Bayesian Analysis (2008) **3**, Number 4, pp. 689–698

Comment on article by Rydén

Sylvia Frühwirth-Schnatter*

First of all, I would like to congratulate the author on an excellent paper providing a very fair comparison of EM versus MCMC estimation for hidden Markov models. I agree with most of what has been said in the paper, nevertheless I would like to comment on several issues, including estimating the unknown number of states using marginal likelihoods, choosing the prior and post-processing the MCMC draws to deal with label switching. Before doing so, I would like to introduce `bayesf`, a MATLAB software package for estimating hidden Markov models.

1 The MATLAB software package `bayesf`

In January 2007 I finished a first version (Version 1.0) of a MATLAB software package called `bayesf`. This package allows to perform Bayesian inference for some of the finite mixture and Markov switching models discussed in my Springer monograph (Frühwirth-Schnatter 2006). An updated version (Version 2.0) is close to be finished and will be available on my web site at

http://www.ifas.jku.at/e2571/e2626/e2632/index_ger.html

by the time this discussion is published. `bayesf` allows MCMC estimation of a hidden Markov model for a variety of conditional distributions, including multivariate normal distributions, Student-t distributions, but also discrete-valued distributions like the Poisson distribution. The unknown number of states is estimated through marginal likelihoods (rather than RJMCMC).

2 Estimating the Unknown Number of States

My first contribution to the discussion concerns estimating the unknown number of states of the hidden Markov model for the S&P 500 data studied in Section 2 of the paper. For these data, I want to discuss two issues, namely choosing an appropriate prior and computing marginal likelihoods.

I consider three different prior choices for the parameters of the hidden Markov model to assess the influence of the prior. "Prior 1" which is closely related to the prior considered in the paper assumes that the state specific standard deviations σ_i are equipped with independent uniform $\mathcal{U}[0, \max|y_k|]$ priors while each row of the transition matrix A follows a $\text{Dir}(1, \ldots, 1)$ prior. "Prior 2" assumes a hierarchical prior for σ_i^2, namely $\sigma_i^{-2} \sim \Gamma(\alpha, \beta)$ where $\alpha = 2$ and $\beta \sim \Gamma(g, h)$ with $g = 0.5$ and $h = 25/R^2$ where $R = \max y_k - \min y_k$ is the data range, while the prior for the rows of the

*Department of Applied Statistics and Econometrics, Johannes Kepler Universität Linz, Linz, Austria, mailto:sylvia.fruehwirth-schnatter@jku.at

 DOI:10.1214/08-BA326A

Table 1: S&P 500 data; log of various estimates of the marginal likelihood $p(\mathbf{y}_{1:n}|d)$ for Prior 1 obtained from three independent MCMC runs; BS ... bridge sampling, IS ... importance sampling, RI ... reciprocal importance sampling; standard errors in parenthesis

d	1	2	3	4	5
BS	-1716.	-1625.1(0.01)	-1625.0(0.04)	-1637.3(0.05)	-1649.6(0.05)
	-1716.	-1625.1(0.01)	-1625.6(0.05)	-1636.5(0.05)	-1649.7(0.05)
	-1716.	-1625.1(0.00)	-1625.1(0.04)	-1636.4(0.05)	-1649.6(0.05)
IS	-1716.	-1625.1(0.01)	-1624.0(0.59)	-1635.7(0.37)	-1644.5(0.62)
	-1716.	-1625.1(0.01)	-1624.1(0.47)	-1635.5(0.26)	-1646.2(0.49)
	-1716.	-1625.1(0.01)	-1623.8(0.63)	-1629.8(0.99)	-1645.4(0.61)
RI	-1716.	-1625.1(0.01)	-1630.6(0.47)	-1646.0(0.72)	-1664.1(0.88)
	-1716.	-1625.1(0.03)	-1630.6(0.22)	-1644.1(0.34)	-1662.8(0.72)
	-1716.	-1625.1(0.01)	-1630.3(0.32)	-1645.8(0.57)	-1661.2(0.70)

transition matrix A is the same as for "Prior 1". This prior is closely related to the prior considered in the paper for Case I. Finally, "Prior 3" combines the hierarchical prior for state specific variances σ_i^2 with a slightly more informative prior on the transition matrix A by assuming that each row $(a_{i1}, \ldots, a_{id})$, $i = 1, \ldots, d$ follows a Dir $(e_{i1}, \ldots, e_{id})$ prior where $e_{ii} = 4$ and $e_{ij} = 1/(d-1)$ for $i \neq j$. By choosing $e_{ii} > e_{ij}$ the hidden Markov model is bounded away from a finite mixture model, see Frühwirth-Schnatter (2006, Subsection 11.5.1). This prior is the default choice in the `bayesf` package. For all priors, I assume that the vector ρ of the initial states is drawn from the ergodic distribution of the hidden Markov chain rather than from an independent Dir $(1, \ldots, 1)$ prior as in the paper.

It will turn out that the choice of the prior is far more influential on the posterior probabilities $p(d|\mathbf{y}_{1:n})$ of the number d of states than one might expect. To compute the posterior probabilities $p(d|\mathbf{y}_{1:n})$, I consider estimating the marginal likelihoods for a certain range of d rather than using RJMCMC as was done in the paper.

2.1 Computing Marginal Likelihoods for Unknown Number of States

As discussed in the paper, there exist many ways to approximate the marginal likelihood using MCMC draws. Ever since Gelfand and Dey (1994) introduced the harmonic mean estimator there has been a desire to find estimators that are based solely on the MCMC draws and the mixture likelihood function evaluated at the MCMC, because such an estimator is extremely cheap to evaluate and very easy to program. Unfortunately, it seems impossible to find such an estimator. Recently Scott (2002) providing a nice survey on recursive algorithms mentioned also in the introduction of the paper suggested an estimator along these lines as did Congdon (2006), however, these estimators are biased as shown by Robert and Marin (2008).

Table 2: S&P 500 data; log of various estimates of the marginal likelihood $p(\mathbf{y}_{1:n}|d)$ for Prior 2 obtained from three independent MCMC runs; BS ... bridge sampling, IS ... importance sampling, RI ... reciprocal importance sampling; standard errors in parenthesis

d	1	2	3	4	5
BS	-1717.6	-1628.4(0.00)	-1630.0(0.04)	-1643.4(0.05)	-1657.2(0.05)
	-1717.6	-1628.4(0.01)	-1630.0(0.04)	-1643.2(0.05)	-1657.6(0.05)
	-1717.6	-1628.4(0.00)	-1629.9(0.04)	-1643.1(0.05)	-1656.8(0.05)
IS	-1717.6	-1628.4(0.01)	-1629.5(0.30)	-1641.9(0.35)	-1652.7(0.73)
	-1717.6	-1628.4(0.01)	-1630.1(0.17)	-1640.1(0.47)	-1655.5(0.29)
	-1717.6	-1628.4(0.01)	-1628.7(0.62)	-1641.7(0.47)	-1653.4(0.58)
RI	-1717.6	-1628.5(0.01)	-1634.7(0.30)	-1653.3(0.45)	-1669.7(0.57)
	-1717.6	-1628.5(0.01)	-1634.1(0.27)	-1650.7(0.28)	-1669.5(0.64)
	-1717.6	-1628.4(0.01)	-1634.3(0.33)	-1651.2(0.38)	-1669.1(0.72)

Thus, I focus on three of the simulation-based approaches discussed in Frühwirth-Schnatter (2006, Section 5.4), namely bridge sampling, importance sampling and reciprocal importance sampling. Bridge sampling is described in detail in the paper, but has not been applied to the S&P 500 data.

The importance density $q(\theta)$ underlying bridge sampling and the other estimators is constructed from the conditional densities appearing in the Gibbs sampler[1] and reads:

$$q(\theta) = \frac{1}{S}\sum_{s=1}^{S} p(\theta|\mathbf{x}_{0:n}^{[s;d]}, \mathbf{y}_{1:n}, d),$$

where the draws $\mathbf{x}_{0:n}^{[s;d]}$, $s = 1, \ldots, S$ are selected randomly from the full MCMC chain $\mathbf{x}_{0:n}^{[m;d]}$, $m = 1, \ldots, M$. Since the complete-data posterior $p(\theta|\mathbf{x}_{0:n}, \mathbf{y}_{1:n}, d)$ usually factors into the product of closed form densities, it is sufficient to store the moments of these densities rather than the entire path of the states $\mathbf{x}_{0:n}$ of the hidden Markov chain.

One advantage of constructing $q(\theta)$ in this way is that the importance density exhibits the same kind of multimodality as the posterior density $p(\theta|\mathbf{y}_{1:n}, d)$, provided that each sweep of the Gibbs sampler is ended by randomly permuting the labels of the states as in Frühwirth-Schnatter (2001). The number m_d of modes in the posterior $p(\theta|\mathbf{y}_{1:n}, d)$ is usually equal to $d!$, if n is large enough and the model is not overfitting, meaning that d is not larger than the true number of states. For small data sets and for overfitting models there may be more or less than $d!$ modes.

It is desirable to choose the number S of components as small as possible, because $q(\theta)$ has to be evaluated at each of the M MCMC draws and at each of the L draws from

[1] Subsequently I use the notation of the present paper rather than the notation applied in Frühwirth-Schnatter (2006).

Table 3: S&P 500 data; log of various estimates of the marginal likelihood $p(\mathbf{y}_{1:n}|d)$ for Prior 3 obtained from three independent MCMC runs; BS ... bridge sampling, IS ... importance sampling, RI ... reciprocal importance sampling; standard errors in parenthesis

d	1	2	3	4	5
BS	-1717.6	-1626.0(0.00)	-1621.3(0.03)	-1626.5(0.05)	-1634.0(0.06)
	-1717.6	-1626.0(0.00)	-1621.3(0.03)	-1626.9(0.05)	-1634.4(0.05)
	-1717.6	-1626.0(0.00)	-1621.3(0.03)	-1626.5(0.05)	-1634.6(0.06)
IS	-1717.6	-1626.0(0.01)	-1621.3(0.17)	-1624.9(0.49)	-1630.3(0.35)
	-1717.6	-1626.0(0.01)	-1621.3(0.08)	-1624.9(0.32)	-1628.2(0.71)
	-1717.6	-1626.0(0.01)	-1621.3(0.14)	-1622.9(0.90)	-1629.9(0.64)
RI	-1717.6	-1626.1(0.02)	-1626.5(0.45)	-1640.9(0.99)	-1647.0(0.40)
	-1717.6	-1626.0(0.01)	-1626.3(0.39)	-1635.1(0.28)	-1654.4(0.95)
	-1717.6	-1626.1(0.02)	-1625.2(0.40)	-1636.2(0.45)	-1648.6(0.50)

the importance density in order to compute the bridge sampling estimator, leading to a total of $S(M+L)$ functional evaluations of the complete-data posterior $p(\theta|\mathbf{x}_{0:n}, \mathbf{y}_{1:n}, d)$. On the other hand, it is essential that all modes of the posterior density are covered by the importance density also for increasing values of d to avoid instability of the resulting estimators. On the average, each mode is covered by $M_0 = S/m_d$ components of $q(\theta)$. This means that the whole approach is limited to moderate values of d, say up to 5 or 6, and does not provide a feasible alternative to RJMCMC if the number of states is larger than that.

The appropriate choice of S also depends on the amount of missing information. Around a particular mode, $p(\theta|\mathbf{y}_{1:n}, d)$ is approximated by a mixture of $M_0 = S/m_d$ complete-data posteriors $p(\theta|\mathbf{x}_{0:n}, \mathbf{y}_{1:n}, d)$. If the complete data $(\mathbf{x}_{0:n}, \mathbf{y}_{1:n})$ are much more informative about θ than the observed data $\mathbf{y}_{1:n}$ alone, then a larger value of M_0 is needed to obtain a reliable importance density than in cases were the amount of missing information is small.

Subsequently, I compute all three estimators of the marginal likelihood for the S&P500 data for $d = 1, \ldots, 5$. The estimators are based on $M = 8,000$ MCMC draws after a burn-in of 4,000 draws and on $L = 8,000$ draws from the importance density $q(\theta)$. The total number of draws is equal to $5(M + L + 4000) = 100,000$ which is comparable to the 100,000 sweeps of the RJMCMC sampler considered in the paper. To construct $q(\theta)$, I choose $S = M_0 d!$ where $M_0 = 100$ for $d \leq 3$, while $M_0 = 25$ for $d = 4$ and $M_0 = 5$ for $d = 5$. Thus for $d \geq 3$ $S = 600$ components are used, regardless of d.

Computations were carried out running the MATLAB package `bayesf` on a notebook with a 2.0 GHz processor. Running MCMC and performing marginal likelihood estimation for $d = 1, \ldots, 5$ required in total 57 CPU minutes which is only a fraction of the 19 hours reported in the paper for RJMCMC. To evaluate the stability of each es-

Table 4: S&P data; posterior distribution of the number of states for various priors on the parameters of the hidden Markov model

d	1	2	3	4	5
Prior 1					
$\hat{p}_{BS}(d\|\mathbf{y}_{1:n})$	0.0	0.482	0.518	0.0	0.0
	0.0	0.633	0.367	0.0	0.0
	0.0	0.512	0.488	0.0	0.0
Prior 2					
$\hat{p}_{BS}(d\|\mathbf{y}_{1:n})$	0.0	0.833	0.167	0.0	0.0
	0.0	0.825	0.175	0.0	0.0
	0.0	0.807	0.193	0.0	0.0
Prior 3					
$\hat{p}_{BS}(d\|\mathbf{y}_{1:n})$	0.0	0.009	0.985	0.006	0.0
	0.0	0.009	0.987	0.004	0.0
	0.0	0.009	0.985	0.006	0.0

timator, MCMC sampling and subsequent marginal likelihood estimation was repeated independently three times.

Table 1 to Table 3 report the various estimators of the (log) marginal likelihood for three independent runs under each prior. Additionally, standard errors were computed as described in Frühwirth-Schnatter (2006, p. 152). While all estimators agree for $d = 1$ and $d = 2$, considerable differences are present for $d \geq 3$. As observed in previous work, we find that bridge sampling is far more stable and more precise than importance or reciprocal importance sampling. For these estimators we find considerable differences in the estimated value for independent runs and much larger standard errors in particular for overfitting models.

2.2 Studying the Influence of the Prior

I will now proceed with studying the influence of the prior on the parameters of the hidden Markov model on the posterior probabilities $p(d|\mathbf{y}_{1:n})$ of the number d of states. For each prior, the marginal likelihoods $p(\mathbf{y}_{1:n}|d)$, estimated through bridge sampling as in Table 1 to Table 3, are combined with a uniform prior over $d \in \{1, \ldots, 5\}$ which is also used in the paper in connection with RJMCMC. From Table 4 we find that the prior exercises considerable influence on the posterior distribution of the number of states and the stability of the estimated posterior probabilities over independent MCMC runs.

For "Prior 1" which is very similar to the prior used in the paper for RJMCMC no clear decision concerning the number of states is possible. In their inability to discriminate clearly between $d = 2$ and $d = 3$ states the posterior distributions are similar to the one reported in the paper for RJMCMC. The estimated posterior probabilities reported

in Table 4 differ considerably over independent runs, the difference between the largest and the smallest value of $\hat{p}_{BS}(2|\mathbf{y}_{1:n})$ being as large as 15.1%. This instability leads to choosing $d = 3$ for the first and $d = 2$ for the remaining two runs, each time with a high risk for a wrong decision.

I would also like to mention that I experienced difficulties with "Prior 1" insofar as Gibbs sampling tended to get stuck at a mode where some of the variances were equal to 0. This happened, for instance, for $d = 3$ for one out of four independent runs. Also, when I tried to combine the uniform prior for $\sigma_1, \ldots, \sigma_d$ with the slightly more informative prior on the transition matrix A I ran into troubles, in particular, for $d = 4$, because the Gibbs sampler got stuck for three independent runs.

No such problems occurred under "Prior 2" and "Prior 3" which both assume a hierarchical prior for the variances $\sigma_1^2, \ldots, \sigma_d^2$ rather than a uniform prior for $\sigma_1, \ldots, \sigma_d$. As discussed in Frühwirth-Schnatter (2006, Subsection 6.2.3) for finite mixture models, under this prior the ratio of any two variances σ_i^2/σ_j^2 follows an $F(2\alpha, 2\alpha)$ distribution and thus is stochastically bounded which helps to overrule spurious local modes and the unboundedness of the mixture likelihood function. The same, of course, applies to hidden Markov models.

For "Prior 2" and "Prior 3" the posterior distributions in Table 4 are much more concentrated than under "Prior 1" and stability over independent runs increases. For "Prior 2" the difference between the largest and the smallest value of $\hat{p}_{BS}(2|\mathbf{y}_{1:n})$ reduces to 2.6%, while the posterior probabilities are extremely stable for "Prior 3".

However, we also find that under a hierarchical prior on the variances the prior on the rows of the transition matrix exercises a tremendous influence on the selected number of states. In combination with the flat prior ("Prior 2") one would choose $d = 2$, in combination with the prior which bounds the hidden Markov model away from a finite mixture model ("Prior 3") one would choose $d = 3$ with high confidence.

If one is willing to view the prior as part of the model, one could consider "prior selection" based on the marginal likelihoods reported in Table 1 to Table 3. We find that "Prior 3" gives larger marginal likelihoods than the other priors and that "Prior 3" together with $d = 3$ has the highest marginal likelihoods among all models considered.

3 Label Switching and Post-processing MCMC

Finally, I would like to comment on handling label switching through post-processing the MCMC output. In (Frühwirth-Schnatter 2001) I suggested to use a point process representation of the MCMC draws, by producing scatter plots of pairs of component specific parameters. Since the MCMC draws scatter around the true point process representation of the mixture (Stephens 2000), a visual inspection of these plots allows to study the differences in the component specific parameters and to formulate an identifiability constraint. Although this works quite well in lower dimensions, I found it difficult or even impossible to extend this method to higher dimensional problems.

Table 5: Data simulated as in Case I with $\mu_1 = -2$, $\mu_2 = 0$, $\mu_3 = 2$, and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$; relabelling the MCMC output using the classification obtained from unsupervised k-means clustering; illustration for $m = 100$ (top), $m = 200$ (middle) and $m = 300$ (bottom); the estimators $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ are the averages over all original and all permuted MCMC draws

	MCMC output			Classification			Permuted MCMC output		
	$i=1$	$i=2$	$i=3$	$i=1$	$i=2$	$i=3$	State 1	State 2	State 3
$\mu_i^{[m]}$	-0.13	2.25	-2.04	3	1	2	2.25	-2.04	-0.13
$(\sigma_i^2)^{[m]}$	1.09	0.84	0.76				0.84	0.76	1.09
$\mu_i^{[m]}$	-1.94	0.15	2.31	2	3	1	2.31	-1.94	0.15
$(\sigma_i^2)^{[m]}$	0.83	1.01	0.59				0.59	0.83	1.01
$\mu_i^{[m]}$	-1.96	2.13	0.03	2	1	3	2.13	-1.96	0.03
$(\sigma_i^2)^{[m]}$	0.78	0.85	1.18				0.85	0.78	1.18
$\hat{\mu}_i$	-0.13	-0.08	-0.08				2	-2.21	-0.09
$\hat{\sigma}_i^2$	1.01	1.01	1.01				0.99	0.99	1.04

Influenced by a paper by Celeux (1998), I prefer now to use standard k-means clustering in the point process representation of the MCMC draws as a routine way to identify finite mixture and hidden Markov models. As opposed to Celeux (1998), clustering is performed in a post-processing manner. This method which is also included in the `bayesf` package is described in detail in (Frühwirth-Schnatter 2006, p. 96f) for finite mixture models, but applies to hidden Markov models as well. For hidden Markov models this method not only allows to identify the state specific parameters, but also to estimate the hidden Markov chain. The method is particularly useful in higher dimensions, for instance, for Markov mixtures of multivariate normal distributions, see (Hahn et al. 2007) for a recent application in finance.

The method is based on the idea that MCMC draws coming from the same state will cluster around the same point in the point process representation. Even if label switching occurred between two draws, the classification sequence resulting from k-means clustering indicates how to rearrange the state specific parameters. In cases where the simulation clusters are well-separated all classification sequences are a permutation of the labels $\{1, \ldots, d\}$ and show how to relabel the MCMC draws in order to obtain draws from an identified model.

For illustration, I consider simulated data of size $n = 1,000$ as in Case I of the paper and fit a hidden Markov model with state specific variances, $Y_k|X_k = i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ for various values of d. Bayesian inference is carried out by combining "Prior 3", defined above, on $\sigma_1^2, \ldots, \sigma_d^2$ and A with the following prior on μ_i: $\mu_i \sim \mathcal{N}\left(m, R^2\right)$, independently for $i = 1, \ldots, d$, where $m = (\min y_k + \max y_k)/2$ and R is the range of the data. Gibbs sampling was run for $M = 10,000$ sweeps after a burn-in of 1,000 draws and each step was ended by a random permutation as in Frühwirth-Schnatter

(2001). Unsupervised clustering is applied to a sample of size Md containing the draws $(\mu_i, \sigma_i^2)^{[m]}, i = 1, \ldots, d, m = 1, \ldots, M$.

First, the method is applied to data where $\sigma_i^2 \equiv 1$ which corresponds to the middle panel of Figure 1 of the paper. Let us start with fitting a three-state hidden Markov model with state specific variances. Table 5 shows MCMC draws of $(\mu_1, \sigma_1^2)^{[m]}$, $(\mu_2, \sigma_2^2)^{[m]}$, and $(\mu_3, \sigma_3^2)^{[m]}$ for $m = 100$, 200, and 300. Evidently label switching took place between these MCMC draws. Nevertheless, each classification sequence suggests a permutation of the labels which evidently leads to identified MCMC draws. For these data, also all remaining classification sequences were permutations of $\{1, 2, 3\}$ and allowed to identify a unique labelling for all MCMC draws including the transition matrices (not reported in the table) and the states of the hidden Markov chain. Estimators of the state specific parameters are obtained as averages of the identified MCMC draws, see e.g. $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ in the last two lines of Table 5.

For overfitting models the classification sequences are not necessarily permutations of the labels $\{1, \ldots, d\}$. When fitting a four-state hidden Markov model to the same data as above, only 175 among the 10,000 classification sequences are permutations of $\{1, \ldots, 4\}$, meaning that it is not possible to find a unique labelling of the MCMC draws for this model and these data. This is not surprising, because a four-state model is overfitting and the (log) marginal likelihoods of $d = 3$ versus $d = 4$ are -1869.1 and -1875.7, respectively, clearly favoring the three-state model.

Not only for this example, but also for many other simulated as well as real data sets, I found that a high fraction of classification sequences that are not permutations of the labels is a reliable hint that the model is overfitting.

To give a further example consider data where $\sigma_i^2 \equiv (1.5)^2$ which corresponds to the right panel of Figure 1 of the paper. There is a strong overlap of the components due to the high variance in each component which makes it difficult to identify the state specific means μ_1, μ_2 and μ_3. When we compare a two-state and three-state model for a data set of 1000 observations using (log) marginal likelihoods we obtain -2054.35 versus -2056.33, favoring the smaller model. The difficulty to identify the state specific means μ_1, μ_2 and μ_3 for this data set is also reflected by unsupervised clustering of the MCMC output. For $d = 2$ all classification sequences are permutations of the labels. However, for $d = 3$ this is the case only for 34 among the 10,000 classification sequences, indicating that it is not possible to find a unique labelling of the MCMC draws for a three-state model for this data set.

For a much larger data set of 10,000 observations the (log) marginal likelihoods of $d = 2$ versus $d = 3$ read -20666.7 and -20609.4, respectively, this time clearly favoring the three-state model. It is possible to identify the state specific means μ_1, μ_2 and μ_3, as indicated also by unsupervised clustering of the MCMC output, where all among the 10,000 classification sequences are permutations of $\{1, 2, 3\}$. The estimators of the state specific means obtained as averages of the identified MCMC draws are $\hat{\mu}_1 = -1.96$, $\hat{\mu}_2 = -0.003$, and $\hat{\mu}_3 = 1.98$.

References

Celeux, G. (1998). Bayesian Inference for Mixture: The Label Switching Problem. In Green, P. J. and Rayne, R. (eds.), *COMPSTAT 98*, 227–232. Heidelberg: Physica. 695

Congdon, P. (2006). Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Computational Statistics and Data Analysis*, 50: 346–357. 690

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96: 194–209. 691, 694, 695

— (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. New York/Berlin/Heidelberg: Springer. 689, 690, 691, 693, 694, 695

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Ser. B*, 56: 501–514. 690

Hahn, M., Frühwirth-Schnatter, S., and Sass, J. (2007). Markov chain Monte Carlo methods for parameter estimation in multidimensional continuous time Markov switching models. RICAM-Report 2007-09, Johannes Kepler University, Linz. 695

Robert, C. P. and Marin, J.-M. (2008). On some difficulties with a posterior probability approximation technique. *Bayesian Analysis*, 3: 427–442. 690

Scott, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97: 337–351. 690

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components – An alternative to reversible jump methods. *The Annals of Statistics*, 28: 40–74. 694