Biometrika (1992), **79**, 4, pp. 823–6
Printed in Great Britain

Miscellanea

Exact score for time series models in state space form

BY SIEM JAN KOOPMAN
Department of Statistical and Mathematical Sciences, London School of Economics and Political Science, London WC2A 2AE, U.K.

AND NEIL SHEPHARD
Nuffield College, Oxford OX1 1NF, U.K.

SUMMARY

This paper shows that the score vector for Gaussian state space models takes on a simple form which can be computed in a single pass of the Kalman filter and a smoother.

Some key words: EM algorithm; Kalman filter; Smoothing; Unobserved component.

1. INTRODUCTION

The Gaussian state space form (Harvey, 1989, Ch. 3) is given by

$$y_t = Z_t\alpha_t + \varepsilon_t, \quad \alpha_t = T_t\alpha_{t-1} + \eta_t \quad (t = 1, \ldots, n),$$

$$\alpha_0 \sim N(a_0, P_0), \quad \varepsilon_t \sim N(0, H_t), \quad \eta_t \sim N(0, Q_t),$$

where α_0, (ε_t) and (η_s) are independent of one another for all t and s and the initial conditions a_0 and P_0 are known. The Kalman filter can be used to deliver the log likelihood which in turn means the score can be evaluated by numerical differentiation.

Engle & Watson (1981) show how to construct a set of filters for computing the exact score analytically, however their approach is cumbersome, difficult to program and typically much more expensive to use than numerically differentiating the likelihood. As a result it is rarely used in practice. In this paper we shall show that the whole score vector can be computed exactly in a single pass of the Kalman filter and a smoother. For many classes of models this markedly increases the speed of algorithms for the numerical maximization of the likelihood (Gill, Murray & Wright, 1981).

The rest of this paper has three sections. Section 2 discusses the general theory behind the evaluation of the score, while § 3 looks at an important special case. The fourth concludes.

2. COMPUTING THE SCORE

Work on the EM algorithm can be coupled with recently developed smoothing algorithms to evaluate the score. To see this we will write

$$Y_n = (y_1', \ldots, y_n')', \quad \alpha = (\alpha_0', \ldots, \alpha_n')'$$

and explicitly express the dependence of Z_t, T_t, H_t and Q_t on θ, the unknown parameters of the model, by writing $Z_t(\theta)$, $T_t(\theta)$, $H_t(\theta)$ and $Q_t(\theta)$. Then using familiar results, and omitting constants, we have for all α

$$\begin{aligned}\log f(Y_n;\theta) &= \log f(Y_n\mid\alpha;\theta)+\log f(\alpha;\theta)-\log f(\alpha\mid Y_n;\theta)\\ &= \sum_{t=1}^{n}\{\log f(y_t\mid\alpha_{t-1};\theta)+\log f(\alpha_t\mid\alpha_{t-1};\theta)\}+\log f(\alpha_0)-\log f(\alpha\mid Y_n;\theta)\\ &= -\frac{1}{2}\sum_{t=1}^{n}(\log|H_t(\theta)|+\log|Q_t(\theta)|)-\frac{1}{2}\sum_{t=1}^{n}\operatorname{tr}[H_t(\theta)^{-1}\{y_t-Z_t(\theta)\alpha_t\}\{y_t-Z_t(\theta)\alpha_t\}']\\ &\quad -\frac{1}{2}\sum_{t=1}^{n}\operatorname{tr}[Q_t(\theta)^{-1}\{\alpha_t-T_t(\theta)\alpha_{t-1}\}\{\alpha_t-T_t(\theta)\alpha_{t-1}\}']\\ &\quad -\tfrac{1}{2}\log|P_0|-\tfrac{1}{2}(\alpha_0-a_0)'P_0^{-1}(\alpha_0-a_0)-\log f(\alpha\mid Y_n;\theta).\end{aligned}$$

To be able to derive the score at a point θ^* we will first integrate both sides with respect to $f(\alpha\mid Y_n;\theta^*)$, the joint density of the smoother, to get

$$\log f(Y_n;\theta)=Q(\theta,\theta^*)+R(\theta^*)-H(\theta,\theta^*), \tag{2·1}$$

where

$$\begin{aligned}Q(\theta,\theta^*) &= -\frac{1}{2}\sum_{t=1}^{n}\{\log|H_t(\theta)|+\log|Q_t(\theta)|\}-\frac{1}{2}\sum_{t=1}^{n}\operatorname{tr}[H_t(\theta)^{-1}\{e_{t|n}e'_{t|n}+Z_t(\theta)P_{t|n}Z_t(\theta)'\}]\\ &\quad -\frac{1}{2}\sum_{t=1}^{n}\operatorname{tr}[Q_t(\theta)^{-1}\{n_{t|n}n'_{t|n}+P_{t|n}-T_t(\theta)P_{t-1,t|n}-P_{t,t-1|n}T_t(\theta)'+T_t(\theta)P_{t-1|n}T_t(\theta)'\}],\end{aligned}$$

$$R(\theta^*)=\int\{-\tfrac{1}{2}\log|P_0|-(\alpha_0-a_0)'P_0^{-1}(\alpha_0-a_0)\}f(\alpha\mid Y_n;\theta^*)\,d\alpha,$$

$$H(\theta,\theta^*)=\int\log f(\alpha\mid Y_n;\theta)f(\alpha\mid Y_n;\theta^*)\,d\alpha,$$

where $a_{t|n}$, $e_{t|n}$ and $n_{t|n}$ are the smoothed estimates of α_t, ε_t and η_t respectively, $P_{t|n}$ is the mean squared error of $a_{t|n}$, and $P_{t,t-1|n}$ is the covariance between the estimators of α_t and α_{t-1}. All these smoothed quantities are computed with θ taken to be θ^* by first running a Kalman filter which is given by

$$\begin{gathered}a_{t+1|t}=T_{t+1}a_{t|t-1}+K_tv_t,\quad P_{t+1|t}=T_{t+1}P_{t|t-1}L_t+Q_{t+1},\\ v_t=y_t-Z_ta_{t|t-1},\quad F_t=Z_tP_{t|t-1}Z'_t+H_t,\\ K_t=T_{t+1}P_{t|t-1}Z'_tF_t^{-1},\quad L_t=T_{t+1}-K_tZ_t\quad(t=1,\ldots,n).\end{gathered} \tag{2·2}$$

Then the de Jong (1989) and de Jong & MacKinnon (1988) smoothing algorithms deliver

$$\begin{gathered}e_{t|n}=y_t-Z_ta_{t|n},\quad n_{t|n}=a_{t|n}-T_ta_{t-1|n},\\ a_{t|n}=a_{t|t-1}+P_{t|t-1}r_{t-1},\quad P_{t|n}=P_{t|t-1}-P_{t|t-1}N_{t-1}P_{t|t-1},\\ r_{t-1}=Z'_tF_t^{-1}v_t+L'_tr_t,\quad N_{t-1}=Z'_tF_t^{-1}Z_t+L'_tN_tL_t,\\ P_{t-1,t|n}=P_{t-1|t-2}L'_{t-1}(I-N_{t-1}P_{t|t-1})\quad(t=n-1,\ldots,1).\end{gathered} \tag{2·3}$$

If (2·1) is differentiated with respect to θ, then

$$\frac{\partial\log f(Y_n;\theta)}{\partial\theta}=\frac{\partial Q(\theta,\theta^*)}{\partial\theta}-\frac{\partial H(\theta,\theta^*)}{\partial\theta},$$

but it can be shown that

$$\left[\frac{\partial H(\theta,\theta^*)}{\partial\theta}\right]_{\theta=\theta^*}=0$$

and so

$$\left[\frac{\partial\log f(Y_n;\theta)}{\partial\theta}\right]_{\theta=\theta^*}=\left[\frac{\partial Q(\theta,\theta^*)}{\partial\theta}\right]_{\theta=\theta^*}.$$

To illustrate the use of this result consider the first order autoregression observed with error

$$y_t = \mu_t + \varepsilon_t, \quad \mu_t = \rho\mu_{t-1} + \eta_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \quad \eta_t \sim N(0, \sigma_\eta^2),$$

so that $\theta = (\rho, \sigma_\varepsilon^2, \sigma_\eta^2)'$. The score vector has elements

$$\frac{\partial \log f(Y_n; \rho^*, \sigma_\varepsilon^{*2}, \sigma_\eta^{*2})}{\partial \sigma_\varepsilon^2} = -\frac{n}{2\sigma_\varepsilon^{*2}} + \frac{\sum (e_{t|n}^2 + p_{t|n})}{2\sigma_\varepsilon^{*4}},$$

$$\frac{\partial \log f(Y_n; \rho^*, \sigma_\varepsilon^{*2}, \sigma_\eta^{*2})}{\partial \sigma_\eta^2} = -\frac{n}{2\sigma_\eta^{*2}} + \frac{\sum (n_{t|n}^2 + p_{t|n} - 2\rho^* p_{t,t-1|n} + \rho^{*2} p_{t-1|n})}{2\sigma_\eta^{*4}},$$

$$\frac{\partial \log f(Y_n; \rho^*, \sigma_\varepsilon^{*2}, \sigma_\eta^{*2})}{\partial \rho} = -\frac{1}{2} \frac{\Sigma 2\rho^* p_{t-1|n} - 2p_{t-1,t|n} - 2a_{t-1|n}(a_{t|n} - \rho^* a_{t-1|n})}{\sigma_\eta^{*2}}.$$

Although these expressions are analytically elegant, in general evaluating the score this way may be slower than numerically differentiating the likelihood. This is because the smoother involves the computation of a considerable number of matrix multiplications, which will be time consuming if the dimension of the state is not short.

3. Special case

Suppose we can partition θ into ψ and λ, where ψ and λ refer to parameters in Z_t and T_t, and H_t and Q_t respectively: in our experience this is almost always possible. Then we are able to find a rapid way of determining the elements of the score vector which corresponds to λ only. Writing $\theta = (\psi^{*\prime}, \lambda')'$ and $\theta^* = (\psi^{*\prime}, \lambda^{*\prime})'$, we have that $Q(\theta, \theta^*)$ becomes

$$Q(\theta, \theta^*) = -\frac{1}{2}\sum_{t=1}^{n} \log|H_t(\lambda)| - \frac{1}{2}\sum_{t=1}^{n} \operatorname{tr}[H_t(\lambda)^{-1}\{e_{t|n}e_{t|n}' + \operatorname{cov}(\varepsilon_t - e_{t|n})\}]$$
$$-\frac{1}{2}\sum_{t=1}^{n} \log|Q_t(\lambda)| - \frac{1}{2}\sum_{t=1}^{n} \operatorname{tr}[Q_t(\lambda)^{-1}\{n_{t|n}n_{t|n}' + \operatorname{cov}(\eta_t - n_{t|n})\}],$$

where the covariance is computed over the density $\alpha \mid Y_n; \theta^*$. The advantage of this formulation is that Koopman's (1993) disturbance smoother can be used to compute $e_{t|n}$, $n_{t|n}$, $\operatorname{cov}(\varepsilon_t - e_{t|n})$ and $\operatorname{cov}(\eta_t - n_{t|n})$. This has

$$e_{t|n} = H_t(\lambda^*)e_t, \quad n_{t|n} = Q_t(\lambda^*)r_{t-1},$$

$$\operatorname{cov}(\varepsilon_t - e_{t|n}) = H_t(\lambda^*) - H_t(\lambda^*)D_t H_t(\lambda^*), \quad \operatorname{cov}(\eta_t - n_{t|n}) = Q_t(\lambda^*) - Q_t(\lambda^*)N_{t-1}Q_t(\lambda^*),$$

where these terms are calculated by starting with $r_n = 0$ and $N_n = 0$ and the backwards recursions

$$e_t = F_t^{-1}v_t - K_t' r_t, \quad r_{t-1} = Z_t' F_t^{-1} v_t + L_t' r_t,$$
$$D_t = F_t^{-1} + K_t' N_t K_t, \quad N_{t-1} = Z_t' F_t^{-1} Z_t + L_t' N_t L_t. \tag{3·1}$$

The storage requirement for this smoother is very minor for it needs just v_t, F_t^{-1} and K_t, while the focusing on ψ avoids the necessity to compute $P_{t,t-1|n}$ altogether. Once again throughout the recursions, and the associated Kalman filter, θ is taken to be θ^*. This implies the simple result that

$$\frac{\partial \log f(Y_n; \theta^*)}{\partial \lambda_i} = \frac{1}{2}\sum_{t=1}^{n} \operatorname{tr}\left\{(e_t e_t' - D_t)\frac{\partial H_t(\lambda^*)}{\partial \lambda_i}\right\} + \frac{1}{2}\sum_{t=1}^{n} \operatorname{tr}\left\{(r_{t-1}r_{t-1}' - N_{t-1})\frac{\partial Q_t(\lambda^*)}{\partial \lambda_i}\right\}, \tag{3·2}$$

which is easy to program, requiring only (3·1) and (3·2) to be added to the Kalman filter (2·2).

Although (3·2) delivers only the part of the score which corresponds to λ, the rest can be filled in by numerical differentiation of the likelihood in the usual way. If the model is rich in parameters in the H_t and Q_t matrices, as for instance the unobserved components models of West & Harrison (1989) and Harvey (1989) are, then the computational savings which result from using this algorithm will be substantial. Our general experience with Koopman's disturbance smoother suggests that it takes about the processing time of a single pass of the Kalman filter. So all the elements of the score corresponding to λ can be computed exactly in the equivalent of two Kalman filters. For multivariate models the computational savings will be important; a four variable local level model (Harvey, 1989, pp. 429-32) has 20 parameters in H_t and Q_t and so the score can be computed more rapidly and accurately using this method than by conventional techniques. Our experience with a number of data sets suggests that, for the time series models we use, numerical maximization routines tend to converge in a marginally smaller number of iterations when using exact, rather than approximate, scores and have very slightly higher likelihoods when they have converged. However, we do not claim either of these results holds in general.

4. CONCLUSION

Recursions for the exact computation of the score of a model in state space form are derived. The first set, developed in § 2, is based on the state smoother and so tends to be slow if the dimension of the state is large. The use of a disturbance smoother allows the computations to be speeded up dramatically. The recursions (3·1) and (3·2) should provide a simple way of computing the score for many models and should be particularly useful in tackling multivariate models.

ACKNOWLEDGEMENT

The work reported here was supported by an ESRC grant. We would like to thank Piet de Jong and Andrew Harvey for their comments on an earlier version of this paper.

REFERENCES

DE JONG, P. (1989). Smoothing and interpolation with the state space model. *J. Am. Statist. Assoc.* **84**, 1085-8.
DE JONG, P. & MACKINNON, M. J. (1988). Covariances for smoothed estimates in state space models. *Biometrika* **75**, 601-2.
ENGLE, R. F. & WATSON, M. W. (1981). A one-factor multivariate time series model of metropolitan wage rates. *J. Am. Statist. Assoc.* **76**, 774-81.
GILL, P. E., MURRAY, W. & WRIGHT, M. H. (1981). *Practical Optimization.* Cambridge: Academic Press.
HARVEY, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press.
KOOPMAN, S. J. (1993). Disturbance smoother for state space models. *Biometrika* **80**. To appear.
WEST, M. & HARRISON, J. (1989). *Bayesian Forecasting and Dynamic Models.* New York: Springer-Verlag.

[*Received May* 1992. *Revised June* 1992]