsilon_0^* = 0. \tag{1}$$

Notice that the ε_t^* are not the true ε_t unless the true ε_0 happened to be exactly zero. The conditional likelihood is very attractive as the corresponding conditional score can be computed as

$$\frac{\partial \log f(y|\varepsilon_0 = 0; \theta)}{\partial\theta} = -\frac{1}{\sigma^2} \sum_{t=1}^{n} \frac{\partial\varepsilon_t^*}{\partial\theta} \varepsilon_t^*,$$

where the derivatives, $\partial\varepsilon_t^*/\partial\theta$, can be computed in parallel with (1) as

$$\frac{\partial\varepsilon_t^*}{\partial\theta} = -\varepsilon_{t-1}^* - \theta\frac{\partial\varepsilon_{t-1}^*}{\partial\theta}, \qquad t = 1, ..., n, \qquad \text{with} \qquad \frac{\partial\varepsilon_0^*}{\partial\theta} = 0. \tag{2}$$

Likewise the observed conditional information is

$$\frac{\partial^2 \log f(y|\varepsilon_0 = 0; \theta)}{\partial\theta\partial\theta'} = -\frac{1}{\sigma^2} \sum_{t=1}^{n} \frac{\partial\varepsilon_t^*}{\partial\theta} \frac{\partial\varepsilon_t^*}{\partial\theta'} - \frac{1}{\sigma^2} \sum_{t=1}^{n} \varepsilon_t^* \frac{\partial^2\varepsilon_t^*}{\partial\theta\partial\theta'}.$$

The second term in this sum is usually ignored as its expectation, conditional on $\varepsilon_0 = 0$, is zero. This suggests numerically computing the CSS estimator $\widetilde{\theta}$ by the recursion

$$\theta_{(i)} = \theta_{(i-1)} - \left(\sum_{t=1}^{n} \frac{\partial\varepsilon_t^*}{\partial\theta} \frac{\partial\varepsilon_t^*}{\partial\theta'} \right)^{-1} \sum_{t=1}^{n} \frac{\partial\varepsilon_t^*}{\partial\theta} \varepsilon_t^*. \tag{3}$$

1·3 *Profile likelihood*

Instead of conditioning on $\varepsilon_0 = 0$ I could have run the CSS recursion (1) starting with any initial ε_0. The result would have been a sequence of errors $\varepsilon_1^+, ..., \varepsilon_n^+$ which depended on the particular choice of ε_0. As the moving average process is a linear model

$$\varepsilon_t^+ = \varepsilon_t^* + \varepsilon_0 r_t, \qquad t = 1, ..., n,$$

where

$$r_t = \left. \frac{\partial\varepsilon_t^+}{\partial\varepsilon_0} \right|_{\varepsilon_0 = 0}.$$

The $r_1, ..., r_n$, can be recursively computed in parallel with ε_t^* as

$$(\varepsilon_t^*, r_t) = (y_t, 0) - \theta \left(\varepsilon_{t-1}^*, r_{t-1} \right), \qquad t = 1, ..., n, \qquad \varepsilon_0^* = 0, r_0 = 1. \tag{4}$$

Thus the more general conditional likelihood is

$$\log f(y|\varepsilon_0; \theta) = -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{n} \varepsilon_t^{+2} = -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \left(\sum_{t=1}^{n} \varepsilon_t^{*2} + 2\varepsilon_0 \sum_{t=1}^{n} \varepsilon_t^* r_t + \varepsilon_0^2 \sum_{t=1}^{n} r_t^2 \right), \tag{5}$$

which explicitly relates the initial condition to the conditional likelihood. If we were to treat ε_0 as an unknown parameter and then used ML methods to estimate it in addition to θ, then the ML estimator of ε_0 is the regression of ε_t^* on $-r_t$, yielding

$$\widehat{\varepsilon_0} = -\frac{\sum_{t=1}^n \varepsilon_t^* r_t}{\sum_{t=1}^n r_t^2}.$$

The implication is that the profile, or concentrated, likelihood is then

$$\log f(y|\widehat{\varepsilon_0};\theta) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^n \varepsilon_t^{*2} + \frac{1}{2\sigma^2}\widehat{\varepsilon_0}^2\sum_{t=1}^n r_t^2. \tag{6}$$

1.·4 *Exact likelihood*

Instead of constructing a profile likelihood function we could just integrate out ε_0 in order to construct the exact likelihood function. Thus

$$f(y;\theta) = \int f(y|s;\theta) f_{\varepsilon_0;\sigma^2}(s)ds, \qquad \text{where} \qquad f_{\varepsilon_0;\sigma^2}(s) = N(0,\sigma^2).$$

An easy way of carrying this out is via the equality

$$\log f(y;\theta) = \log f(y|\varepsilon_0 = \varphi;\theta) + \log f_{\varepsilon_0;\sigma^2}(\varphi) - \log f_{\varepsilon_0|y;\theta}(\varphi).$$

It is convenient to use the special case of where $\varphi = 0$. The only term which is non-trivial is the posterior density of $\varepsilon_0|y;\theta$. But keeping θ fixed, (5) is the corresponding likelihood for this posterior, while the prior is $\varepsilon_0 \sim N(0,\sigma^2)$. Therefore straightforward use of Bayes theorem yields

$$\varepsilon_0|y;\theta \sim N(\mu_p,\sigma_p^2), \qquad \sigma_p^2 = \frac{\sigma^2}{1+\sum_{t=1}^n r_t^2}, \qquad \mu_p = -\frac{\sum_{t=1}^n \varepsilon_t^* r_t}{1+\sum_{t=1}^n r_t^2}.$$

That is μ_p is a Bayesian regression mean of ε_t^* on $-r_t$ with a measurement variance of σ^2 and a prior $\varepsilon_0 \sim N(0,\sigma^2)$.

This gives

$$\begin{aligned}\log f(y;\theta) &= -\frac{n+1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^n \varepsilon_t^{*2} + \frac{1}{2}\log 2\pi\sigma_p^2 + \frac{\mu_p^2}{2\sigma_p^2} \\ &= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^n \varepsilon_t^{*2} + \frac{1}{2}\log\left(1+\sum_{t=1}^n r_t^2\right) + \frac{1}{2\sigma^2}\mu_p^2\left(1+\sum_{t=1}^n r_t^2\right).\end{aligned}$$

Hence the likelihood is a simple correction of the CSS. It is also algebraically very similar to the profile likelihood function (6).

1.·5 *Score*

Likewise the score can be computed using a result due to Louis (1982), which puts the problem in an EM algorithm framework suggested by Dempster et al. (1977). Define

$$Q(\theta,\theta_1) = E\log f(y|\varepsilon_0;\theta),$$

where the expectation is with respect to $\varepsilon_0|y_1,...,y_n;\theta_1$. Then

$$Q(\theta,\theta_1)=-\frac{n}{2}\log 2\pi\sigma^2-\frac{1}{2\sigma^2}\left\{\sum_{t=1}^{n}\varepsilon_t^{*2}+2\mu_p\sum_{t=1}^{n}\varepsilon_t^{*}r_t+E(\varepsilon_0^2)\sum_{t=1}^{n}r_t^2\right\}. \tag{7}$$

This is helpful as

$$\begin{aligned}\left.\frac{\partial\log f(y;\theta)}{\partial\theta}\right|_{\theta=\theta_1} &= \left.\frac{\partial Q(\theta,\theta_1)}{\partial\theta}\right|_{\theta=\theta_1}\\ &= -\frac{1}{\sigma^2}\sum_{t=1}^{n}\left[\left\{\frac{\partial\varepsilon_t^*}{\partial\theta}+\frac{\partial r_t}{\partial\theta}\mu_p\right\}\varepsilon_t^*+\left\{\frac{\partial r_t}{\partial\theta}E(\varepsilon_0^2)+\frac{\partial\varepsilon_t^*}{\partial\theta}\mu_p\right\}r_t\right].\end{aligned}$$

The required derivatives can be computed using the recursion

$$\left(\frac{\partial\varepsilon_t^*}{\partial\theta},\frac{\partial r_t}{\partial\theta}\right)=-\left(\varepsilon_t^*,r_t\right)-\theta\left(\frac{\partial\varepsilon_{t-1}^*}{\partial\theta},\frac{\partial r_{t-1}}{\partial\theta}\right),\quad \frac{\partial\varepsilon_0^*}{\partial\theta}=0,\frac{\partial r_0}{\partial\theta}=0.$$

The score can be used in a quasi-Newton algorithm or the likelihood could be maximized via an EM algorithm using (7). The second of these is particularly attractive in this context as the second derivative could be approximated by

$$-\frac{1}{\sigma^2}\sum_{t=1}^{n}\left\{\left(\frac{\partial\varepsilon_t^*}{\partial\theta}+\frac{\partial r_t}{\partial\theta}\mu_p\right)\left(\frac{\partial\varepsilon_t^*}{\partial\theta}+\frac{\partial r_t}{\partial\theta}\mu_p\right)'+\sigma_p^2\frac{\partial r_t}{\partial\theta}\frac{\partial r_t}{\partial\theta'}\right\}.$$

Then, writing $\theta_{(0)}$ as some initial value of a numerical optimization routine, the M-step could be performed via

$$\begin{aligned}\theta_{(i)} &= \theta_{(i-1)}-\left\{\sum_{t=1}^{n}\left(\frac{\partial\varepsilon_t^*}{\partial\theta}+\frac{\partial r_t}{\partial\theta}\mu_p\right)\left(\frac{\partial\varepsilon_t^*}{\partial\theta}+\frac{\partial r_t}{\partial\theta}\mu_p\right)'+\sigma_p^2\frac{\partial r_t}{\partial\theta}\frac{\partial r_t}{\partial\theta'}\right\}^{-1} \\ &\quad \sum_{t=1}^{n}\left\{\left(\frac{\partial\varepsilon_t^*}{\partial\theta}+\frac{\partial r_t}{\partial\theta}\mu_p\right)(\varepsilon_t^*+\mu_p r_t)+\left(\frac{\partial r_t}{\partial\theta}\sigma_p\right)r_t\sigma_p\right\}.\end{aligned} \tag{8}$$

This generalizes (3).

In general there is no need to iterate this numerical optimization until convergence for each EM step. As Dempster et al. (1977) noted to maximize the likelihood function the M-step only needs to find a value of θ such that $Q(\theta,\theta_1)>Q(\theta_1,\theta_1)$, where θ_1 is value of the parameter use in the E-step. Dempster et al. (1977) called such a procedure the generalized EM (GEM) algorithm and it is particularly convenient here for there seems little to be gained by iterating (8) until precise convergence in the first couple of GEM iterations.

In this setup of the EM algorithm, ε_0 is used as the artificial missing data. Given the CSS is known to be close to the ML estimator, the information in this missing data must be small and so the EM algorithm should converge quickly.

1.·6 *Numerical example*

To illustrate the steps of the EM algorithm I simulated a moving average with $n = 20, \theta = -0.7$, $\sigma^2 = 1$. Throughout I assumed σ^2 is known. The ML estimator of θ is `-0.74271`, while I used an initial value, for the optimisation, of `0.5`. The results are displayed in Table 1. It shows small changes in the parameters as the procedure is iterated with the likelihood increasing with each iteration.

Iteration	$\theta_{(0)}$	$\log L(\theta_{(0)})$	score
0	-0.73686	-112.40	-0.34519
Iteration, i	$\theta_{(i)}$	$\log L(\theta_{(i)})/L(\theta_{(i-1)})$	score
1	-0.74186	0.00099079	-0.050280
2	-0.74258	2.0821e-005	-0.0073666
3	-0.74269	4.4631e-007	-0.0010802
4	-0.74270	9.5941e-009	-0.00015841
5	-0.74271	2.0634e-010	-2.3231e-005
6	-0.74271	4.4054e-012	-3.4070e-006
7	-0.74271	1.1369e-013	-4.9981e-007
8	-0.74271	0.00000	-7.3367e-008

Table 1: *Iteration of the EM algorithm. For the first iteration I set $\mu_p = 0$ and $\sigma_p = 0$, so it gives the CSS estimator. The first figure of the third column is the likelihood at $\theta_{(i)}$. For other iterations I display the likelihood improvement over the previous iteration.*

2. GENERALIZATION

2.·1 *The model*

This analysis generalizes to more complicated pure moving averages in a straightforward way. Here I look at the problem of dealing with autoregressive moving average (ARMA) models of order p, q where

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t + \lambda_1 \varepsilon_{t-1} + ... + \lambda_q \varepsilon_{t-q},$$

where stationarity and invertibility is imposed on the model. Throughout I will write θ to denote the vector of parameters. It will be convenient to write this model in companion form

$$y_t = (1, 0, ..., 0)\, \alpha_t = z\alpha_t, \quad m = \max(p+1, q),$$

with

$$\alpha_{t+1} = \begin{pmatrix} y_t \\ \phi_2 y_{t-1} + ... + \phi_p y_{t-p+1} + \lambda_1 \varepsilon_t + ... + \lambda_{m-1} \varepsilon_{t-m+2} \\ \phi_3 y_{t-1} + ... + \phi_p y_{t-p+2} + \lambda_2 \varepsilon_t + ... + \lambda_{m-1} \varepsilon_{t-m+3} \\ \vdots \\ \phi_m y_{t-1} + \lambda_{m-1} \varepsilon_t \end{pmatrix}$$

$$= \begin{pmatrix} \phi_1 & 1 & 0 & \cdots & 0 \\ \phi_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{m-1} & 0 & 0 & \cdots & 1 \\ \phi_m & 0 & 0 & & 0 \end{pmatrix} \alpha_t + \begin{pmatrix} 1 \\ \lambda_1 \\ \vdots \\ \lambda_{m-2} \\ \lambda_{m-1} \end{pmatrix} \varepsilon_t = T\alpha_t + h\varepsilon_t.$$

This representation is often used when placing a model into state space form in order to compute the likelihood function of ARMA models (see, for example, Harvey (1993, p. 96)). Now under the assumption that y_t is stationary, α_0 has a mean of zero with a covariance of $\sigma^2\Sigma_\alpha$, while writing $G = (I - T \otimes T)^{-1}$, then $vec\{\Sigma_\alpha\} = Gvec(hh') = G(h \otimes h)$.

The likelihood can be computed in a number of ways, including the Kalman filter. A recent paper which references the literature on evaluating the likelihood of ARMA models is Mauricio (1995) who focuses on the multivariate case.

2.·2 *Conditional sum of squares*

Now

$$\log f(y|\alpha_0 = 0) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\varepsilon_t^{*2},$$

where

$$\varepsilon_t^* = y_t - za_{t|t-1}, \qquad \text{and} \qquad a_{t|t-1} = Ta_{t-1|t-2} + k\varepsilon_{t-1}^*. \tag{9}$$

Here

$$k = Th, \qquad \text{and} \qquad a_{1|0} = 0.$$

This is a special case of the Kalman filter exploiting the fact that there is need to carry out the Riccati equation as α_t is a linear combination of $\alpha_0, y_1, ..., y_t$.

Clearly, $\partial\varepsilon_t^*/\partial\theta_i = -z\partial a_{t|t-1}/\partial\theta_i$, where

$$\frac{\partial a_{t|t-1}}{\partial\theta_i} = \frac{\partial T}{\partial\theta_i}a_{t-1|t-2} + T\frac{\partial a_{t-1|t-2}}{\partial\theta_i} + \frac{\partial k}{\partial\theta_i}\varepsilon_{t-1}^* + k\frac{\partial\varepsilon_{t-1}^*}{\partial\theta_i}. \tag{10}$$

Here $\partial a_{1|0}/\partial\theta_i = 0$. Then the score is

$$\frac{\partial \log f(y|\alpha_0 = 0;\theta)}{\partial\theta} = -\frac{1}{\sigma^2}\sum_{t=1}^{n}\frac{\partial\varepsilon_t^*}{\partial\theta}\varepsilon_t^*,$$

while it is straightforward to compute an approximation to the observed information as

$$-\frac{1}{\sigma^2}\sum_{t=1}^{n}\frac{\partial\varepsilon_t^*}{\partial\theta}\frac{\partial\varepsilon_t^*}{\partial\theta'}.$$

2.·3 *Likelihood*

Define $\varepsilon_t^+ = \varepsilon_t^* + \alpha_0' V_t$ and $v_t' = (\varepsilon_t^*, V_t')$. Then v_t is calculated via

$$\begin{aligned} v_t' &= \left(y_t, \widetilde{0}\right) - zB_{t|t-1}, \qquad t = 1, ..., n \\ B_{t|t-1} &= TB_{t-1|t-2} + kv_{t-1}', \qquad t = 2, ..., n, \end{aligned} \tag{11}$$

where $B_{1|0} = (T, 0)$. We record only

$$\begin{pmatrix} s\varepsilon_n & s_n' \\ s_n & S_n \end{pmatrix} = \sum_{t=1}^n v_t v_t'.$$

This implies the conditional likelihood function is

$$\begin{aligned} \log f(y|\alpha_0;\theta) &= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^n \varepsilon_t^{+2} \\ &= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left(\sum_{t=1}^n \varepsilon_t^{*2} + 2\alpha_0' \sum_{t=1}^n \varepsilon_t^* V_t + \alpha_0' \sum_{t=1}^n V_t V_t' \alpha_0\right) \\ &= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left(s\varepsilon_n + 2\alpha_0' s_n + \alpha_0' S_n \alpha_0\right). \end{aligned}$$

Straightforwardly, if we regard α_0 as an unknown parameter then the ML estimator of α_0 is the regression of ε_t^* on $-V_t$, which is $\widehat{\alpha_0} = -S_n^{-1} s_n$. The resulting profile likelihood function is

$$\log f(y|\widehat{\alpha_0};\theta) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2} s\varepsilon_n + \frac{1}{2\sigma^2} s_n' S_n^{-1} s_n.$$

The exact likelihood function can be computed by using a Bayesian regression

$$\alpha_0|y \sim N(\mu_{\alpha|y}, \sigma^2\Sigma_{\alpha|y}), \quad \Sigma_{\alpha|y} = \left(\Sigma_\alpha^{-1} + S_n\right)^{-1}, \quad \mu_{\alpha|y} = -\Sigma_{\alpha|y} s_n, \quad \Theta_{\alpha|y} = \Sigma_{\alpha|y} + \mu_{\alpha|y}\mu_{\alpha|y}'.$$

Then as the likelihood is $f(y) = f(y|\alpha_0 = \varphi) f_{\alpha_0}(\varphi)/f_{\alpha_0|y}(\varphi)$ for any value of φ, we can set $\varphi = 0$. This gives the expression

$$\begin{aligned} \log f(y;\theta) &= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2} s\varepsilon_n - \frac{1}{2}\log|\sigma^2\Sigma_\alpha| + \frac{1}{2\sigma^2} s_n' \Sigma_{\alpha|y} s_n + \frac{1}{2}\log|\sigma^2\Sigma_{\alpha|y}| \\ &= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2} s\varepsilon_n - \frac{1}{2}\log|\Sigma_\alpha| + \frac{1}{2\sigma^2} s_n' \Sigma_{\alpha|y} s_n + \frac{1}{2}\log|\Sigma_{\alpha|y}|. \end{aligned}$$

2.·4 *Score*

The evaluation of the score requires the computation of

$$\partial \begin{pmatrix} s\varepsilon_n & s_n' \\ s_n & S_n \end{pmatrix} /\partial\theta_i = \sum_{t=1}^n \left(\frac{\partial v_t}{\partial\theta_i} v_t' + v_t \frac{\partial v_t'}{\partial\theta_i}\right),$$

where $\partial v_t'/\partial\theta_i = -z\partial B_{t|t-1}/\partial\theta_i$. Then starting with $\partial B_{1|0}/\partial\theta_i = 0, \partial T/\partial\theta_i$

$$B_{t|t-1}^* = \left(\frac{\partial B_{t|t-1}}{\partial\theta_1}, ..., \frac{\partial B_{t|t-1}}{\partial\theta_p}\right) = T^* B_{t-1|t-2} + TB_{t-1|t-2}^* + k^* v_{t-1}' + kv_{t-1}^{*\prime},,$$

where

$$T^* = \left(\frac{\partial T}{\partial \theta_1}, ..., \frac{\partial T}{\partial \theta_w} \right), \quad k^* = \left(\frac{\partial k}{\partial \theta_1}, ..., \frac{\partial k}{\partial \theta_w} \right), \quad v_{t-1}^{*\prime} = \left(\frac{\partial v_{t-1}^{\prime}}{\partial \theta_1}, ..., \frac{\partial v_{t-1}^{\prime}}{\partial \theta_w} \right).$$

Here w is the dimension of θ. Notice in pure moving average models $T^* = 0$, while for autoregressions $k^* = T^* g$. In all cases, T^* and T will be very sparse matrices.

The score, $\partial \log f / \partial \theta_i$, can be expressed analytically as

$$\begin{aligned} & -\frac{1}{2\sigma^2} \left[\begin{array}{c} \frac{\partial s\varepsilon_n}{\partial \theta_i} - 2s_n^{\prime} \Sigma_{\alpha|y} \frac{\partial s_n}{\partial \theta_i} + s_n^{\prime} \Sigma_{\alpha|y} \left(\frac{\partial \Sigma_\alpha^{-1}}{\partial \theta_i} + \frac{\partial S_n}{\partial \theta_i} \right) \Sigma_{\alpha|y} s_n + \frac{\partial \log |\Sigma_\alpha|}{\partial \theta_i} \\ + tr \left\{ \Sigma_{\alpha|y} \left(\frac{\partial \Sigma_\alpha^{-1}}{\partial \theta_i} + \frac{\partial S_n}{\partial \theta_i} \right) \right\} \end{array} \right] \qquad (12) \\ = & -\frac{1}{2\sigma^2} \left[\begin{array}{c} \frac{\partial s\varepsilon_n}{\partial \theta_i} + 2\mu_{\alpha|y}^{\prime} \frac{\partial s_n}{\partial \theta_i} + \mu_{\alpha|y}^{\prime} \left(\frac{\partial \Sigma_\alpha^{-1}}{\partial \theta_i} + \frac{\partial S_n}{\partial \theta_i} \right) \mu_{\alpha|y} + \frac{\partial \log |\Sigma_\alpha|}{\partial \theta_i} \\ + tr \left\{ \Sigma_{\alpha|y} \left(\frac{\partial \Sigma_\alpha^{-1}}{\partial \theta_i} + \frac{\partial S_n}{\partial \theta_i} \right) \right\} \end{array} \right]. \end{aligned}$$

The derivatives of Σ_α^{-1} and $\log |\Sigma_\alpha|$ can be derived through

$$\begin{aligned} \frac{\partial vec \{ Var(\alpha_0) \}}{\partial \theta_i} &= -G \frac{\partial \left(I - T \otimes T \right)}{\partial \theta_i} G vec(hh^{\prime}) + G \frac{\partial h \otimes h}{\partial \theta_i} \\ &= G \left(\frac{\partial T}{\partial \theta_i} \otimes T + T \otimes \frac{\partial T}{\partial \theta_i} \right) \Sigma_\alpha + G \left(\frac{\partial h}{\partial \theta_i} \otimes h + h \otimes \frac{\partial h}{\partial \theta_i} \right) \end{aligned}$$

as (see, for example, Magnus & Neudecker (1988))

$$\frac{\partial \Sigma_\alpha^{-1}}{\partial \theta_i} = -\Sigma_\alpha \frac{\partial \Sigma_\alpha}{\partial \theta_i} \Sigma_\alpha, \qquad \text{and} \qquad \frac{\partial |\Sigma_\alpha|}{\partial \theta_i} = tr \left\{ \Sigma_\alpha^{-1} \frac{\partial \Sigma_\alpha}{\partial \theta_i} \right\}.$$

ACKNOWLEDGEMENTS

Neil Shephard thanks Siem Jan Koopman for comments on an earlier version of this note. He also thanks the ESRC for their financial support through the project 'Estimation via Simulation in Econometrics' and the EU through their grant 'Econometric inference using Simulation Techniques'.

References

Box, G. E. P. & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco, CA, 2nd edition.

de Jong, P. (1989). Smoothing and interpolation with the state space model. *J. Am. Statist. Assoc.* **84**, 1085–8.

Dempster, A. P., Laird, N. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* **39**, 1–38.

Harvey, A. C. (1993). *Time Series Models.* Harvester Wheatsheaf, Hemel Hempstead, 2nd edition.

Koopman, S. J. & Shephard, N. (1992). Exact score for time series models in state space form. *Biometrika* **79**, 823–6.

Louis, T. A. (1982). Finding observed information using the EM algorithm. *J. R. Statist. Soc. B* **44**, 98–103.

Magnus, J. R. & Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley, New York.

Mauricio, J. A. (1995). Exact maximum likelihood of stationary vector ARMA models. *J. Am. Statist. Assoc.* **90**, 282–291.