Institute of Mathematical Statistics

**LECTURE NOTES — MONOGRAPH SERIES**

# Optimal Estimating Equations for State Vectors in Non-Gaussian and Nonlinear State Space Time Series Models

J. Durbin
London School of Economics and Political Science

**ABSTRACT**

In state space times series models the development over time of the observed series is determined by an unobserved series of state vectors. The paper considers the estimation of these vectors by the mode of the posterior distribution of the state vectors given the data. It is shown that the estimates are the solution of an optimal unbiased estimating equation.

**Key words:** Nonlinear time series; non-Gaussian time series; posterior mode estimates; estimating functions.

## 1 Introduction

State space models are a very general class of models which are increasingly used in applied time series analysis. In such models we have a series $y_1, \ldots, y_n$ of vector observations, a series $\alpha_1, \ldots, \alpha_n$ of unobserved state vectors and a vector $\psi$ of parameters which we assume to be known or to have been estimated efficiently. This paper is concerned with the problem of estimating $\alpha_1, \ldots, \alpha_n$ given the observations $y_1, \ldots, y_n$.

Most of the work that has been done on such models hitherto has been based essentially on the linear Gaussian case. See for example the book by Harvey (1989) for a comprehensive treatment of linear Gaussian state space models. However, for many practical applications the assumptions of linearity and normality seem inappropriate. For example, if the data consist of the number of car drivers killed per month in road accidents in a particular region, the Poisson distribution would seem to provide a more appropriate model for the data than the normal distribution. Similarly, if the observations appear to come from distributions with heavy tails, as is common with economic and many other types of data, the t-distribution with a low number of degrees of frequency would seem a more appropriate model than the normal distribution. A further desirable relaxation is to allow departures from linearity in the model.

For the linear Gaussian case the standard estimates of the $\alpha_t$'s are their conditional expectations given the $y_t$'s. As is to be expected, these have an unbiased minimum variance property. For the non-Gaussian or non-linear cases, however, the problem of calculating these conditional expectations by analytical methods is intractable. An alternative considered by some authors is to use the mode instead of the mean of the conditional density of $[\alpha_1', ..., \alpha_n']'$ given the $y_t$'s since it is easier to handle. While this approach is intuitively attractive since the resulting estimates are the most probable values of the $\alpha_t$'s given the observations, it does not lead to estimation errors which are unbiased with minimum variance matrix. However, we shall show in this paper that the estimates have an analogous property, namely that they are the solution of unbiased estimating equations with minimum variance matrix. Our results are derived from the estimating equations approach to the estimation of fixed parameters of Godambe (1960) and Durbin (1960). They are also related to results of Ferreira (1982) on the application of estimating equation theory to the estimation of a single random parameter.

The next section begins by considering the standard linear Gaussian state space model and uses this as a basis for discussing the classes of non-Gaussian and nonlinear models considered in the paper. It introduces the idea of estimating the $\alpha_t$'s by their posterior mode and obtains an estimating equation for it in an appropriate form. Section 3 derives optimal estimating equations for models of the kind under consideration and shows that the estimating equation for the posterior mode belongs to this class.

# 2 State Space Models for Non-Gaussian and Nonlinear Time Series

The purpose of this section is to outline a broad class of models to which the results of the next section apply. Our starting point is the standard linear Gaussian state space model for an observed vector time series $y_1, ..., y_n$, namely

$$y_t = Z_t\alpha_t + \epsilon_t, \qquad \epsilon_t \sim N(0, H_t) \tag{2.1}$$

$$\alpha_t = T_t\alpha_{t-1} + R_t\eta_t, \qquad \eta_t \sim N(0, Q_t) \tag{2.2}$$

for $t = 1, ..., n$ where $\epsilon_t$ and $\eta_t$ are independent error series and $Z_t.H_t, T_t.R_t$ and $Q_t$ are known matrices. The remaining series $\alpha_t$ is an unobserved series of state vectors which represent the development over time of the underlying system. This is a very general model which includes as special cases many specific models used in time series analysis, such as ARIMA models. For the purpose of this paper we assume that the object of the analysis is to estimate $\alpha_1, ..., \alpha_n$ for this and other models discussed later in this section.

Denote the stacked vectors $[y_1', ..., y_n']'$ and $[\alpha_1', \ldots, \alpha_n']'$ by $y$ and $\alpha$; also denote the joint density of $\alpha$ and $y$ by $p(\alpha, y)$ and the conditional density of $\alpha$ given $y$ by $p(\alpha|y)$. For model (2.1) and (2.2) it is standard to estimate $\alpha$ by $E(\alpha|y)$, which we denote by $\overline{\alpha}$. This estimation is carried out by the well known Kalman filter and smoother (KFS).

We shall consider only departures from normality in the observational part (1) of the model, while retaining the linear Gaussian form (2) for the development of $\alpha_t$. The first class of non-Gaussian observations we shall consider is the exponential family, for example Poisson or binomial observations, for which the density has the general form

$$p(y_t|\alpha_t) = \exp[\theta_t' y_t - b_t(\theta_t) + c_t(y_t)] \tag{2.3}$$

where $\theta_t = Z_t\alpha_t$, $t = 1, ..., n$ and where $b_t$ and $c_t$ are known functions. It turns out that the task of calculating $\overline{\alpha}$ for model 2.3 by analytical techniques is intractable. Fahrmeir (1992) therefore suggested estimating $\alpha$ by the mode $\hat{\alpha}$ of $p(\alpha|y)$ and he gave an approximation to $\hat{\alpha}$ based on the extended Kalman filter. He called $\hat{\alpha}$ the posterior mode estimate (PME). Durbin and Koopman (1993) showed how to compute $\hat{\alpha}$ accurately in a few iterations by applying the KFS to a linearised form of the estimating equation for $\alpha$.

A second important class of non-Gaussian models retains the same form as equation (2.1) but requires $\epsilon_t$ to have a non-Gaussian distribution, for example a t-distribution or a mixture of normals. Such distributions allow heavy-tailed observational densities to be handled. Again, $\hat{\alpha}$ is easily calculated by a few iterations of the KFS as shown by Durbin and Koopman (1993).

Finally, we consider nonlinear models where (2.1) is replaced by the equation $y_t = Z_t(\alpha_t) + \epsilon_t$ where $Z_t$ is a nonlinear function of $\alpha_t$ and $\epsilon_t$ may be Gaussian or non-Gaussian, for example a time series made up of the product of trend and seasonal plus random error.

For all these models we shall assume that $\hat{\alpha}$ is the unique solution to the equation $\partial \log p(\alpha|y)/\partial\alpha = 0$. But $\log p(\alpha|y) = \log p(\alpha, y) - \log p(y)$ where $p(y)$ is the marginal density of $y$. It follows that $\hat{\alpha}$ is the solution of the equation $\partial \log p(\alpha, y)/\partial\alpha = 0$. This is the form of the estimating equation for $\hat{\alpha}$ that we shall use in this paper.

The estimate $\hat{\alpha}$ has the attractive intuitive property that it is the most probable value of $\alpha$ given the data. This might be sufficient grounds for using it for some workers. However, $\overline{\alpha}$ has the objective optimality property that $E(\overline{\alpha} - \alpha) = 0$ and if $\overline{\alpha}^*$ is any other estimate of $\alpha$ such that $E(\overline{\alpha}^* - \alpha) = 0$, with $MSE(\overline{\alpha}) = V$, $MSE(\overline{\alpha}^*) = V^*$, then $V^* - V$ is non-negative definite. In the next section we shall seek an analogous optimality property for $\hat{\alpha}$ based on the theory of optimal unbiased estimating equations.

# 3 An Optimality Property of the Posterior Mode Estimate

We begin with some preliminaries. If $\hat{\alpha}$ is the unique solution for $\alpha$ of the $(m \times 1)$ vector equation $H(\alpha, y) = 0$ and if $E[H(\alpha, y)] = 0$, where expectation is taken with respect to the joint density of $\alpha$ and $y$, we say that $H(\alpha, y) = 0$ is an unbiased estimating equation. We want to establish a minimum variance property for such functions $H(\alpha, y)$ but obviously the equation can be multiplied through by an arbitrary nonsingular matrix and still give the same value $\hat{\alpha}$ as its solution. We therefore standardise $H(\alpha, y)$ in the usual way in estimating equation theory by multiplying it by $[E\{\dot{H}(\alpha, y)\}]^{-1}$ where $\dot{H}(\alpha, y) = \partial H(\alpha, y)/\partial \alpha'$ and we then seek a minimum variance property for the function $h(\alpha, y) = [E\{\dot{H}(\alpha, y)\}]^{-1} H(\alpha, y)$.

Let

$$\int H(\alpha, y) p(\alpha, y) dy = k(\alpha) \tag{3.1}$$

where $\int$ indicates integration over the domain of $y$ and where $dy = \prod_{t=1}^{n} \prod_{i=1}^{p} dy_{ti}$, $p$ being the dimensionality of $y_t$. Denote the $i$th element of $\alpha$ by $\alpha_{(i)}$, $i = 1, ..., mn$ where $m$ is the dimensionality of $\alpha_t$, and note that $k(\alpha)$ is an $(m \times 1)$ vector. We make the following assumption.

**Assumption A**

For each $i$, and for all $\alpha_{(j)}$ fixed, $j \neq i$, $\lim_{\alpha_{(i)} \to \pm\infty} k(\alpha) = 0$.

In view of the normality of the marginal distribution of $\alpha$, this requirement would appear selfevidently satisfied for reasonable functions $H(\alpha, y)$; however it seems sensible to make the requirement explicit. It is an appropriate reformulation for the present problem of Ferreira's (1960) condition $B(g) = 0$. Differentiating (3.1) under the integral sign, and assuming that this operation is valid, we obtain

$$\int \dot{H}(\alpha, y) p(\alpha, y) dy + \int H(\alpha.y) \frac{\partial \log p(\alpha, y)}{\partial \alpha'} p(\alpha, y) dy = \frac{\partial k(\alpha)}{\partial \alpha'}. \tag{3.2}$$

We now wish to integrate this with respect to $\alpha_{(1)}, ..., \alpha_{(mn)}$.

Writing

$$\frac{dk(\alpha)}{d\alpha'} = \left[ \frac{\partial k(\alpha)}{\partial \alpha_{(1)}}, ..., \frac{\partial k(\alpha)}{\partial \alpha_{(mn)}} \right],$$

it follows from Assumption A that

$$\int_{-\infty}^{\infty} \frac{\partial k(\alpha)}{\partial \alpha_{(i)}} d\alpha_{(i)} = [k(\alpha)]_{-\infty}^{\infty} = 0.$$

Thus

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial k(\alpha)}{\partial \alpha_{(i)}} ... d\alpha_{(mn)} = 0$$

and hence

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial k(\alpha)}{\partial \alpha'} d\alpha_{(1)} \dots d\alpha_{(mn)} = 0.$$

Integrating both sides of (3.2) with respect to $\alpha_{(1)}, ..., \alpha_{(mn)}$ we therefore have

$$E[\dot{H}(\alpha, y)] + E\left[H(\alpha, y)\frac{\partial \log p(\alpha, y)}{\partial \alpha'}\right] = 0.$$

Multiplying both sides by $[E\{\dot{H}(\alpha, y)\}]^{-1}$ gives

$$I + E\left[h(\alpha, y)\frac{\partial \log p(\alpha, y)}{\partial \alpha'}\right] = 0. \tag{3.3}$$

Let $\mathrm{Var}[h(x,y)] = E[h(\alpha,y)h(\alpha,y)']$ and let $\mathcal{T} = E[\frac{\partial \log(\alpha,y)}{\partial \alpha}\frac{\partial \log p(\alpha,y)'}{\partial \alpha}]$.

**Result 1**

If $E[H(\alpha, y)] = 0$, $H$ is differentiable with respect to $\alpha$ and Assumption A holds, $\mathrm{Var}[h(\alpha, y)] - \mathcal{T}^{-1}$ is non-negative definite.

We need the following further assumption.

**Assumption B.**

If $k(\alpha) = \frac{\partial \log p(\alpha)}{\partial \alpha} p(\alpha)$ where $p(\alpha)$ is the marginal density of $\alpha$ then Assumption A is satisfied.

**Result 2**

If Assumption B holds, the minimum is attained when $H(\alpha, y) = \frac{\partial p(\alpha,y)}{\partial \alpha}$.

**Proof of Result 1**

This follows immediately from (3.3) by the Cauchy-Schwarz inequality.

**Proof of Result 2**

Let $p(y|\alpha)$ be the conditional density of $y$ given $\alpha$. Then

$$\frac{\partial \log p(\alpha, y)}{\partial \alpha} = \frac{\partial \log p(y|\alpha)}{\partial \alpha} + \frac{\partial \log p(\alpha)}{\partial \alpha}.$$

Substituting this for $H(\alpha, y)$ in (3.1) gives

$$p(\alpha) \int \frac{\partial \log p(y|\alpha)}{\partial \alpha} p(y|\alpha) dy + p(\alpha) \frac{\partial \log p(\alpha)}{\partial \alpha} = k(\alpha).$$

Since $\partial \log p(y|\alpha)/\partial\alpha$ is the score function when $\alpha$ is regarded as fixed, the first term is zero so $k(\alpha) = p(\alpha)\partial \log p(\alpha)/\partial\alpha$. It follows from Assumption B that Assumption A is satisfied. Also, differentiation of the identity $\int p(\alpha, y) d\alpha dy = 1$ under the integral sign with respect to $\alpha$ shows that $E[\partial \log p(\alpha, y)/\partial\alpha] = 0$. Now if $H(\alpha, y) = \partial \log p(\alpha, y)/\partial\alpha$ then $E[\dot{H}(\alpha, y)] = E[\partial^2 \log p(\alpha, y)/\partial\alpha\partial\alpha'] = -\mathcal{T}$ as is shown by differentiating the identity $\int[\partial \log p(\alpha, y)/\partial\alpha] d\alpha dy = 0$ under the integral sign with respect to $\alpha'$.

Thus

$$h(\alpha, y) = -\mathcal{T}^{-1}\partial \log(\alpha, y)/\partial\alpha$$

so

$$Var[h(\alpha, y)] = \mathcal{T}^{-1} Var[\partial \log p(\alpha, y)/\partial\alpha]\mathcal{T}^{-1} = \mathcal{T}^{-1}.$$

This proves Result 2.

When Result 2 holds we say that the estimating equation is an optimal estimating equation.

These results can be regarded as an extension of the following result due independently to Godambe (1960) and G. A. Barnard, to whom it is attributed on p. 145 of Durbin (1960). If under suitable regularity conditions $x$ is an observational vector with density $f(x, \theta)$ where $\theta$ is a fixed scalar parameter, and if $G(x, \theta) = 0$ is an estimating equation for $\theta$ satisfying $E[G(x, \theta)] = 0$, then defining $g(x, \theta) = [E\{\frac{\partial G(x,\theta)}{\partial\theta}\}]^{-1} G(x, \theta)$ we have that

$$E[g(x, \theta)g(x, \theta)'] - \mathcal{T}_\theta^{-1}$$

is non-negative definite, where $\mathcal{T}_\theta = E[\frac{\partial \log f(x,\theta)}{\partial\theta}]^2$ and the minimum is attained when the equation $G(x, \theta) = 0$ is the maximum likelihood equation $\partial \log f(x, \theta)/\partial\theta = 0$. The straightforward extension to the case where $\theta$ is a vector was indicated on p.145 of Durbin (1960).

The Godambe-Barnard result was extended to the Baysian context in which $\theta$ is a random scalar parameter by Ferreira (1982). Thus the present findings could be interpreted as an extension of Ferreira's results, although we have not regarded $\alpha$ as a parameter vector in this paper.

Returning to the PME $\hat{\alpha}$ for nonlinear or non-Gaussian state space models, this is the solution of the estimating equation $\partial \log p(\alpha, y)/\partial\alpha = 0$. Since $k(\alpha) = \partial \log p(\alpha)/\partial\alpha$ and $p(\alpha)$ is a Gaussian density. Assumption A is satisfied. Thus Result 2 holds and $\hat{\alpha}$ is the solution of an optimal estimating equation. When the state space model has the linear Gaussian form (2.1) and (2.2) the estimating equation is linear in $\alpha$ and we have $h(\alpha, y) = \alpha - \overline{\alpha}$ so the mode $\hat{\alpha}$ is equal to $\overline{\alpha}$. This result is in fact obvious since in this case $p(\alpha|y)$ is normal and for a normal distribution the mode is equal to the mean.

## References

Durbin, J. (1960). Estimation of parameters in time series regression models. *J. R. Statist. Soc. B*, 22, 139-153.

Durbin, J. and Koopman, S. J. (1993). Filtering, smoothing and estimation for time series models when the observations come from exponential family distributions. LSE discussion paper.

Fahrmeir, L. (1992). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalised linear models. *J. Amer. Statist. Ass.,* 87, 501-509.

Ferreira, P. E. (1982). Estimating equations in the presence of prior knowledge. *Biometrika*, 69, 667-669.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, 31, 1208-1211.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press.