TI 2008-040/4
Tinbergen Institute Discussion Paper

Likelihood Functions for State Space Models with Diffuse Initial Conditions

Marc K. Francke
*Siem Jan Koopman**
Aart F. de Vos

Department of Econometrics, VU University Amsterdam.

* *Tinbergen Institute.*

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Likelihood functions for state space models with diffuse initial conditions

By Marc K. Francke, Siem Jan Koopman and Aart F. de Vos

Department of Econometrics, VU University Amsterdam,

De Boelelaan 1105, NL-1081 HV Amsterdam,

The Netherlands

mfrancke@feweb.vu.nl s.j.koopman@feweb.vu.nl avos@feweb.vu.nl

SUMMARY

State space models with nonstationary processes and fixed regression effects require a state vector with diffuse initial conditions. Different likelihood functions can be adopted for the estimation of parameters in time series models with diffuse initial conditions. In this paper we consider profile, diffuse and marginal likelihood functions. The marginal likelihood is defined as the likelihood function of a transformation of the data vector. The transformation is not unique. The diffuse likelihood is a marginal likelihood for a specific data transformation that may depend on parameters. Therefore, the diffuse likelihood can not be used generally for parameter estimation. Our newly proposed marginal likelihood function is based on an orthonormal transformation that does not depend on parameters. Likelihood functions for state space models are evaluated using the Kalman filter. The diffuse Kalman filter is specifically designed for computing the diffuse likelihood function. We show that a modification of the diffuse Kalman filter is needed for the evaluation of our proposed marginal likelihood function. Diffuse and marginal likelihood functions have better small sample properties compared to the profile likelihood function for the estimation of parameters in linear time series models. The results in our paper confirm the earlier findings and show that the diffuse likelihood function is not appropriate for a range of state space model specifications.

Some key words: Diffuse likelihood; Kalman filter; Marginal likelihood; Multivariate time series models; Profile likelihood.

1 Introduction

Consider the linear regression model $y = X\beta + u$ with observation vector y, covariate matrix X, regression coefficient vector β and disturbance vector $u \sim N(0, \sigma^2\Omega)$ where σ is the scaling factor and Ω is a variance matrix depending on the vector of nuisance parameters θ. We therefore may write $\Omega = \Omega(\theta)$ and possibly $X = X(\theta)$. The marginal likelihood function is defined as the likelihood function of a transformation of the observations in y such that the transformed data is orthogonal in X and therefore independent of β. The profile likelihood function for the linear regression model is the likelihood function evaluated at the maximum likelihood estimate of β. In econometrics, the profile likelihood function is also known as the concentrated likelihood function. Among others, Cooper and Thompson (1977) and Tunnicliffe-Wilson (1989) argue that the marginal likelihood is superior to the profile likelihood for the inference of nuisance parameters collected in vector θ. The marginal likelihood is for a (transformed) random variable and therefore its score vector has expectation zero, see, for example, Shephard (1993), Rahman and King (1997) and Francke and de Vos (2007).

The state space form for linear Gaussian time series models is convenient for likelihood-based estimation, signal extraction and forecasting. State space models can be represented as linear regression models with specifically designed matrices X and Ω, see Durbin and Koopman (2001, section 4.11). The likelihood function for stationary time series models can be evaluated by the Kalman filter as it effectively carries out the prediction error decomposition, see Schweppe (1965) and Harvey (1989). Nuisance parameter vector θ can be estimated by directly maximising the likelihood function. Time series models with (time-varying) regression parameters and nonstationary latent factors require state space formulations with unknown initial

conditions. In cases where the initial conditions are treated as fixed regression coefficients, the profile likelihood function can be computed as in Rosenberg (1973). When they are treated as random variables with large variances converging to infinity, a so-called diffuse likelihood function can be defined and be computed as described in, among others, Harvey (1989, section 3.4.3), Ansley and Kohn (1985, 1990), De Jong (1988, 1991) and Koopman (1997). The diffuse likelihood function is a marginal likelihood function based on a transformation that is not necessarily invariant to the parameter vector θ. In this paper we develop a marginal likelihood function for the state space model that is always invariant to θ in linear models. The evaluation of the marginal likelihood requires a modification of the diffuse Kalman filter. We further discuss its relation with profile and diffuse likelihood functions.

In section 2 we develop general expressions for the profile, diffuse and marginal likelihood functions and we discuss their merits. Section 3 shows how the Kalman filter needs to be modified for the computation of the marginal likelihood function. Illustrations are given in Section 4. It is shown that different specifications of the same model lead to different diffuse likelihood functions while the marginal likelihood functions remain equal. Section 5 concludes.

2 Likelihood functions for state space models

For the $N_t \times 1$ vector of time series y_t, with $t = 1, \ldots, T$, the state space model is given by

$$y_t = Z_t \alpha_t + \varepsilon_t, \qquad \alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \tag{1}$$

with $p \times 1$ state vector α_t and where the system matrices Z_t, T_t and R_t are fixed but may depend on known functions of parameter vector θ. The disturbance vectors ε_t and η_t are mutually and

serially independent and distributed by

$$\varepsilon_t \sim NID(0, \sigma^2 H_t), \qquad \eta_t \sim NID(0, \sigma^2 Q_t), \tag{2}$$

where σ^2 is a scaling factor and variance matrices H_t and Q_t are fixed but may depend on θ as well. The state space model specification is completed with the initial state vector modelled by

$$\alpha_1 = a + A\beta + C\xi, \qquad \xi \sim N(0, \sigma^2 Q_0), \tag{3}$$

where vector a and matrices A, C and Q_0 are fixed system variables of appropriate dimensions. The random vector ξ is independent of the other disturbances. The $k \times 1$ vector of coefficients β can be treated in two ways: (i) as a fixed and unknown vector; (ii) as a diffuse random vector, distributed by $\beta \sim N(0, \sigma^2 \Sigma)$ where $\Sigma^{-1} \to 0$. The initial state constant a is for known effects, the coefficient vector β is for unknown regression effects and for initial effects in nonstationary processes while the random vector ξ is for the exact initialisation of stationary processes. Since ξ is a random vector with a properly defined variance matrix, we are not interested in case (ii) with Σ as a regular variance matrix and therefore we assume always that $\Sigma^{-1} \to 0$ and $\mathrm{E}(\beta) = 0$ without loss of generality. Finally, the (possibly time-varying) system matrices are fixed and known functions of the vector of nuisance parameters θ. Textbook treatments of state space time series models are, amongst others, given by Anderson and Moore (1979), Harvey (1989) and Durbin and Koopman (2001).

The state space model (1) can be represented as a linear regression model. In particular, we can consider the formulation

$$y = c + X\beta + u, \qquad u \sim N(0, \sigma^2 \Omega). \tag{4}$$

The equivalence of (4) with the state space model is obtained by defining

$$y = (y_1', \ldots, y_T')', \qquad (c, X) = Z \left[I \ , \ T_1 \ , \ldots, \ \prod_{t=T-1}^{1} T_t \right]' (a, A), \tag{5}$$

where $Z = \text{diag}(Z_1, \ldots, Z_T)$ and with Ω representing the covariance structure implied by the state space model and depending on all system matrices. The dimension of y is $n \times 1$ with $n = \sum_{t=1}^{T} N_t$ and the dimension of X is $n \times k$. As system matrices may depend on θ, the explanatory variable matrix $X = X(\theta)$ and covariance matrix $\Omega = \Omega(\theta)$ may also depend on θ.

2.1 Profile likelihood function

In terms of the linear regression model (4) with a fixed and unknown β, the likelihood function is denoted by $L = \exp\{\ell(y; \beta, \sigma, \theta)\}$ and the scaled loglikelihood function is given by

$$\begin{aligned} -2 \log L &= -2\ell(y; \beta, \sigma, \theta) \\ &= n \log 2\pi + n \log \sigma^2 + \log |\Omega| + \sigma^{-2} (y - c - X\beta)' \Omega^{-1} (y - c - X\beta). \end{aligned} \tag{6}$$

Analytical expressions for the maximum likelihood estimators for β and σ can be obtained and are given by the generalized least squares expressions

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1}(y - c), \qquad \hat{\sigma}^2 = n^{-1}\text{RSS}, \qquad \text{RSS} = (y - c)'\Omega^{-1} M_\Omega (y - c), \tag{7}$$

where $M_\Omega = I - X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}$. The loglikelihood function (6) at the maximized location of $\beta = \hat{\beta}$ is given by

$$-2 \log L^P = -2\ell(y; \hat{\beta}, \sigma, \theta) = n \log 2\pi + n \log \sigma^2 + \log |\Omega| + \sigma^{-2}\text{RSS}, \tag{8}$$

and is defined as the *profile* loglikelihood function. We obtain the concentrated profile log-likelihood function by replacing σ^2 by its maximum likelihood estimator $\hat{\sigma}^2 = \text{RSS} \ / \ n$, that is

$$-2 \log L_c^P = -2\ell(y; \hat{\beta}, \hat{\sigma}, \theta) = n \log 2\pi + n \log \text{RSS} - n \log n + \log |\Omega| + n. \tag{9}$$

2.2 Diffuse likelihood function

In terms of the linear regression model with a random vector $\beta \sim N(0, \sigma^2\Sigma)$, the loglikelihood function is given by

$$\ell(y;\sigma,\theta) = \ell(y|\beta;\sigma,\theta) + \ell(\beta;\sigma,\theta) - \ell(\beta|y;\sigma,\theta), \tag{10}$$

where $\ell(y|\beta;\sigma,\theta) = \ell(y;\beta,\sigma,\theta)$ is given in (6) while $\ell(\beta;\sigma,\theta) = \ell(\beta;\sigma)$ with

$$-2\ell(\beta;\sigma) = k\log 2\pi + k\log\sigma^2 + \log|\Sigma| + \sigma^{-2}\beta'\Sigma^{-1}\beta.$$

The density implied by $\ell(\beta|y;\sigma,\theta)$ is obtained as follows. Since $\mathrm{E}(y) = c + X\mathrm{E}(\beta) = c$, $\mathrm{Var}(y) = \sigma^2(X\Sigma X' + \Omega)$, $\mathrm{E}(\beta) = 0$, $\mathrm{Var}(\beta) = \sigma^2\Sigma$ and $\mathrm{E}(\beta y') = \sigma^2\Sigma X'$, we obtain

$$\begin{aligned}
\mathrm{E}(\beta|y) &= \mathrm{E}(\beta y')\mathrm{Var}(y)^{-1}[y - \mathrm{E}(y)] \\
&= \Sigma X'(X\Sigma X' + \Omega)^{-1}(y - c) \\
&= (\Sigma^{-1} + X'\Omega^{-1}X)^{-1}X'\Omega^{-1}(y - c), \\
\mathrm{Var}(\beta|y) &= \mathrm{Var}(\beta) - \mathrm{E}(\beta y')\mathrm{Var}(y)^{-1}\mathrm{E}(y\beta') \\
&= \sigma^2\Sigma - \sigma^2\Sigma X'(X\Sigma X' + \Omega)^{-1}X\Sigma \\
&= \sigma^2(\Sigma^{-1} + X'\Omega^{-1}X)^{-1},
\end{aligned}$$

where we have suppressed the dependence on σ and θ. These results follow from a matrix inversion lemma and some minor manipulations. The first term in the right-hand side of (10) becomes

$$\begin{aligned}
-2\ell(\beta|y;\sigma,\theta) &= k\log 2\pi + k\log\sigma^2 - \log|\Sigma^{-1} + X'\Omega^{-1}X| + \sigma^{-2}\beta'(\Sigma^{-1} + X'\Omega^{-1}X)\beta \\
&+ \sigma^{-2}(y - c)'\Omega^{-1}X(\Sigma^{-1} + X'\Omega^{-1}X)^{-1}X'\Omega^{-1}(y - c) - 2\sigma^{-2}(y - c)'\Omega^{-1}X\beta.
\end{aligned}$$

By re-arranging the different terms of the loglikelihood function (10), we obtain

$$\begin{aligned}
-2\ell(y;\sigma,\theta) &= n\log 2\pi + n\log\sigma^2 + \log|\Omega| + \log|\Sigma| + \log|\Sigma^{-1} + X'\Omega^{-1}X| \\
&+ \sigma^{-2}(y - c)'[\Omega^{-1} - \Omega^{-1}X(\Sigma^{-1} + X'\Omega^{-1}X)^{-1}X'\Omega^{-1}](y - c).
\end{aligned}$$

The *diffuse* loglikelihood function $\log L^D$ is defined as

$$\ell_\infty(y;\sigma,\theta) = \lim_{\Sigma^{-1}\to 0} \ell(y;\sigma,\theta) + \frac{1}{2}\log|\Sigma|, \tag{11}$$

from which it follows that

$$-2\log L^D = -2\ell_\infty(y;\sigma,\theta) = n\log 2\pi + n\log\sigma^2 + \log|\Omega| + \log|X'\Omega^{-1}X| + \sigma^{-2}\text{RSS}, \tag{12}$$

which is equivalent to (8) apart from the term $\log|X'\Omega^{-1}X|$. This result is due to De Jong (1991). The loglikelihood function (12) at the maximized location of $\sigma = \hat{\sigma}$ is given by

$$-2\log L_c^D = -2\ell_\infty(y;\hat{\sigma},\theta) = n\log 2\pi + n\log\text{RSS} - n\log n + \log|\Omega| + \log|X'\Omega^{-1}X| + n. \tag{13}$$

which is equivalent to (9) apart from the term $\log|X'\Omega^{-1}X|$.

The definition of the diffuse loglikelihood function (11) may be regarded as somewhat *ad hoc*. For example, an alternative suggestion is to define the diffuse loglikelihood function as

$$\ell_\infty^*(y;\sigma,\theta) = \lim_{\Sigma^{-1}\to 0} \ell(y;\sigma,\theta) + \frac{1}{2}\log|2\pi\sigma^2\Sigma|, \tag{14}$$

see De Jong and Chu-Chun Lin (1994). In light of definition (14), the likelihood functions (12) and (13) remain the same but with n replaced by $m = n - k$. The alternative definition in (14) becomes relevant in the discussion of the marginal likelihood function in the next subsection.

2.3 Marginal likelihood function

The concept of marginal likelihood has been introduced by Kalbfleisch and Sprott (1970). The marginal likelihood function for model (4) is defined as the likelihood function that is invariant to the regression coefficient vector β. Many contributions in the statistics literature have developed the concept of marginal likelihoods further and have investigated this approach

in more detail, for example, see Patterson and Thompson (1971), Harville (1974), King (1980), Smyth and Verbyla (1996), and Rahman and King (1997). In particular, McCullagh and Nelder (1989) consider the marginal likelihood function for the generalized linear model. The marginal likelihood function has also been adopted for the inference of nuisance parameters in time series models, for example, see Levenbach (1972), Cooper and Thompson (1977) and Tunnicliffe-Wilson (1989). In the linear model $y = c + X\beta + u$ where $u \sim N(0, \Omega)$ with $X = X(\theta)$ and $\Omega = \Omega(\theta)$, the marginal likelihood function is for a transformed data vector $y^* = A'y$ that does not depend on β. The transformation matrix A has dimension $n \times m$ with $m = n - k$, is of full column rank and is subject to $A'X = 0$. Apart from these conditions, the choice of matrix A is irrelevant. In our context of likelihood-based inference for θ, it is important to assume that matrix A does not depend on θ.

The scaled log-density function of y^* is given by

$$-2\ell(y^*; \sigma, \theta) = m \log 2\pi + m \log \sigma^2 + \log |A'\Omega A| + \sigma^{-2}(y - c)'A(A'\Omega A)^{-1}A'(y - c), \qquad (15)$$

since $A'X = 0$. The equalities

$$(\Omega A, X)'A(A'\Omega A)^{-1}A' = (A, 0)', \quad \Leftrightarrow \quad (\Omega A, X)'\Omega^{-1}M_\Omega = (A, 0)',$$

imply that $A(A'\Omega A)^{-1}A' = \Omega^{-1}M_\Omega$. Furthermore, since

$$\begin{aligned} |\Omega| \cdot |A'A| \cdot |X'X| &= |(A, X)'\Omega(A, X)| \\ &= |A'\Omega A| \cdot |X'\Omega X - X'\Omega A(A'\Omega A)^{-1}A'\Omega X| \\ &= |A'\Omega A| \cdot |X'\Omega X - X'M_\Omega \Omega X| \\ &= |A'\Omega A| \cdot |X'X|^2 \cdot |X'\Omega^{-1}X|^{-1}, \end{aligned}$$

the determinental term in the density is $|A'\Omega A| = |\Omega| \cdot |A'A| \cdot |X'X|^{-1}|X'\Omega^{-1}X|$. Following Harville (1974) we normalize matrix A such that $A'A = I_m$ and $|A'A| = 1$. The marginal

likelihood function with respect to β is based on the density of $y^* = A'y$. The scaled marginal loglikelihood function is then given by

$$\begin{aligned} -2\log L^M &= -2\ell(y^*;\sigma,\theta) \qquad (16)\\ &= m\log 2\pi + m\log\sigma^2 + \log|\Omega| + \log|X'\Omega^{-1}X| - \log|X'X| + \sigma^{-2}\text{RSS}. \end{aligned}$$

The marginal likelihood (16) is equivalent to (12) apart from the term $\log|X'X|$ and n replaced by m. When the diffuse likelihood function is defined as in (14), the marginal likelihood only differs by the term $\log|X'X|$.

The variance scalar σ^2 can also be concentrated out from the marginal likelihood function. The marginal likelihood evaluated at the maximized value of σ is given by

$$\begin{aligned} -2\log L_c^M &= -2\ell(y^*;\hat\sigma,\theta) \qquad (17)\\ &= m\log 2\pi + m\log\text{RSS} - m\log m + \log|\Omega| + \log|X'\Omega^{-1}X| - \log|X'X| + m, \end{aligned}$$

and is equivalent to (13) apart from the term $\log|X'\Omega^{-1}X|$ and n replaced by m. Expressions (16) and (17) are new and convenient for our purposes below.

2.4 Discussion of likelihood functions

The close resemblance of the diffuse and marginal likelihoods has been discussed by Shephard (1993) and Kuo (1999). Their marginal likelihood function does not have the term $\log|X'X|$ in (16) and the marginal and diffuse likelihood functions are proportional. They also argue that the marginal likelihood function is based on the density of a random variable and therefore the score function has zero expectation. Given that the difference between the profile and marginal likelihoods is the term $\log|X'\Omega^{-1}X| - \log|X'X|$ where $\Omega = \Omega(\theta)$ and $X = X(\theta)$, it is obvious

that the score of the profile likelihood function is non-zero and the profile likelihood is subject to a bias term. As a result, the use of the profile likelihood function introduces bias in the estimation of θ.

In cases where X does *not* depend on θ, the marginal and diffuse likelihoods are indeed proportional to each other and the choice between the two likelihoods is irrelevant for the inference of θ. This fact is recognised by Ansley and Kohn (1985) in their treatment of the diffuse likelihood function and they explicitly assume that θ does not influence the transformation matrix. However, in the next section we consider cases where matrix X *does* depend on θ, that is $X = X(\theta)$. Then, the data transformation implied by the diffuse likelihood function of Shephard (1993) and Kuo (1999) is based on some matrix A^* for which we can assume that $|A^{*\prime}A^*| \propto |X'X|$ without loss of generality. In case $X = X(\theta)$, the diffuse likelihood function is not appropriate for a likelihood-based analysis with respect to θ. The marginal likelihood function defined by (16) is based on the transformation matrix A with $A'A = I$ as shown in the previous subsection. The orthonormal transformation does not depend on θ in linear models and therefore can be used for the inference of θ. In other words, the term $\log|X'X|$ in (16) and (17) cannot be ignored.

In case the regression model (4) implies a time series model in the state space form (1), matrix X and its dependence on θ should be considered carefully. In case of stationary time series models without regression effects, this issue does not arise as β is not present. In case regression effects are present and in case the model includes nonstationary processes, coefficient vector β is present and the dependence of θ on covariate matrix X must be taken into account. The use of the marginal likelihood function is recommended for this class of linear time series models.

3 Evaluation of likelihood functions

The Kalman filter effectively carries out the prediction error decomposition for time series models in the state space representation (1), see Schweppe (1965) and Harvey (1989). The prediction error decomposition is based on

$$\ell(y) = \ell(y_1, \ldots, y_T) = \ell(y_1) \prod_{t=2}^{T} \ell(y_t|Y_{t-1}),$$

where $Y_t = \{y_1, \ldots, y_t\}$. The prediction error $v_t = y_t - \mathrm{E}(y_t|Y_{t-1})$, with its variance matrix $F_t = \mathrm{Var}(y_t|Y_{t-1}) = \mathrm{Var}(v_t)$, is serially uncorrelated when the model is correctly specified. This implies that $\mathrm{Var}(v) = F$ is block-diagonal with prediction error vector $v = (v_1, \ldots, v_T)'$ and associated variance matrix $F = \mathrm{diag}(F_1, \ldots, F_T)$. The Kalman filter therefore carries out the Cholesky decomposition $\Omega = L^{-1}FL'^{-1}$, or $F = L\Omega L'$, where $\Omega = \Omega(\theta)$ is implied by state space model (1) and $n \times n$ matrix L is a lower block unity triangular matrix with $|L| = 1$. It also implicitly follows that $v = L(y - c)$.

The Kalman filter for the state space model (1) with $\beta = 0$ in the initial state specification (3) is given by

$$\begin{aligned} v_t &= y_t - Z_t a_t, & F_t &= Z_t P_t Z_t' + H_t, \\ & & K_t &= T_t P_t Z_t' F_t^{-1}, \\ a_{t+1} &= T_t a_t + K_t v_t, & P_{t+1} &= T_t P_t T_t' - K_t F_t K_t' + R_t Q_t R_t', \end{aligned} \tag{18}$$

for $t = 1, \ldots, T$ and with $a_1 = a$ and $P_1 = CQ_0C'$. The likelihood function (6) with $\beta = 0$ can be written as

$$\begin{aligned} -2\log L &= n\log 2\pi + n\log\sigma^2 + \log(|L||\Omega||L'|) + \sigma^{-2}(y-c)'L'L'^{-1}\Omega^{-1}L^{-1}L(y-c) \\ &= n\log 2\pi + n\log\sigma^2 + \log|F| + \sigma^{-2}v'F^{-1}v \\ &= n\log 2\pi + n\log\sigma^2 + \textstyle\sum_{t=1}^{T}\log|F_t| + \sigma^{-2}\sum_{t=1}^{T} v_t'F_t^{-1}v_t. \end{aligned}$$

It follows that the Kalman filter can evaluate the likelihood function (6) with $\beta = 0$ in a computationally efficient way.

3.1 Evaluation of profile likelihood

The evaluation of the profile likelihood functions (8) and (9) focuses on

$$\log|\Omega| = \log|L\Omega L'| = \log|F|, \qquad \text{RSS} = (y-c)'L'L'^{-1}\Omega^{-1}L^{-1}LM_\Omega L^{-1}L(y-c) = v'F^{-1}M^*v,$$

where

$$M^* = LM_\Omega L^{-1} = I - LX(X'L'L'^{-1}\Omega^{-1}L^{-1}LX)^{-1}X'L'L'^{-1}\Omega L^{-1} = I - V(V'F^{-1}V)^{-1}V'F^{-1},$$

with $V = LX$. It follows that

$$\text{RSS} = q - s'S^{-1}s, \quad \text{where} \quad q = v'F^{-1}v, \qquad s = V'F^{-1}v, \qquad S = V'F^{-1}V. \tag{19}$$

We note that $q \equiv (y-c)'\Omega^{-1}(y-c)$, $s \equiv X'\Omega^{-1}(y-c)$ and $S \equiv X'\Omega^{-1}X$. Given that the Kalman filter evaluates the block elements of $v = L(y-c)$ recursively, the columns of matrix $V = LX = L(X^1, \ldots, X^k)$, where X^i is the ith column of X for $i = 1, \ldots, k$, can be evaluated simultaneously and recursively in the following way

$$V_t = X_t - Z_t A_t, \qquad A_{t+1} = T_t A_t + K_t V_t, \tag{20}$$

with $A_1 = A$ and $V = (V_1', \ldots, V_T')'$. Further, we have

$$q = \sum_{t=1}^{T} v_t' F_t^{-1} v_t, \qquad s = \sum_{t=1}^{T} V_t' F_t^{-1} v_t, \qquad S = \sum_{t=1}^{T} V_t' F_t^{-1} V_t.$$

The Kalman filter with the additional recursion (20) is referred to as the diffuse Kalman filter and is developed by De Jong (1991).

The likelihood function (6), for any β, and the profile loglikelihood functions $\log L^P$ and $\log L_c^P$ can be expressed by

$$\begin{aligned}
-2\log L &= n\log 2\pi + n\log\sigma^2 + \log|F| + \sigma^{-2}(v - V\beta)'F^{-1}(v - V\beta),\\
-2\log L^P &= n\log 2\pi + n\log\sigma^2 + \log|F| + \sigma^{-2}(q - s'S^{-1}s),\\
-2\log L_c^P &= n\log 2\pi + n\log(q - s'S^{-1}s) - n\log n + \log|F| + n,
\end{aligned}$$

which can be evaluated by the diffuse Kalman filter in a computationally efficient way.

3.2 Evaluation of diffuse likelihood

The diffuse loglikelihood functions (12) and (13) are evaluated by

$$\begin{aligned}
-2\log L^D &= m\log 2\pi + m\log\sigma^2 + \log|F| + \log|S| + \sigma^{-2}(q - s'S^{-1}s),\\
-2\log L_c^D &= m\log 2\pi + m\log(q - s'S^{-1}s) - m\log m + \log|F| + \log|S| + m,
\end{aligned}$$

respectively. Here we have replaced n by m and in effect have adopted definition (14) for the diffuse likelihood function. All terms can be evaluated by the diffuse Kalman filter.

3.3 Evaluation of marginal likelihood

The marginal loglikelihood differs from the diffuse loglikelihood by the term $\frac{1}{2}\log|X'X|$. It follows from the design of X in (5), implied by the state space model (1), that the $k \times k$ matrix $S^* = X'X$ can be evaluated by the recursion

$$V_t^* = Z_t A_t^*, \qquad A_{t+1}^* = T_t A_t^*, \qquad t = 1, \ldots, T, \tag{21}$$

with $A_1^* = A^*$ and $S^* = \sum_{t=1}^n V_t^{*\prime} V_t^*$. The marginal loglikelihood functions are given by

$$\begin{aligned}
-2\log L^M &= m\log 2\pi + m\log\sigma^2 + \log|F| + \log|S| - \log|S^*| + \sigma^{-2}(q - s'S^{-1}s),\\
-2\log L_c^M &= m\log 2\pi + m\log(q - s'S^{-1}s) - m\log m + \log|F| + \log|S| - \log|S^*| + m,
\end{aligned}$$

and are evaluated by the diffuse Kalman filter together with the additional recursion (21).

4 Illustrations

In this section we explore the differences between estimation based on the profile, diffuse and marginal likelihood functions. The diffuse and marginal likelihood functions have score functions with zero expectations since they are based on a random variable (the transformed data vector). As a result, the profile likelihood function does not have this property. The non-zero expectation of the score for the profile likelihood leads to a bias in the estimation of θ. Shephard and Harvey (1990), Shephard (1993) and Kuo (1999) have investigated this in more detail in the context of estimating the signal-to-noise ratio of the stochastic trend model $y_t = \mu_t + \epsilon_t$ with trend μ_t as the random walk process $\mu_{t+1} = \mu_t + \eta_t$ and signal-to-noise ratio $q = \text{var}(\eta_t) \,/\, \text{var}(\epsilon_t)$. Based on a set of Monte Carlo studies, it is found that the estimation of the signal-to-noise ratio q based on the profile likelihood leads to many zero estimates while the underlying data generating process used a strictly positive q value. Estimation based on the diffuse/marginal likelihood function reduces this bias substantially. In this section we confirm these findings and review the consequences of considering stationary, nonstationary and multivariate time series models. Furthermore, we argue that in cases of interest the marginal likelihood function (16) should be used rather than profile or diffuse likelihood functions for parameter estimation. Since we focus on differences between likelihood functions, we present them explicitly in Table 1.

4.1 Stationary time series models

The state space form of a linear stationary time series model without regression effects has a state vector depending only on stationary processes and with initial condition (3) given by

Table 1: *Differences between the loglikelihood functions. The loglikelihood functions* $\log L^P$, $\log L^D$ *and* $\log L^M$ *refer to* (8), (12) *and* (16), *respectively, while* $\log L^{D*}$ *refers to the diffuse loglikelihood function as defined by* (14) *which is equal to* (12) *with* n *replaced by* m. *Matrices* S *and* S^* *are defined below* (20) *and* (21), *respectively. The lower triangular part of the table represents the differences of the loglikelihood functions. The upper triangular part reports the differences in the data vector dimensions.*

	$-2\log L^P$	$-2\log L^D$	$-2\log L^{D*}$	$-2\log L^M$
$-2\log L^P$	0	0	$n-m$	$n-m$
$-2\log L^D$	$\log\lvert S\rvert$	0	$n-m$	$n-m$
$-2\log L^{D*}$	$\log\lvert S\rvert$	0	0	0
$-2\log L^M$	$\log\lvert S\rvert - \log\lvert S^*\rvert$	$-\log\lvert S^*\rvert$	$-\log\lvert S^*\rvert$	0

$\alpha_1 = a + C\xi$, that is $\beta = 0$. As a result, the matrix X is non-existent and the profile, marginal and diffuse likelihood functions are equivalent. In case the stationary time series model contains linear regression effects, the vector $\beta \neq 0$ in (3) represents the regression coefficients in the model. The resulting matrix X in (5) is exogenous and does not depend on θ. The profile likelihood does not have the term $\log|S| = \log|X'\Omega^{-1}X|$ while only the marginal likelihood functions has the term $\log|S^*| = \log|X'X|$. Since $|X'X|$ is fixed, the diffuse and marginal likelihood functions are proportional to each other and the estimation of θ is not affected by the choice between the two. The profile likelihood function will lead to a maximum likelihood estimator of θ that is different from the one based on the diffuse/marginal likelihood function.

4.2 Nonstationary time series models

The initial conditions of nonstationary components in a time series model must depend on the vector β in (3). In such cases, $\beta \neq 0$ and as long as vector θ does not enter X, the diffuse and marginal likelihoods are proportional and provide the same maximum likelihood estimates of θ. The profile likelihood function leads to a different estimate of θ. Shephard and Harvey (1990) and Shephard (1993) carry out Monte Carlo studies using the stochastic trend model with a strictly positive signal-to-noise ratio as the data generating process. They show that the number of zero estimates of the signal-to-noise ratio based on the profile likelihood is considerably higher than based on the marginal likelihood. They obtain similar results when regression effects are introduced in the model, requiring the extension of β with regression coefficients. We have been able to reproduce their findings.

Testing for unit roots in autoregressive models also provides an illustration of the difference between profile and marginal likelihood functions. For example, the first-order autoregressive model with a constant is given by

$$y_t = \mu + u_t, \qquad u_{t+1} = \rho u_t + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2), \tag{22}$$

for $t = 1, \ldots, T$, where

$$u_1 = \begin{cases} \xi & \text{for } \rho = 1, \\ N\{0, \sigma_\varepsilon^2 \ / \ (1-\rho^2)\} & \text{for } |\rho| < 1, \end{cases}$$

with ξ as an unknown scalar. The specification of the initial condition (22) is coherent as the variance of u_1 goes to infinity for $\rho \uparrow 1$. The core of this problem is that the profile likelihood degenerates in the unit root. The marginal likelihood is well-defined for $-1 < \rho \leq 1$ where the profile likelihood is zero when $\rho = 1$. Francke and de Vos (2007) show that unit root tests

based on the marginal likelihood ratio outperform other well-known tests in the literature. This result holds specifically for small samples.

4.3 Multivariate nonstationary time series models

The generality of the state space framework allows different state space representations of the same time series model. The likelihood value should not depend on the particular state space formulation that is used. However, we will show that this can be the case for the diffuse likelihood function while this is not the case for the profile and marginal likelihood functions. A convenient illustration is given in the context of multivariate time series models. Consider a model with random walk trends from which some trends are possibly common to all series. The $N \times 1$ vector of observations y_t is then modelled by

$$y_t = \gamma + \Lambda \mu_t + \varepsilon_t, \qquad \mu_{t+1} = \mu_t + \eta_t, \qquad \eta_t \sim N(0, I_r), \tag{23}$$

for $t = 1, \ldots, T$, where μ_t is an $r \times 1$ vector of independent random walks with $r < N$ and γ is an $N \times 1$ fixed unknown vector for which the first r elements are zero, $\gamma = (0, \ldots, 0, \gamma_{r+1}, \ldots, \gamma_N)'$. The $N \times r$ matrix of factor loadings Λ has unknown fixed elements which are collected in the parameter vector θ. The properties of disturbance vector ε_t are not relevant for this illustration but ε_t is assumed Gaussian and independent of η_s for $t, s = 1, \ldots, T$.

A valid state space formulation (1) of model (23) can be based on the $N \times 1$ state vector $\alpha_t = (\mu_t', \gamma_{r+1}, \ldots, \gamma_N)'$ and with system matrices

$$Z_t = \begin{bmatrix} \Lambda_1 & 0 \\ \Lambda_2 & I_{N-r} \end{bmatrix}, \qquad T_t = I_N, \qquad R_t = \begin{bmatrix} I_r \\ 0 \end{bmatrix}, \qquad Q_t = I_r, \tag{24}$$

where Λ_1 consists of the first r rows of Λ and Λ_2 are the remaining $N - r$ rows of Λ. Given

the nonstationary process for μ_t, all initial values in α_t at $t = 1$ are treated as unknown coefficients and collected in vector β of (3). The initial state condition for this time series model is therefore given by (3) with $a = 0$, $B = I_N$ and $C = 0$. As a result, we have matrix $X = (Z_1', \ldots, Z_T')'$ in (5) that depends on Λ and therefore $X = X(\theta)$. For this state space formulation, the marginal and diffuse likelihood functions are different. It is easily shown that $|S^*| = |X'X| = T|\Lambda_1'\Lambda_1| = T|\Lambda_1|^2$ where S^* is formally defined below (21).

Alternatively, a state space formulation (1) of model (23) can be based on the $N \times 1$ state vector $\alpha_t = \gamma + \Lambda\mu_t$ and with system matrices $Z_t = I_N$, $T_t = I_N$, $R_t = \Lambda$ and $Q_t = I_r$. The initial state conditions in (3) remain the same with $a = 0$, $B = I_N$ and $C = 0$. In this case, $n \times N$ matrix $X = (I_N, \ldots, I_N)'$ in (5), with $n = N \cdot T$, does not depend on θ and the marginal and diffuse likelihoods are proportional to each other. It can be shown that the marginal likelihood functions for both state space representations are proportional. The diffuse likelihood functions are different for the two alternative state space formulations. In the first case, we have, say, $S = S_1$ and in the second case, we have, say, $S = S_2$. It then follows that $S_2 = (I_N, \ldots, I_N)\Omega^{-1}(I_N, \ldots, I_N)'$ and $S_1 = Z'S_2Z$ where $Z = \text{diag}(Z_1, \ldots, Z_T)$ and with Z_t as defined in the first state space representation (24) for $t = 1, \ldots, T$. The determinental terms $|S_1|$ and $|S_2|$ therefore differ by the term $T|\Lambda_1|^2$. This term is equal to $|S^*|$ for the first state space representation. In other words, the marginal likelihood for the first state space representation is equivalent to the marginal likelihood and (upto proportionality) to the diffuse likelihood for the second state space representation. The diffuse likelihood for the first representation is different. Finally, the transformation matrix A, underlying the marginal likelihood function and subject to $A'X = 0$, does not depend on θ (as required) since X does not depend on θ. In cases that X depends on θ in a linear way, matrix A does still not depend on θ.

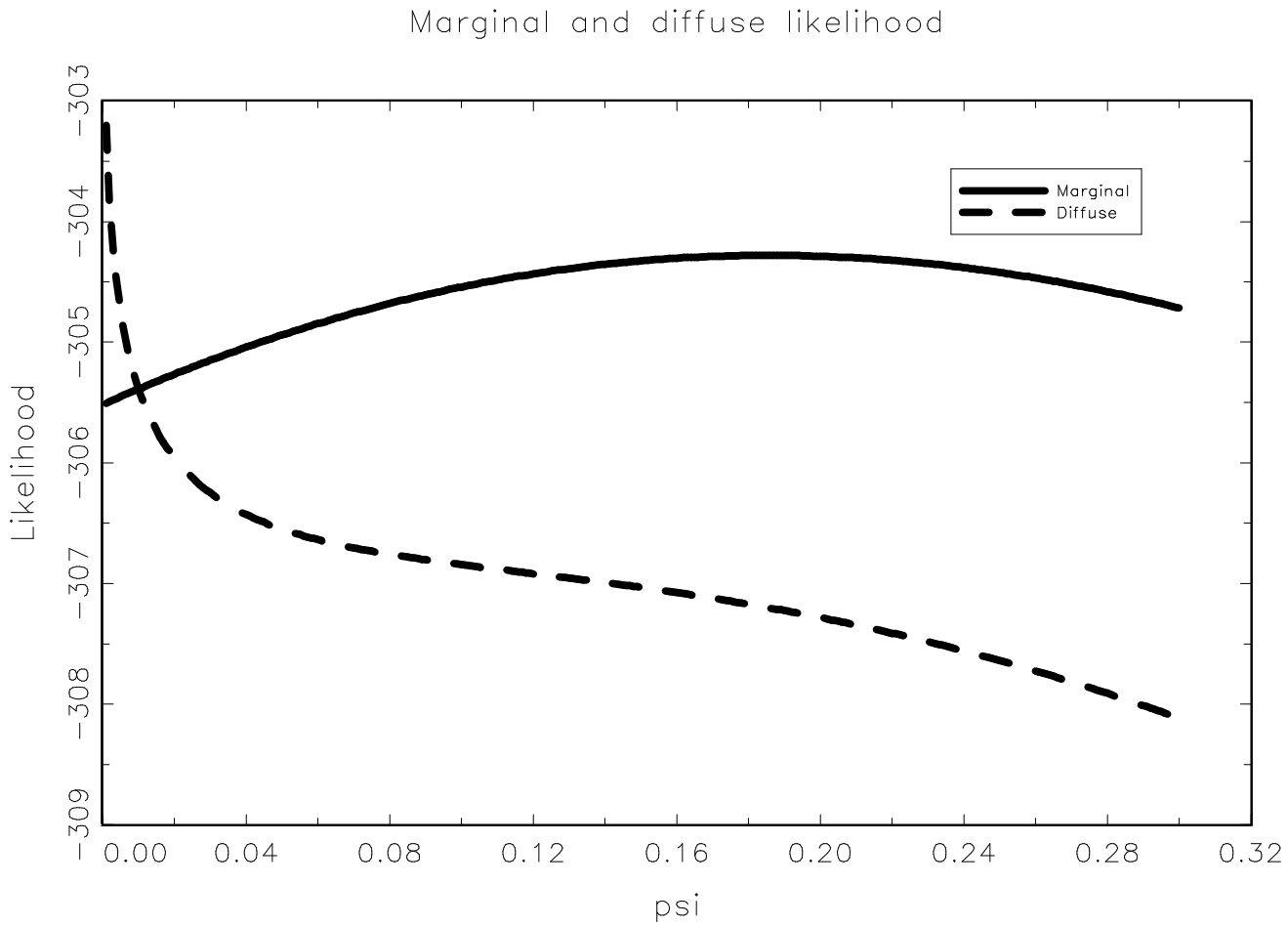

Figure 1: Marginal and diffuse loglikelihood functions for a bivariate version of the model (23), represented by the state space form (24), as functions of $\psi = \sqrt{\text{Var}(\eta_t)}$. The true value of ψ is 0.25.

To illustrate that the diffuse likelihood function may be inappropriate, we consider state space representation (24) for model (23) with $N = 2$ and $r = 1$. We simulate $T = 100$ observations from the bivariate common trend model (23) with $\gamma = (0, 1)'$, $\Lambda = (1, 0.1)'$, $\text{Var}(\epsilon_t) = I_2$ and $\text{Var}(\eta_t) = 0.25^2$. Figure 1 presents the marginal and diffuse loglikehoods as functions of $\psi = \sqrt{\text{Var}(\eta_t)}$. The diffuse likelihood is clearly not proportional to the marginal likelihood while the maximum of the latter is in the neighborhood of the true value $\psi = 0.25$. The diffuse and marginal loglikelihood functions for the second state space representation are proportional to the marginal loglikelihood as depicted in Figure 1.

5 Conclusion

In this paper we have argued for the preference of the marginal likelihood function over the profile and diffuse likelihood functions when we estimate parameters in time series models with nonstationary components and unknown regression effects. In many cases, the diffuse and marginal likelihood functions are proportional to each other. However, in cases where the implied data transformation for the diffuse likelihood function depends on parameters, estimation based on the diffuse likelihood function will lead to unreliable results. For these cases, the marginal likelihood as defined by Harville (1974) and adapted for state space models in this paper should be considered since the implied data transformation does not depend on parameters in linear models.

ACKNOWLEDGEMENTS

We would like to thank Jacques J.F. Commandeur and Borus Jungbacker for their comments on an earlier version of this paper. All errors are our own.

References

Anderson, B. D. O. and J. B. Moore (1979). *Optimal Filtering*. Englewood Cliffs: Prentice-Hall.

Ansley, C. F. and R. Kohn (1985). Estimation, filtering and smoothing in state space models with incompletely specified initial conditions. *Annals of Statistics 13*, 1286–1316.

Ansley, C. F. and R. Kohn (1990). Filtering and smoothing in state space models with partially diffuse initial conditions. *J. Time Series Analysis 11*, 275–93.

Cooper, D. M. and R. Thompson (1977). A note on the estimation of the parameters of the autoregressive-moving average process. *Biometrika 64*, 625–628.

De Jong, P. (1988). The likelihood for a state-space model. *Biometrika 75*, 165–169.

De Jong, P. (1991). The diffuse Kalman filter. *The Annals of Statistics 2*, 1073–1083.

De Jong, P. and S. Chu-Chun Lin (1994). Stationary and non-stationary state space models. *Journal of Time Series Analysis 15*, 151–166.

Durbin, J. and S. J. Koopman (2001). *Times Series Analysis by State Space Methods*. Oxford University Press, Oxford.

Francke, M. K. and A. F. de Vos (2007). Marginal likelihood and unit roots. *Journal of Econometrics 137*, 708–728.

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

Harville, D. A. (1974). Bayesian inference for variance components using only error contrast. *Biometrika 61*, 383–385.

Kalbfleisch, J. D. and D. A. Sprott (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society B 32*, 175–208.

King, M. L. (1980). Robust tests for spherical symmetry and their application to least squares regression. *The Annals of Statistics 8*, 1265–1271.

Koopman, S. J. (1997). Exact initial Kalman filtering and smoothing for nonstationary time series models. *Journal of the American Statistical Assocation 92*, 1630–1638.

Kuo, B. S. (1999). Asymptotics of ML estimator for regression models with a stochastic trend component. *Econometric Theory 15*, 24–29.

Levenbach, H. (1972). Estimation of autoregressive parameters from a marginal likelihood function. *Biometrika 59*, 61–71.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2 ed.). London: Chapman & Hall.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*, 545–554.

Rahman, S. and M. L. King (1997). Marginal-likelihood score-based tests of regression disturbances in the presence of nuisance parameters. *Journal of Econometrics 82*, 81–106.

Rosenberg, B. (1973). Random coefficients models: the analysis of a cross-section of time series by stochastically convergent parameter regression. *Annals of Economic and Social Measurement 2*, 399–428.

Schweppe, F. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Transac-*

tions on Information Theory 11, 61–70.

Shephard, N. (1993). Maximum likelihood estimation of regression models with stochastic trend components. *Journal of the American Statistical Association 88*, 590–595.

Shephard, N. and A. C. Harvey (1990). On the probability of estimating a deterministic component in the local level model. *Journal of Time Series Analysis 11*, 339–347.

Smyth, G. K. and A. P. Verbyla (1996). A conditional likelihood approach to REML in generalized linear models. *Journal of the Royal Statistical Society B 58*, 565–572.

Tunnicliffe-Wilson, G. (1989). On the use of marginal likelihood in time series model estimation. *Journal of the Royal Statistical Society B 51*, 15–27.