Bayesian Model Discrimination and Bayes Factors for Normal Linear State Space Models

Frühwirth-Schnatter, Sylvia

Published: 01/01/1993

Document Version
Publisher's PDF, also known as Version of record

Bayesian Model Discrimination and Bayes Factors for Normal Linear State Space Models

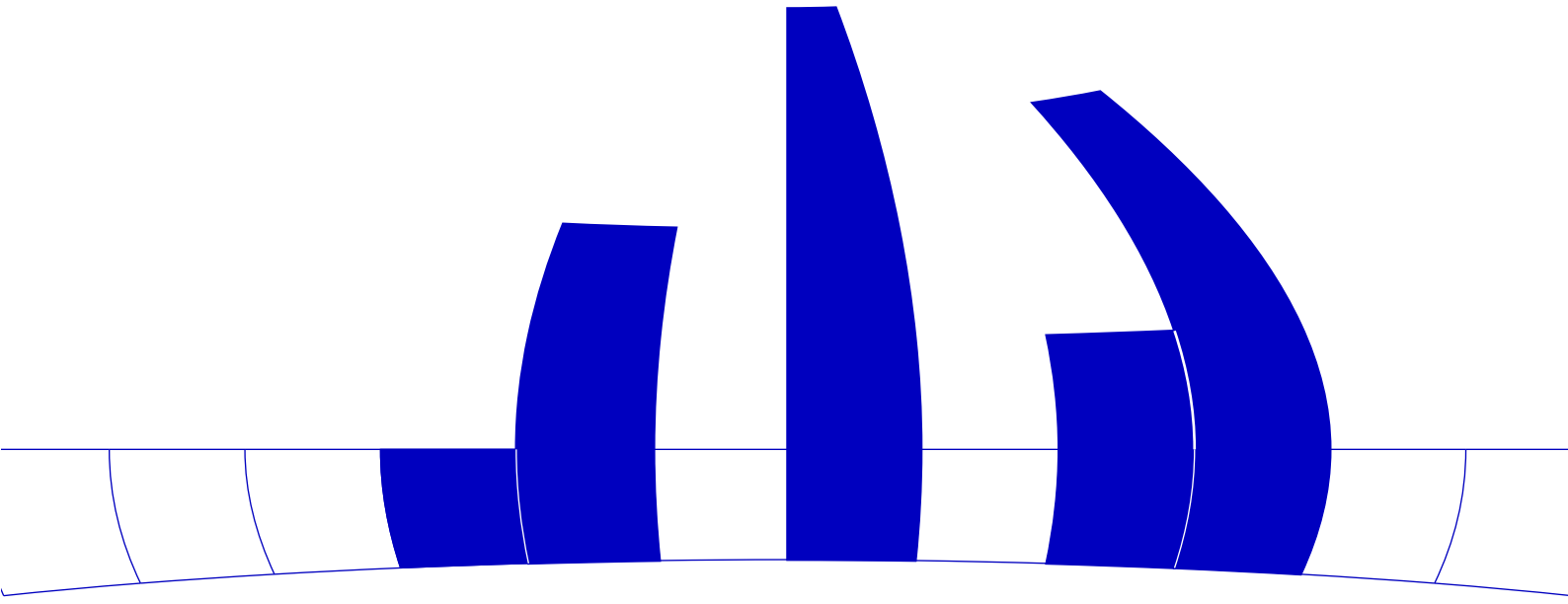
Sylvia Frühwirth-Schnatter

Institut für Statistik
Wirtschaftsuniversität Wien

Forschungsberichte

Bericht 33
May 1993

http://statmath.wu-wien.ac.at/



Abstract

It is suggested to discriminate between different state space models for a given time series by means of a Bayesian approach which chooses the model that minimizes the expected loss. Practical implementation of this procedure requires a fully Bayesian analysis for both the state vector and the unknown hyperparameters which is carried out by Markov chain Monte Carlo methods. Application to some non-standard situations such as testing hypotheses on the boundary of the parameter space, discriminating non-nested models and discrimination of more than two models is discussed in detail.

Keywords: Bayes factors, Markov chain Monte Carlo, model discrimination, model likelihood, state space models, training sample priors.

1 Introduction

Suppose that L different state space models $\mathcal{M}_1, \cdots \mathcal{M}_L$ are possible candidates for modelling a time series $y^N = \{y_1, \ldots, y_N\}$. We suggest to discriminate between these models by means of a Bayesian discrimination procedure (e.g. Geisser and Eddy, 1979). Assume that the following quantities are known: the prior probabilities $P(\mathcal{M}_l|y^0)$, $l = 1, 2, \ldots, L$ and the loss $r(\mathcal{M}_q, \mathcal{M}_l)$, $1 \leq q \leq L$, $1 \leq l \leq L$ caused by a decision for $\mathcal{M}_q$, if $\mathcal{M}_l$ holds. One then decides for that model $\mathcal{M}^\star$ which minimizes the expected loss given the time series $y^N = \{y_1, \ldots, y_N\}$:

$$\sum_{l=1}^{L} r(\mathcal{M}^\star, \mathcal{M}_l) P(\mathcal{M}_l|y^N) \leq \sum_{l=1}^{L} r(\mathcal{M}_q, \mathcal{M}_l) P(\mathcal{M}_l|y^N), \quad \forall q = 1, \ldots, L. \tag{1}$$

Discrimination rule (1) requires to compute the posterior probabilities $P(\mathcal{M}_l|y^N)$ for all models $\mathcal{M}_l$ given the time series y^N from Bayes' theorem:

$$P(\mathcal{M}_l|y^N) \propto p(y_1, \ldots, y_N|y^0, \mathcal{M}_l) P(\mathcal{M}_l|y^0). \tag{2}$$

The factor $L(y^N|y^0, \mathcal{M}_l) := p(y_1, \ldots, y_N|y^0, \mathcal{M}_l)$ is called model likelihood. Practical implementation of this approach requires algorithms to compute the model likelihood which is standard only for normal linear state space models ("dynamic linear models") with known hyperparameters. In Section 2 we will discuss Markov chain Monte Carlo approximation of the model likelihood for dynamic linear models with unknown hyperparameters. Section 3 deals with the issues of deriving training sample priors both for the state vector and for the unknown hyperparameters.

In Section 4 we apply the Bayesian discrimination rule (1) to some non-standard situations where likelihood based methods seem problematical, among them non-regular cases such as hypothesis testing on the boundary of the parameter space, discriminating non-nested models, and discrimination of more than two models.

2 Markov Chain Monte Carlo Approximation of Model Likelihoods

For practical implementation of the Bayesian discrimination rule (1) we need to compute the model likelihood $L(y^N|y^0,\mathcal{M})$ of a state space model $\mathcal{M}$. In this section we will illustrate how for dynamic linear models with unknown hyperparameter $\boldsymbol{\theta}$ the model likelihood is available as a by-product of a fully Bayesian analysis.

The model likelihood is given by:

$$L(y^N|y^0,\mathcal{M}) = \int p(y_1,\ldots,y_N|y^0,\boldsymbol{\theta},\mathcal{M})p(\boldsymbol{\theta}|y^0,\mathcal{M})\,d\boldsymbol{\theta}. \tag{3}$$

Except for very special cases, e.g. for a fully static model with unknown observation variance, there exists no simple analytical solution of the integration arising in (3) and some method of approximation has to be worked out.

First we rewrite (3) as

$$\begin{aligned} L(y^N|y^0,\mathcal{M}) &= \int c_{\mathcal{M}}(\boldsymbol{\theta})g_{\mathcal{M}}(\boldsymbol{\theta})d\boldsymbol{\theta}, \qquad (4)\\ c_{\mathcal{M}}(\boldsymbol{\theta}) &= \frac{p(y_1,\ldots,y_N|y^0,\boldsymbol{\theta},\mathcal{M})\cdot p(\boldsymbol{\theta}|y^0,\mathcal{M})}{g_{\mathcal{M}}(\boldsymbol{\theta})}. \end{aligned}$$

If we draw an i.i.d. sample $\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(M)}$ from the importance function $g_{\mathcal{M}}(\boldsymbol{\theta})$, then importance sampling Monte Carlo integration of (4) leads to the following Monte Carlo estimate of $L(y^N|y^0,\mathcal{M})$:

$$\hat{L}(y^N|y^0,\mathcal{M}) = \frac{1}{M}\sum_{m=1}^{M}\frac{p(y_1,\ldots,y_N|y^0,\boldsymbol{\theta}^{(m)},\mathcal{M})p(\boldsymbol{\theta}^{(m)}|y^0,\mathcal{M})}{g_{\mathcal{M}}(\boldsymbol{\theta}^{(m)})} = \frac{1}{M}\sum_{m=1}^{M}c_{\mathcal{M}}(\boldsymbol{\theta}^{(m)}). \tag{5}$$

This estimate has the following properties:

$$E_{g_{\mathcal{M}}}(\hat{L}) = L, \tag{6}$$

$$V_{g_{\mathcal{M}}}(\hat{L}) = \frac{L^2}{M}E_{p(\boldsymbol{\theta}|y^N,\mathcal{M})}\left(\frac{p(\boldsymbol{\theta}|y^N,\mathcal{M})}{g_{\mathcal{M}}(\boldsymbol{\theta})} - 1\right), \tag{7}$$

where $p(\boldsymbol{\theta}|y^N,\mathcal{M})$ is the posterior of the unknown hyperparameter $\boldsymbol{\theta}$ given the time series. A proof of (6) and (7) is given in the appendix. If the posterior $p(\boldsymbol{\theta}|y^N,\mathcal{M})$ were known exactly we would choose $g_{\mathcal{M}}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|y^N,\mathcal{M})$ in which

case the approximation is exact even for $M = 1$. Thus if the posterior $p(\boldsymbol{\theta}|y^N, \mathcal{M})$ is available from a fully Bayesian analysis, the model likelihood is in fact an easily available by-product.

Although for dynamic linear models with unknown hyperparameters the posterior $p(\boldsymbol{\theta}|y^N, \mathcal{M})$ has no tractable analytical form except for very special cases, an accurate and yet tractable approximation $g_{\mathcal{M}}(\boldsymbol{\theta})$ may be obtained from Markov chain Monte Carlo methods (Carlin *et al.*, 1992; Frühwirth-Schnatter, 1992). From (7) it is clear that the better $g_{\mathcal{M}}(\cdot)$ approximates the untractable posterior the less the variance of the estimated model likelihood will be.

We now proceed with a description of the Markov chain Monte Carlo approximation of the posterior. $p(\boldsymbol{\theta}|y^N, \mathcal{M})$ is given as an infinite mixture over the posterior density of $\boldsymbol{\theta}$ given the observations y^N *and* a trajectory $\boldsymbol{x}^N = (\boldsymbol{x}_0, \ldots, \boldsymbol{x}_N)$ of the unobservable state process:

$$p(\boldsymbol{\theta}|y^N, \mathcal{M}) = \int p(\boldsymbol{\theta}|\boldsymbol{x}^N, y^N, \mathcal{M}) p(\boldsymbol{x}^N, \boldsymbol{\theta}'|y^N, \mathcal{M}) d\boldsymbol{x}^N d\boldsymbol{\theta}'. \tag{8}$$

Dynamic linear models with unknown hyperparameters are an important special case of what has been called incomplete data problems (Dempster *et al.*, 1976). Whereas the posterior $p(\boldsymbol{\theta}|y^N, \mathcal{M})$ is untractable, the conditional density $p(\boldsymbol{\theta}|\boldsymbol{x}^N, y^N, \mathcal{M})$,

$$p(\boldsymbol{\theta}|\boldsymbol{x}^N, y^N, \mathcal{M}) \propto \prod_{t=1}^{N} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta}, \mathcal{M}) \cdot \prod_{t=1}^{N} p(y_t|\boldsymbol{x}_t, \boldsymbol{\theta}, \mathcal{M}) \cdot p(\boldsymbol{\theta}|\boldsymbol{x}_0, y^0, \mathcal{M}), \tag{9}$$

quite often belongs to a conjugate family. $p(\boldsymbol{\theta}|\boldsymbol{x}_0, y^0, \mathcal{M}) = \pi(\boldsymbol{\theta}; \boldsymbol{\beta}_0)$ and $p(\boldsymbol{\theta}|\boldsymbol{x}^N, y^N, \mathcal{M})$ $= \pi(\boldsymbol{\theta}; \boldsymbol{\beta}_N(\boldsymbol{x}^N, y^N, \boldsymbol{\beta}_0))$ then are members of the same distribution family. For more details see Example 2.1 below.

Based on a sample $(\boldsymbol{x}^N, \boldsymbol{\theta}')^{(m)}$ from the joint posterior $p(\boldsymbol{x}^N, \boldsymbol{\theta}'|y^N, \mathcal{M})$ the infinite mixture (8) is approximated by a finite one:

$$\begin{aligned} p(\boldsymbol{\theta}|y^N, \mathcal{M}) &\approx g_{\mathcal{M}}(\boldsymbol{\theta}), \\ g_{\mathcal{M}}(\boldsymbol{\theta}) &= \frac{1}{M} \sum_{m=1}^{M} \pi(\boldsymbol{\theta}; \boldsymbol{\beta}_N^{(m)}), \quad \boldsymbol{\beta}_N^{(m)} = \boldsymbol{\beta}_N((\boldsymbol{x}^N)^{(m)}, y^N, \boldsymbol{\beta}_0). \end{aligned} \tag{10}$$

A sample from $p(\boldsymbol{x}^N, \boldsymbol{\theta}'|y^N, \mathcal{M})$ is obtained by two different Markov chain Monte Carlo methods. First, we can apply the Gibbs sampler which for each m iterates the two following steps for $n = 1, 2, \ldots$: given $(\boldsymbol{\theta}')_{n-1}^{(m)}$ sample $(\boldsymbol{x}^N)_{n-1}^{(m)}$ from $p(\boldsymbol{x}^N|y^N, (\boldsymbol{\theta}')_{n-1}^{(m)}, \mathcal{M})$ and then sample $(\boldsymbol{\theta}')_n^{(m)}$ from $\pi(\boldsymbol{\theta}; \boldsymbol{\beta}_N((\boldsymbol{x}^N)_{n-1}^{(m)}, y^N, \boldsymbol{\beta}_0))$. An easily implemented forward-filtering-backward-sampling algorithm for drawing a path of the state process $(\boldsymbol{x}^N)^{(\cdot)}$ from $p(\boldsymbol{x}^N|y^N, \boldsymbol{\theta}, \mathcal{M})$ is suggested in Frühwirth-Schnatter (1992, Proposition 1 and 2). This algorithm extends the Gibbs sampler of Carlin *et al.* (1992), which is limited to state space models with regular system variances, to more general models.

An alternative approach is to apply data augmentation methods (Tanner and Wong, 1987) to dynamic linear models (Frühwirth-Schnatter, 1992) where one searches for a fixed-point solution of (8). Based on an 'old' approximation $g_{\mathcal{M}}^{(n-1)}(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}|y^N, \mathcal{M})$, $(\boldsymbol{x}^N, \boldsymbol{\theta}')^{(m)}$ is sampled M_n times from the approximate joint posterior $p(\boldsymbol{x}^N|\boldsymbol{\theta}, y^N, \mathcal{M}) \cdot g_{\mathcal{M}}^{(n-1)}(\boldsymbol{\theta})$ by the method of substitution. A new approximation $g_{\mathcal{M}}^{(n)}(\boldsymbol{\theta})$ is given by the finite mixture (10). The next sample is drawn from this new approximation and so on. This second, adaptive method seems to be more efficient than the non-adaptive Gibbs sampler.

Example 2.1: d-inverse-gamma dynamic linear models. For a large class of dynamic linear models with unknown variances $p(\boldsymbol{\theta}|\boldsymbol{x}^N, y^N, \mathcal{M})$ splits into the product of d densities from an inverted gamma distribution ("d-inverse-gamma models"):

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}^N, y^N, \mathcal{M}) &= \textstyle\prod_{j=1}^{d} p(\theta_j|\boldsymbol{x}^N, y^N, \mathcal{M}), \\ \theta_j|\boldsymbol{x}^N, y^N &\sim \mathrm{IG}(\alpha_N^{(j)}, \beta_N^{(j)}(\boldsymbol{x}^N, y^N)), \end{aligned} \tag{11}$$

if the conjugate prior, too, takes this form:

$$\theta_j|\boldsymbol{x}_0, y^0 \sim \theta_j|y^0 \sim \mathrm{IG}(\alpha_0^{(j)}, \beta_0^{(j)}), \quad j = 1, 2, \ldots, d. \tag{12}$$

A fully Bayesian analysis of this class is carried out in detail in Frühwirth-Schnatter (1992). There it is proved that the adaptive data augmentation procedure described above converges provided that the parameters $\beta_0^{(j)}$ in (12) are positive.

3 Training Sample Priors

For a fully Bayesian analysis of dynamic linear model with unknown hyperparameters one has to choose the joint prior $p(\boldsymbol{x}_0, \boldsymbol{\theta}|y^0, \mathcal{M})$ of the state vector $\boldsymbol{x}_0$ and the unknown hyperparameter $\boldsymbol{\theta}$. One of the main difficulties when computing model likelihoods is to deal with vague priors.

$p(\boldsymbol{x}_0|y^0, \boldsymbol{\theta}, \mathcal{M})$ serves as a starting prior to determine the moments of the conditional posterior $p(\boldsymbol{x}^N|y^N, \boldsymbol{\theta}, \mathcal{M})$ and to compute the conditional likelihood function $p(y_1, \ldots, y_N|y^0, \boldsymbol{\theta}, \mathcal{M})$ in (3) by one run of a Kalman filter. One should avoid modeling vagueness of priors by the large k-approximation (e.g. Harvey, 1989) as it may result in numerical problems. An uninformative prior of the type

$$p(\boldsymbol{x}_0|y^0, \boldsymbol{\theta}, \mathcal{M}) \propto c$$

will not cause any numerical problem, if the likelihood $p(y_1, \ldots, y_N|y^0, \boldsymbol{\theta}, \mathcal{M})$ is computed e.g. by the algorithm of de Jong (1991) and the moments of the conditional posterior $p(\boldsymbol{x}^N|y^N, \boldsymbol{\theta}, \mathcal{M})$ are determined by the forward-filtering-backward-sampling algorithm of Frühwirth-Schnatter (1992).

Whereas one may cope with uninformative prior on the state variable, uninformative priors on the hyperparameter are cumbersome. $p(\boldsymbol{\theta}|\boldsymbol{x}_0, y^0, \mathcal{M})$ serves as

a prior for the conjugate analysis appearing in (9) and its marginal $p(\boldsymbol{\theta}|y^0, \mathcal{M})$ acts as a weight function when computing the model likelihood from (3). If the prior $p(\boldsymbol{\theta}|y^0, \mathcal{M})$ is improper including an unspecified constant, the model likelihood will be proportional to that constant. This seems to introduce some arbitrariness. Spiegelhalter and Smith (1982) introduced "imaginary observations" to cope with vague priors. Following Atkinson (1978) we will use a training sample prior approach. We derive a proper prior for the hyperparameter and the state variable from the first p observations of the time series and then use these proper priors and the remaining (non-training) observations of the time series for model discrimination.

We denote the complete time series by (z^p, y^N) where $z^p = (z_1, \ldots, z_p)$ is the training sample. $\tilde{\boldsymbol{x}}_t$ denotes the state vector for the training sample. The first step is to derive the posterior $p(\tilde{\boldsymbol{x}}_p, \boldsymbol{\theta}|z^p, \mathcal{M}) = p(\tilde{\boldsymbol{x}}_p|z^p, \boldsymbol{\theta}, \mathcal{M}) \cdot p(\boldsymbol{\theta}|z^p, \mathcal{M})$ of the training sample from the improper prior

$$p(\tilde{\boldsymbol{x}}_0, \boldsymbol{\theta}|z^0, \mathcal{M}) \propto c \cdot \pi(\boldsymbol{\theta}; \boldsymbol{\beta}_0),$$

where $\pi(\boldsymbol{\theta}; \boldsymbol{\beta}_0)$ is chosen to be uninformative for the conjugate analysis appearing in (9). The second step is to use this posterior as starting prior for the remaining time series y^N:

$$p(\boldsymbol{x}_0, \boldsymbol{\theta}|y^0, \mathcal{M}) = p(\tilde{\boldsymbol{x}}_p, \boldsymbol{\theta}|z^p, \mathcal{M}).$$

The conditional likelihood $p(y_1, \ldots y_N|y^0, \boldsymbol{\theta})$ is computed from the prior $p(\boldsymbol{x}_0|y^0, \boldsymbol{\theta}, \mathcal{M})$ $= p(\tilde{\boldsymbol{x}}_p|z^p, \boldsymbol{\theta}, \mathcal{M})$ the moments of which are obtained by an inversion filter (Anderson and Moore, 1979). The weight function $p(\boldsymbol{\theta}|y^0, \mathcal{M}) = p(\boldsymbol{\theta}|z^p, \mathcal{M})$ in (3) is approximated by a mixture as in (10):

$$p(\boldsymbol{\theta}|y^0, \mathcal{M}) \approx \frac{1}{M} \sum_{m=1}^{M} \pi(\boldsymbol{\theta}; \boldsymbol{\beta}_p((\tilde{\boldsymbol{x}}^p)^{(m)}, z^p, \boldsymbol{\beta}_0)). \tag{13}$$

Finally, the hyperparameters $\boldsymbol{\theta}^{(m)}$ for which the weight function $p(\boldsymbol{\theta}|y^0, \mathcal{M})$ and the conditional likelihood $p(y_1, \ldots y_N|y^0, \boldsymbol{\theta}, \mathcal{M})$ are evaluated in (5), are sampled from the posterior $p(\boldsymbol{\theta}|z^p, y^N, \mathcal{M})$ which is derived from the whole time series (z^p, y^N) and the prior $p(\boldsymbol{\theta}|z^0, \mathcal{M}) = \pi(\boldsymbol{\theta}; \boldsymbol{\beta}_0)$.

4 Applications

In this section we apply the Bayesian discrimination rule (1) to various non-standard situations where classical likelihood methods seem problematical even asymptotically. Numerical case studies are carried out for two time series published previously in Harvey (1989).

4.1 Bayes Factors for nested and non-nested dynamic linear models

The Bayesian approach of selecting one of two models $\mathcal{M}_1$ and $\mathcal{M}_2$ from the Bayes factor

$$B = \frac{L(y^N|y^0, \mathcal{M}_1)}{L(y^N|y^0, \mathcal{M}_2)}$$

has been extensively discussed in the literature (see among many others: Jeffrey, 1961; Smith and Spiegelhalter, 1980; Spiegelhalter and Smith, 1981; Berger and Delambady, 1987). Bayes factors work for both nested and non-nested models, whereas for non-nested models the use of likelihood ratio statistics seems problematical even asymptotically.

First, we would like to emphazise the close relationship between the Bayesian discrimination rule (1) and the Bayes factor B. If the loss function takes the form $r(\mathcal{M}_1, \mathcal{M}_1) = r(\mathcal{M}_2, \mathcal{M}_2) = 0$, $r(\mathcal{M}_1, \mathcal{M}_2) = r(\mathcal{M}_2, \mathcal{M}_1) = r_0$ and no model is preferred a priori, then from (1) and (2) it is easy to derive that one decides in favour of $\mathcal{M}_1$ if $B \geq 1$:

$$\frac{L(y^N|y^0, \mathcal{M}_1)}{L(y^N|y^0, \mathcal{M}_2)} \geq 1$$

and for $\mathcal{M}_2$ otherwise. Furthermore from (2) we obtain a simple relationship between the posterior probabilities of model $\mathcal{M}_1$ and $\mathcal{M}_2$ and the Bayes factor:

$$P(\mathcal{M}_1|y^N) = \frac{B}{B+1}, \quad P(\mathcal{M}_2|y^N) = \frac{1}{B+1}.$$

From Table 1 it is clear that a Bayes factor much smaller than one gives evidence against the "null hypothesis" $\mathcal{M}_1$ in favour of the alternative $\mathcal{M}_2$, as does a likelihood ratio statistic much smaller that one. On the other hand a Bayes factor much bigger than one gives *evidence in favour of the null hypothesis* $\mathcal{M}_1$ which is never possible with a frequentistic likelihood ratio test.

B	99	19	9	4	2	1	0.5	0.25	$\frac{1}{9}$	$\frac{1}{19}$	$\frac{1}{99}$
$P(\mathcal{M}_1\|y^N)$	0.99	0.95	0.9	0.8	0.667	0.5	0.333	0.2	0.1	0.05	0.01
$P(\mathcal{M}_2\|y^N)$	0.01	0.05	0.1	0.2	0.333	0.5	0.667	0.8	0.9	0.95	0.99

Table 1: Relation between the Bayes factor and the posterior probability of both models

Computing Bayes factors is standard for linear models (see e.g. Smith and Spiegelhalter, 1980). If the model is non-linear in the parameter as is the case for dynamic models with unknown hyperparameters only approximate Bayes factors are

available. Laplace approximations of Bayes factors (Kass and Vaidyanathan, 1992) rely on the asumption of asymptotic normality of the posterior $p(\boldsymbol{\theta}|y^N, \mathcal{M})$. As this regularity quite often is lost in the context of dynamic linear models with unknown hyperparameters (see section 4.2), Laplace's method seems to be of limited use for state space models.

We now proceed with a description of how the material of Section 2 may be applied to approximating Bayes factors for dynamic linear models with unknown hyperparameters. For non-nested models the procedure of Section 2 has to be applied to model $\mathcal{M}_1$ and $\mathcal{M}_2$, seperately. A joint analysis may be possible for nested model $\mathcal{M}_1$ and $\mathcal{M}_2$, where $\mathcal{M}_1$ is obtained by restricting some components of the hyperparameter $\boldsymbol{\theta}_2$ of $\mathcal{M}_2$ to a fixed value $\tilde{\boldsymbol{\theta}}_1^0$. Let $\tilde{\boldsymbol{\theta}}_1$ denote the components of $\boldsymbol{\theta}_2$ which are restricted to $\tilde{\boldsymbol{\theta}}_1^0$. The remaining components of $\boldsymbol{\theta}_2$ clearly are identical with the unknown hyperparameter $\boldsymbol{\theta}_1$ of $\mathcal{M}_1$. If the marginal $\pi(\tilde{\boldsymbol{\theta}}_1; \boldsymbol{\beta})$ of the complete data posterior $\pi(\boldsymbol{\theta}_2; \boldsymbol{\beta}) = \pi(\tilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_1; \boldsymbol{\beta})$ is positive at $\tilde{\boldsymbol{\theta}}_1 = \tilde{\boldsymbol{\theta}}_1^0$:

$$\pi(\tilde{\boldsymbol{\theta}}_1^0; \boldsymbol{\beta}) = \int \pi(\tilde{\boldsymbol{\theta}}_1^0, \boldsymbol{\theta}_1; \boldsymbol{\beta}) d\boldsymbol{\theta}_1 > 0, \quad \forall \boldsymbol{\beta}, \tag{14}$$

there is no need to reiterate for the posterior $p(\boldsymbol{\theta}_1|y^N, \mathcal{M}_1)$ of $\boldsymbol{\theta}_1$ under the smaller model. An approximation $g_{\mathcal{M}_1}(\boldsymbol{\theta}_1)$ to the posterior is directly available from:

$$g_{\mathcal{M}_1}(\boldsymbol{\theta}_1) = g_{\mathcal{M}_2}(\boldsymbol{\theta}_1|\tilde{\boldsymbol{\theta}}_1 = \tilde{\boldsymbol{\theta}}_1^0) = \frac{g_{\mathcal{M}_2}(\tilde{\boldsymbol{\theta}}_1^0, \boldsymbol{\theta}_1)}{g_{\mathcal{M}_2}(\tilde{\boldsymbol{\theta}}_1^0)},$$

$$g_{\mathcal{M}_2}(\tilde{\boldsymbol{\theta}}_1^0) = \frac{1}{M} \sum_{m=1}^{M} \pi(\tilde{\boldsymbol{\theta}}_1^0; \boldsymbol{\beta}_N^{(m)}) > 0.$$

If condition (14) is violated the procedure of section 2 has to be applied to both models, seperately. This is typically the case, if we restrict some components of the bigger model to values on the boundary of the parameter space.

Example 4.1: Purse Snatching in Chicago. Harvey (1989,pp.217,p.516) reanalyzes a time series of reported purse snatchings in the Hyde Park neighbourhood of Chicago by structural time series models:

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2).$$

He compares the local level or steady state model $\mathcal{M}_1$

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

with the local trend model $\mathcal{M}_2$

$$\begin{aligned} \mu_t &= \mu_{t-1} + a_{t-1} + \eta_t, \quad \eta_t \sim \mathrm{N}(0, \sigma_\eta^2), \\ a_t &= a_{t-1} + \xi_t, \quad \xi_t \sim \mathrm{N}(0, \sigma_\xi^2). \end{aligned} \tag{15}$$

We are going to discriminate between these non-nested models by the discrimination procedure described above.

Table 2 displays the model likelihoods for both models together with the posterior probabilities based on equal prior weights $P(\mathcal{M}_1|y^0) = P(\mathcal{M}_2|y^0) = 0.5$ and the Bayes factor. There is substantial support for the local level model $\mathcal{M}_1$.

$\mathcal{M}_i$	$\log \tilde{L}(y^N\|\mathcal{M}_i, y^0)$	$P(\mathcal{M}_i\|y^N)$	Bayes factor
$\mathcal{M}_1$	-234.29	0.8436	5.394
$\mathcal{M}_2$	-235.98	0.1564	0.185

Table 2: Discriminating between the local level and the local trend model for Purse Snatching data

The model likelihoods were estimated from (5) with $M = 1000$. For the first model the hyperparameter equals $\boldsymbol{\theta}_1 = (\sigma_\eta^2, \sigma_\varepsilon^2)$, for the second model $\boldsymbol{\theta}_2 = (\sigma_\eta^2, \sigma_\xi^2, \sigma_\varepsilon^2)$. For both models the first five observations served as training sample.

As both models belong to the d-inverse-gamma-class mentioned in Example 2.1, both the training sample priors $p(\boldsymbol{\theta}_1|y^0, \mathcal{M}_1)$ and $p(\boldsymbol{\theta}_2|y^0, \mathcal{M}_2)$ as well as the approximate posteriors $g_{\mathcal{M}_1}(\boldsymbol{\theta}_1)$ and $g_{\mathcal{M}_2}(\boldsymbol{\theta}_2)$ were derived by the iterative method suggested in Frühwirth-Schnatter (1992). By the way, Harvey (1989, p.218) reports problems with estimating the variance σ_ξ^2 in the local trend model $\mathcal{M}_2$. These problems become clear from the Bayesian approach. Figure 1 compares the marginal posteriors $p(\sigma_\varepsilon^2|y^N, \mathcal{M}_2)$ and $p(\sigma_\xi^2|y^N, \mathcal{M}_2)$. Whereas the first one is already close to approximate normality, the posterior of σ_ξ^2 is skew, has several peaks and is concentrated close to 0. No wonder that methods that rely on asymptotic normality fail.

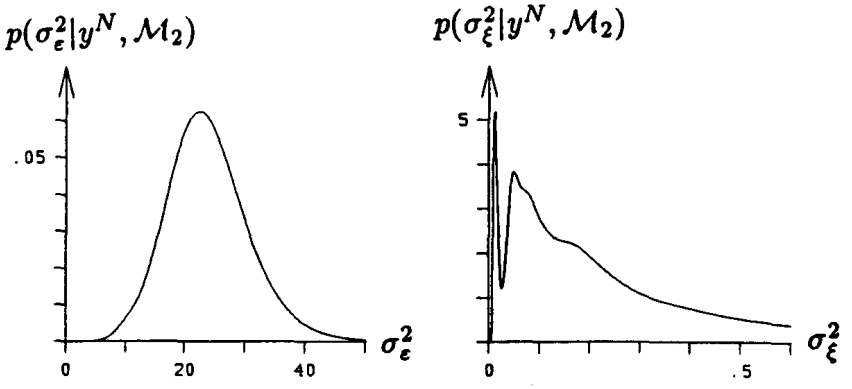


Figure 1: Marginal posteriors $p(\sigma_\varepsilon^2|y^N, \mathcal{M}_2)$ and $p(\sigma_\xi^2|y^N, \mathcal{M}_2)$ of the variances of a trend model fitted to the purse snatching data

4.2 Testing Hypotheses on the Boundary of the Parameter Space

We will now discuss a problem which may occur in the context of state space modelling and which – as far as we know – has not been discussed before from a Bayesian point of view. We mean testing hypotheses on the variances of a structural time series model. This problem is discussed in detail from a frequentistic point of view in Harvey (1989).

To give an idea of the problem consider e.g. the *basic structural model* defined in Harvey (1989) to model a time series $\{y_1, y_2, \ldots, y_N\}$:

$$\begin{aligned} y_t &= \mu_t + s_t + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2), \\ \mu_t &= \mu_{t-1} + a_{t-1} + \eta_t, \quad \eta_t \sim \mathrm{N}(0, \sigma_\eta^2), \\ a_t &= a_{t-1} + \xi_t, \quad \xi_t \sim \mathrm{N}(0, \sigma_\xi^2), \\ s_t &= -\textstyle\sum_{j=1}^{k-1} s_{t-j} + \omega_t, \quad \omega_t \sim \mathrm{N}(0, \sigma_\omega^2). \end{aligned} \tag{16}$$

If all variances are positive, the model combines a random walk mean μ_t which is shifted by a dynamic stochastic trend component a_t with a dynamic saisonal component s_t. It may be of practical interest to decide whether parts of this model are really dynamic for a given time series. This problem is equivalent to testing whether some of the variances appearing in (16) are equal to 0 or not. To decide e.g. whether the saisonal pattern changes in time or not corresponds to test the hypothesis $\sigma_\omega^2 > 0$ against $\sigma_\omega^2 = 0$. It is well known that classical testing procedures such as likelihood ratio tests may loose their asymptotic χ^2-distribution, if some components of the parameter of interest lie on the boundary of the parameter space – as is the case when testing whether a variance is 0 or not. Harvey (1989) points out that in some cases (e.g. testing $\sigma_\eta^2 > 0$ against $\sigma_\eta^2 = 0$, if all other variances are positiv) modified likelihood ratio tests may be constructed, whereas in other cases (e.g. testing $\sigma_\eta^2 > 0$ against $\sigma_\eta^2 = 0$, if σ_ω^2 and σ_ε^2 are positiv and σ_ξ^2 is 0) no such modifications are possible.

Let $\boldsymbol{\theta}_2$ denote the vector containing all unknown variances of the bigger (unrestricted model) $\mathcal{M}_2$. $\mathcal{M}_2$ itself may be the restriction of an even bigger model. Let $\tilde{\boldsymbol{\theta}}_1$ denote the vector containing those variances which are restricted to 0. We then suggest to solve the problem of testing between the hypotheses $\mathcal{M}_1 : \tilde{\boldsymbol{\theta}}_1 = 0$ and $\mathcal{M}_2 : \tilde{\boldsymbol{\theta}}_1 > 0$ by means of the Bayesian discrimination rule (1).

It should be pointed out that hypotheses on variances typically correspond to models from the d-inverse-gamma class (see Example 2.1). If this is the case we have to compute the Bayes factor seperately for $\mathcal{M}_1$ and $\mathcal{M}_2$ as condition (14) is violated. From (11) we realize that the marginal $\pi(\tilde{\boldsymbol{\theta}}_1; \boldsymbol{\beta})$ of the complete data posterior $p(\boldsymbol{\theta}_2 | \boldsymbol{x}^N, y^N, \mathcal{M}_2)$ is a product of inverted gamma densities and therefore equal to 0 at $\tilde{\boldsymbol{\theta}}_1 = \tilde{\boldsymbol{\theta}}_1^0 = 0$. We would like to emphazise once more that even if $\mathcal{M}_1$ holds and asymptotic normality of $p(\tilde{\boldsymbol{\theta}}_1 | y^N, \mathcal{M}_2)$ under $\mathcal{M}_2$ is lost, iteration for the posterior will converge to the true non-normal density (see Figure 1). Therefore the non-regularity of the problem will not cause any additional problems.

Example 4.2: UK-coal consumption. For illustration we consider the time series of UK-coal consumption modelled by a basic structural model in Harvey (1989, p.95, pp.512) A fully Bayesian analysis has been carried out in Frühwirth-Schnatter (1992) where evidence for the hypothesis $\sigma_\eta^2 = \sigma_\omega^2 = 0$ was found from visual inspection of the approximate posterior $p(\sigma_\eta^2, \sigma_\xi^2, \sigma_\omega^2, \sigma_\varepsilon^2 | y^N)$.

For a more formal decision between the hypotheses

$$\begin{aligned} \mathcal{M}_1 : \quad & \sigma_\eta^2 = 0, \sigma_\omega^2 = 0, \\ \mathcal{M}_2 : \quad & \sigma_\eta^2 > 0, \sigma_\omega^2 > 0, \end{aligned}$$

we will apply the discrimination procedure described above. Table 3 displays the model likelihoods for both models together with the posterior probabilities based on equal prior weights $P(\mathcal{M}_1|y^0) = P(\mathcal{M}_2|y^0) = 0.5$ and the Bayes factor. There is a strong support for the "null hypothesis" $\mathcal{M}_1$. The model likelihoods were estimated from (5) with $M = 1000$. For the first model the hyperparameter equals $\boldsymbol{\theta}_1 = (\sigma_\xi^2, \sigma_\varepsilon^2)$, for the second model $\boldsymbol{\theta}_2 = (\sigma_\eta^2, \sigma_\xi^2, \sigma_\omega^2, \sigma_\varepsilon^2)$. For both models the first 8 observations served as training sample.

$\mathcal{M}_i$	$\log \hat{L}(y^N \mid \mathcal{M}_i, y^0)$	$P(\mathcal{M}_i \mid y^N)$	Bayes factor
$\mathcal{M}_1$	41.2	0.937	14.87
$\mathcal{M}_2$	38.5	0.063	0.067

Table 3: Testing on the boundary of the parameter space for the UK-coal consumption time series

4.3 Discrimination between more than two models

Obviously, Bayesian discrimination is not restricted to two models. An interesting special case of discrimination of more than two models occurs, if the risk function takes the form

$$r(\mathcal{M}_q, \mathcal{M}_l) = (1 - \delta_{ql}) \cdot r_0. \tag{17}$$

It is easy to verify that the expected loss of decision for $\mathcal{M}_q$ is equal to:

$$r_0 \cdot \sum_{l=1, l \neq q}^{L} P(\mathcal{M}_l | y^N) = r_0 \cdot \left(1 - P(\mathcal{M}_q | y^N)\right).$$

Therefore the loss is minimal for the model with maximum posterior probability. Thus the heuristic strategy of taking the model with the biggest posterior probability turns out to be the optimal strategy under risk function (17). If no model is preferred apriori, the model with the biggest model likelihood turns out to be the best choice under risk function (17).

To carry out discrimination in practice, we again suggest to apply the procedure of section 2. If none of the L models are nested, we have to carry out the procedure seperately for all L models. If some of the models are nested and conditions comparable to condition (14) hold between the nested models, one could save a lot of computing time by starting with the biggest model in a tree and approximate the posterior of the unknown hyperparameter of the smaller model from the approximate posterior of the bigger one.

Example 4.3: UK-coal consumption (continued). Apart from the two hypotheses studied in Example 4.2 it seems interesting to take other hypotheses, such as e.g. $\mathcal{M}_3 : \sigma_\xi^2 = 0$ or $\mathcal{M}_4 : \sigma_\eta^2 = 0$ into consideration. For a basic structural model there exist 8 possibilities of combining static and dynamic components of the state variable (see Table 4). For illustration of the material of this subsection we show how to discriminate among them for the UK-coal consumption time series based on equal prior chances $P(\mathcal{M}_i|y^0) = 0.125$. Table 4 compares the model likelihood and the posterior probabilities of each model. Based on the risk function (17) we decide for model $\mathcal{M}_1$ with the biggest posterior probability. This strategy leads to the choice of a basic structural model with a dynamic trend component, a static saisonal component and σ_η^2 =0.

$\mathcal{M}_i$	$\log \hat{L}(y^N\|\mathcal{M}_i, y^0)$	$P(\mathcal{M}_i\|y^N)$
$\mathcal{M}_1 : \sigma_\eta^2 = 0, \sigma_\xi^2 > 0, \sigma_\omega^2 = 0, \sigma_\varepsilon^2 > 0$	41.18	0.678
$\mathcal{M}_2 : \sigma_\eta^2 > 0, \sigma_\xi^2 > 0, \sigma_\omega^2 > 0, \sigma_\varepsilon^2 > 0$	38.50	0.0464
$\mathcal{M}_3 : \sigma_\eta^2 > 0, \sigma_\xi^2 = 0, \sigma_\omega^2 > 0, \sigma_\varepsilon^2 > 0$	38.25	0.0361
$\mathcal{M}_4 : \sigma_\eta^2 = 0, \sigma_\xi^2 > 0, \sigma_\omega^2 > 0, \sigma_\varepsilon^2 > 0$	39.20	0.0938
$\mathcal{M}_5 : \sigma_\eta^2 > 0, \sigma_\xi^2 > 0, \sigma_\omega^2 = 0, \sigma_\varepsilon^2 > 0$	38.67	0.055
$\mathcal{M}_6 : \sigma_\eta^2 = 0, \sigma_\xi^2 = 0, \sigma_\omega^2 > 0, \sigma_\varepsilon^2 > 0$	27.75	$0.1 \cdot 10^{-5}$
$\mathcal{M}_7 : \sigma_\eta^2 > 0, \sigma_\xi^2 = 0, \sigma_\omega^2 = 0, \sigma_\varepsilon^2 > 0$	39.17	0.0907
$\mathcal{M}_8 : \sigma_\eta^2 = 0, \sigma_\xi^2 = 0, \sigma_\omega^2 = 0, \sigma_\varepsilon^2 > 0$	30.72	$0.2 \cdot 10^{-4}$

Table 4: Discriminating eight models for the UK-coal consumption time series

The numerical values obtained for the model likelihood illustrate how the model likelihood $L(y^N|y^0, \mathcal{M}_i)$ differs in an important way from the maximum of the conditional likelihood $p(y_1, \ldots, y_N|y^0, \hat{\theta}_i, \mathcal{M}_i)$. It is well known that the maximum of the conditional likelihood would increase within nested models, if the number of parameters increases. Some penalty term including the number of parameters has to be introduced to allow model selection from the maximum of the conditional likelihood (Akaike, 1973; Schwarz, 1978). Following the path $\mathcal{M}_8 \to \mathcal{M}_1 \to \mathcal{M}_5 \to \mathcal{M}_2$ in Table 4 nicely demonstrates that penalty of too many parameters is an intrinsic property of the model likelihood which is derived as a *weighted* and not as a *maximum* conditional likelihood.

5 Concluding Remarks

Among further possible applications we would like to mention the problem of testing constancy of regression coefficients over time (see e.g. Hackl, 1980). The Bayesian discrimination procedure may be applied to a set of different dynamic regressison models, which combine static and dynamic parameters in various ways. If the dimension d of the parameter is high, it may be tiresome to check all 2^d different possibilities. In such a case visual inspection of the $\binom{d}{2}$ bivariate marginal posteriors of the variances in a fully dynamic regression model could be of great help in preselecting possible hypotheses.

In this paper we have confined ourselves to normal linear state space models. The extension to non-normal and/or non-linear state space models will follow along the same lines (e.g. use a density estimate of the posterior of hyperparameters as importance function), however some technical details still need to be clarified. The paper of Carlin *et al.* seems to be a fruitful starting point for doing so.

In the paper the problem of choosing the prior $p(\boldsymbol{\theta}|y^0, \mathcal{M})$ was circumvented by a training sample prior approach. It would be extremely interesting to analyze how this choice compares to subjective, non-training sample priors. Table 5 summarizes the results of a small simulation experiment on the power of discriminating between the hypotheses $\mathcal{M}_1 : \sigma_\eta^2 = 0$ and $\mathcal{M}_2 : \sigma_\eta^2 > 0$ for time series from a local linear trend model (15) with $\sigma_\xi^2 = 0$ and $\sigma_\eta^2 \geq 0$. 100 time series were simulated from the model with $\sigma_\varepsilon^2 = 1$, $\mu_0 = 100$, $a_0 = 3$ and various values of σ_η^2. Table 5 shows the relative frequency of decision for $\mathcal{M}_2$ based on $P(\mathcal{M}_1|y^0) = P(\mathcal{M}_2|y^0) = 0.5$ and symmetric risk function (17) for various priors. The training sample prior performed better than the first subjective prior. However, both priors lead to a conservative test: they never reject the "null hypothesis" within 100 trials of time series for which the "null hypothesis" hold. The second subjective prior has a higher power of decting the "alternative hypothesis" than the other two priors at the cost of a higher probability of rejecting a true "null hypothesis". A formal discussion of the proposed procedure's power is far beyond the scope of this paper. This surely is a challenging problem for theoretical statisticians.

priors	true value of σ_η^2			
	$\sigma_\eta^2 = 0$	$\sigma_\eta^2 = 0.01$	$\sigma_\eta^2 = 0.1$	$\sigma_\eta^2 = 0.5$
training sample prior	0.	0.17	0.85	1.
subjective prior 1: $\alpha_0^{(j)} = 0.01, \beta_0^{(j)} = 10^{-5}$	0.	0.11	0.72	0.98
subjective prior 2: $\alpha_0^{(j)} = 1, \beta_0^{(j)} = 0.01$	0.09	0.39	0.92	1.

Table 5: Comparing the power of Bayesian discrimination for different prior choices

Appendix: Proof of (6) and (7)

1. $E_{g_{\mathcal{M}}}(\hat{L}) = \frac{1}{M}\sum_{m=1}^{M}\int p(y_1,\ldots,y_N|y^0,\boldsymbol{\theta},\mathcal{M})p(\boldsymbol{\theta}|y^0,\mathcal{M})d\boldsymbol{\theta} = L.$

2. As $\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(M)}$ are independent, the following holds:

$$\begin{aligned} V_{g_{\mathcal{M}}}(\hat{L}) &= \frac{1}{M^2}\sum_{m=1}^{M}\int \frac{p(y_1,\ldots,y_N|y^0,\boldsymbol{\theta},\mathcal{M})p(\boldsymbol{\theta}|y^0,\mathcal{M})}{g_{\mathcal{M}}(\boldsymbol{\theta})}d\boldsymbol{\theta} - \frac{1}{M}L^2 = \\ &= \frac{L^2}{M}\left[\int \frac{p(\boldsymbol{\theta}|y^N,\mathcal{M})}{g_{\mathcal{M}}(\boldsymbol{\theta})}p(\boldsymbol{\theta}|y^N,\mathcal{M})d\boldsymbol{\theta} - 1\right] = \\ &= \frac{L^2}{M}E_{p(\boldsymbol{\theta}|y^N,\mathcal{M})}\left(\frac{p(\boldsymbol{\theta}|y^N,\mathcal{M})}{g_{\mathcal{M}}(\boldsymbol{\theta})} - 1\right). \end{aligned}$$

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *2nd Int. Symp. Inf. Theory* (B.N. Petrov and F. Czáki, eds), 267-281. Budapest: Akad. Kiadoó.

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

Atkinson, A.C. (1978). Posterior Probabilities for Choosing a Regression Model. *Biometrica*, **65**, 39-48.

Berger, J. and Delambady, M. (1987). Testing Precise Hypothesis (with discussion). *Statist. Sci.*, **2**, 317-335.

Carlin, B., Polson, N. and Stoffer, D. (1992). A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling. *J.Am.Statist.Ass.*, **87**, 493-500.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1976). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *J.R.Stat.Soc. B*, **39**, 1-38.

Frühwirth-Schnatter, S. (1992). *Data Augmentation and Dynamic Linear Models.* Preprint submitted for publication.

Geisser, S. and Eddy, W.F. (1979). A Predictive Approach to Model Selection. *J.Am.Statist.Ass.*, **74**, 153–160.

Hackl, P. (1980). *Testing the Constancy of Regression Models over Time.* Göttingen: Vandenhoeck & Ruprecht.

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

Jeffreys, H. (1961). *Theory of Probability.* Oxford: Oxford University Press.

de Jong, P. (1991). The Diffuse Kalman Filter. *Ann. Statist.*, **19**, 1037-1083.

Kass, R. and S.K. Vaidyanathan (1992). Approximate Bayes Factors and Orthogonal Parameters, with Application to Testing Equality of Two Binomial Proportions. *J.R. Statist. Soc. B*, **54**, 129-144.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Statist.*, **6**, 461-464.

Smith, A.F.M. and Spiegelhalter, D.J. (1980). Bayes Factors and Choice Criteria for Linear Models. *J.R. Statist. Soc. B*, **42**, 213-220 .

Spiegelhalter, D.J. and Smith, A.F.M. (1982). Bayes Factors for Linear and Log-linear Models with Vague Prior Information. *J.R. Statist. Soc. B*, **44**, 377-387.

Tanner, M. (1991). *Tools for Statistical Inference. Observed Data and Data Augmentation Methods.* Lecture Notes in Statistics, **67**. New York: Springer.

Tanner, M. and Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *J.Am.Statist.Ass.*, **82**, 528–550.

West, M. and Harrison, P.J. (1989). *Bayesian Forecasting and Dynamic Models.* New York/Heidelberg/Berlin: Springer.