Bayesian Analysis (2008) **3**, Number 4, pp. 707–716

Rejoinder

Tobias Rydén[1]

First of all I want to thank Sylvia Frühwirth-Schnatter, Sergey Kirshner and Padhraic Smyth for their many interesting, thoughtful and constructive comments. In this rejoinder I will try to respond to most, although not all, of them.

1 Convergence of EM and further comparisons between EM and the Gibbs sampler

Smyth and Kirshner point out that for $\sigma = 1.5$ in Figure 2, the estimate of μ_1 (in my notation), for which the true value is -2, appears to diverge after about 30 iterations of EM. It is absolutely correct that the estimate drifts away from -2, but it does converge to a limit. This is shown in Figure A (left plot), which is similar to Figure 2 but extends over 200 iterations of EM. Obviously, even if convergence occurs for EM, it can happen that convergence is not monotone, not even after a moderately large number of iterations. The right plot of Figure A corresponds to the right plot of Figure 2. The irregularities in the curves for $\sigma = 1$ are caused by sign changes in $\mu_i^{(m+1)} - \mu_i^{(m)}$. Interestingly enough, the rate of convergence changes slightly after these sign changes, becoming a little bit slower. From a practical point of view this is of little importance, at least in the present example, as the irregularity appears when differences are of the order 10^{-7}, but it is still an observation that forces us to think carefully about the practical meaning of the word 'asymptotics'.

Another question raised by Smyth and Kirshner is that of convergence of the log-likelihood along the sequence of estimates produced by EM, and whether the limiting log-likelihood exceeds that at the true parameter. Figure B shows that for all three values of σ^2, this is indeed the case. It also shows that the log-likelihood increases much more dramatically in the first few iterations of EM than later, an observation that has been done repeatedly for the EM algorithm. We also see that the change from large to small improvements of the log-likelihood becomes more pronounced (sharp), the smaller σ^2 is. Comparing Figure A to Figure B, we find that features of the former one such as the sign changes of the differences $\mu_i^{(m+1)} - \mu_i^{(m)}$ have no counterpart in the latter figure. In fact, the second order differences of the log-likelihood sequences are all negative, for all values of σ^2. Thus all curves in Figure B are concave. From a practical point of view this seems to imply that if we are only interested in the maximal value of the log-likelihood, and not in the parameters where this maximum occurs (perhaps when doing model selection using penalised likelihoods), we can terminate EM earlier without worrying that the log-likelihood would increase notably had we continued EM for some more iterations. If EM is initialised far from the point to which is converges,

[1]Centre for Mathematical Sciences, Lund University, Lund, Sweden, mailto:tobias.ryden@matstat.lu.se

 DOI:10.1214/08-BA326REJ

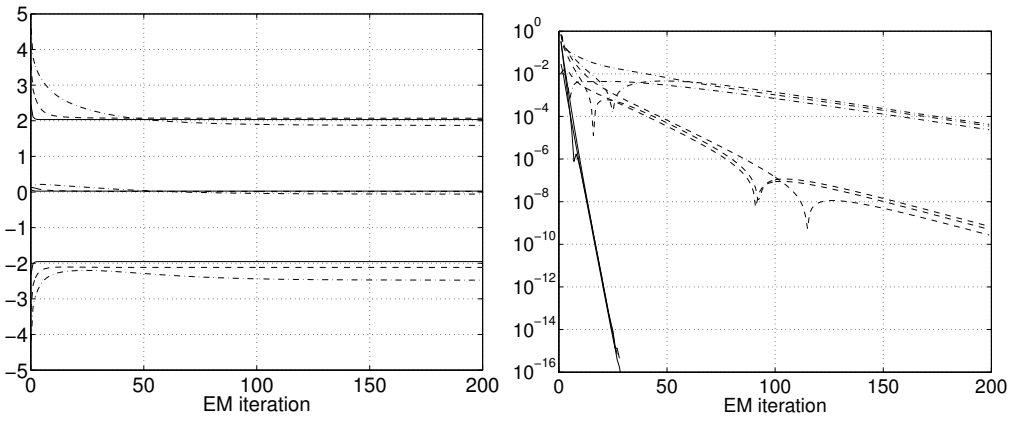

Figure A: *Left plot: trajectories of* $\mu_i^{(m)}$ *for EM iterations* $m = 0, 1, \ldots, 200$*, component indices* $i = 1, 2, 3$ *(bottom to top curves) and data simulated with* $\sigma = 0.5$ *(solid lines),* $\sigma = 1$ *(dashed lines) and* $\sigma = 1.5$ *(dash-dotted lines). Right plot: absolute differences* $|\mu_i^{(m+1)} - \mu_i^{(m)}|$*; same components and line symbols. Model and data were as in Case I.*

the log-likelihood curve could however have inflection points and hence not be concave. Combating such problems can be done by initialising EM from multiple starting points.

A further remark concerns the (empirical) posterior densities of the μ_i produced by the Gibbs sampler. These are all plotted in Figure C. For $\sigma = 1.5$ there is a clear overlap between the densities for different μ_i, which is in turn related to the label-switching seen in Figure 3; in addition the densities are not (as good as) symmetric, as for smaller σ, and at least for μ_1 not unimodal. The EM estimates (MLEs) are close to the posterior modes, with the exception for μ_3 when $\sigma = 1.5$ where there is a somewhat larger discrepancy. We also see that the MLEs are close to the posterior means, again except for μ_1 when $\sigma = 1.5$, where the label-switching causes a heavier tail to the right that pulls the posterior mean in the same direction.

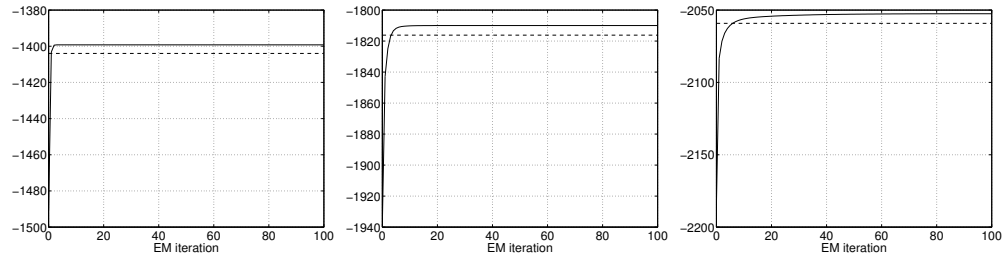

Figure B: *Log-likelihoods for parameters produced by the EM algorithm (solid lines) for the model and data of Case I with* $\sigma = 0.5$ *(left panel),* $\sigma = 1$ *(middle panel) and* $\sigma = 1.5$ *(right panel); dashed lines mark the log-likelihood at the true parameters.*

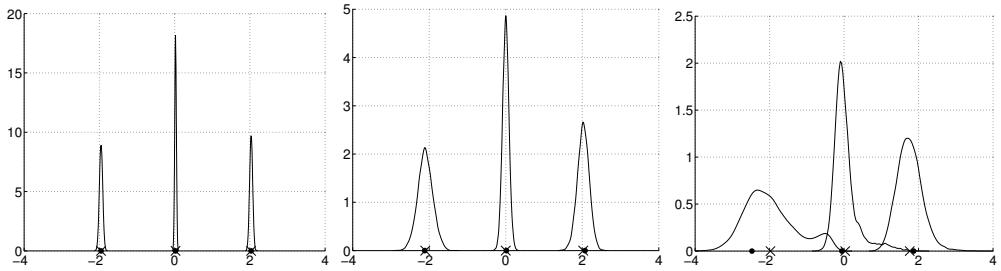

Figure C: *Empirical posterior densities of the means μ_i, sorted according to the constraint $\mu_1 < \mu_2 < \mu_3$, for the model and data of Case I with $\sigma = 0.5$ (left panel), $\sigma = 1$ (middle panel) and $\sigma = 1.5$ (right panel). The density estimates were computed from 10,000 sweeps of the Gibbs sampler, using the Matlab function* `ksdensity` *with default options. Crosses and bullets on the x-axes mark empirical posterior means and estimates after 200 iterations of EM, respectively.*

Yet another issue raised by Smyth and Kirshner is that transition probabilities are often more difficult to estimate than are the means. It is certainly a folklore among HMM users that parameters related to the dynamics to the Markov chain are more difficult to estimate than parameters governing the link between the x's and the y's, the intuition being that the former type of parameters relate to the hidden variables only, whereas the latter type of parameters also relate to the observations. In fact, such a claim can also be put on theoretical footing (Mitrophanov et al. 2005, Example 5). There is no unique or obviously best way to quantify this difference in difficulty, but here is one way. Taking one observation from $\mathrm{N}(\mu, \sigma^2)$ and estimating μ with this single number, provides an estimator with variance σ^2. Denote by ν_{μ_i} the sum of sample covariances in the range $-30, \ldots, 0, \ldots, 30$ of the $\mu_i^{[t]}$ over the 10,000 sweeps of the Gibbs sampler illustrated in Figure 3. Then ν_{μ_i} is a measure of uncertainty of the estimate of μ_i, and the ratio σ^2/ν_{μ_i} corresponds to an effective number of (i.i.d.) observations from $\mathrm{N}(\mu_i, \sigma^2)$ (cf. Section 2.1 of the main paper). For $\sigma = 0.5$, this ratio was 91, 399 and 126 respectively (sum 616) for μ_1, μ_2 and μ_3.

Reasoning in a similar way for the transition probabilities, by observing one jump from state i, the variance of an estimate of a_{ij} is $a_{ij}(1 - a_{ij})$; this is the variance of a Bernoulli random variable indicating whether the jump was to state j or not (cf. Basawa and Prakasa Rao 1980, Eq. (7) of Chapter 4, although that expression also accounts for the fraction π_i spent in the respective states). Computing the ratios $a_{ij}(1 - a_{ij})/\nu_{a_{ij}}$, with notation analogous to that above, yields the numbers 137, 381 and 186 (sum 704) respectively for a_{11}, a_{22} and a_{33}. This analysis does hence not support that for this particular model and data, estimating the a_{ii} is more difficult than is estimating the μ_i. An analogous analysis for $\sigma = 1$ yielded the same conclusion.

Related to the uncertainty of the parameter estimates is the construction of confidence or credibility intervals, and confidence or credibility multi-dimensional regions. Smyth and Kirshner pose the question whether the two approaches should produce sim-

ilar intervals and regions, or whether they capture different notions of uncertainty. I would say that confidence and credibility intervals or regions indeed answer essentially one single question. We are faced with a model, comprising certain parameters, and also a set of data assumed to (with reasonable accuracy) having been produced by this model. The question we want to address is: what were the parameter values of the model that produced this data, and can we somehow compute intervals or regions that in a suitable sense are likely to cover these parameter values? In a frequentist setting we have no additional information to guide us, but in a Bayesian setting we have a prior distribution on the parameters that adds some further information. Yet, as the size of the data becomes large, the posterior will be dominated by the likelihood part and hence the influence of the prior will be small. Therefore, when the size of the data is large, in comparison to the size of the model, the posterior and likelihood surfaces will look alike (see e.g. Cox and Hinkley 1974, Section 10.6). However, in a real problem data is not of infinite size, and in general confidence intervals need to be approximated using say a normality approximation of the MLE, or bootstrap (which is only approximate whatever the number of replicates). Therefore, for any real set of data, confidence and credibility intervals will be different, even though they address the same question.

2 Searching for modes and algorithmic considerations

Let us return to the problem of computing a point estimate, e.g. the MLE or the maximum aposteriori (MAP) estimate. Smyth and Kirshner propose to use MCMC to search for the mode, and consider whether MCMC as a mode-searching algorithm could have advantages to EM—in particular since EM may converge to local maxima. Using MCMC for this purpose is, I would say, a very reasonable idea, maybe most naturally for searching for the posterior mode (i.e. the MAP estimate), but also for searching for the mode of the log-likelihood (i.e. the MLE). Whether this approach has distinct advantages compared to EM is more difficult to say. Obviously EM can get trapped at local maxima and saddle points, which is the reason for often initialising EM at multiple starting points. If the posterior surface is multi-modal, which it commonly is for HMMs and which can be expected to happen when EM has trouble finding the global mode, an MCMC algorithm may however also spend a considerable fraction of the iterations away from the global mode; hence I find it difficult to make a general statement regarding the pros of cons of using MCMC for mode-searching.

Mode-searching, in the sense of finding the MLE, is also a key ingredient of the bootstrap. Smyth and Kirshner note that if EM stops at a local maxima, this may inflate confidence intervals in an undesired way. To combat this one may again initialise EM at multiple starting points also for the bootstrap replicates. When bootstrapped MLEs (or other estimates) are used to compute confidence intervals, it is important to eliminate label-switching, i.e. to align the components of the different estimates in a coherent way. For one-dimensional parameters such as individual one-dimensional means, this can be done by sorting, but, as pointed out by Smyth and Kirshner, such an alignment can be more difficult to achieve for multi-dimensional parameters. Presumably one could re-use techniques proposed for alignment of multi-dimensional parameters from

MCMC output. One such technique—clustering—is detailed by Frühwirth-Schnatter in her discussion. Another one, based on loss-function minimisation, is described by Stephens (2000); see also the survey by Jasra et al. (2005).

Smyth and Kirshner also bring up a criticism against Monte Carlo EM: why spend computational efforts on sampling latent variables given certain parameter values, when these parameter values are not optimal (in a likelihood sense) and will be updated in the next M-step anyway? One possible answer to this question is that if the number of Monte Carlo replications is increased in each iteration of MCEM in an appropriate way, the sequence of parameter estimates will in fact have convergence properties analogous to those of the standard EM algorithm (Fort and Moulines 2003). Thus, if one prefers the ML approach to inference and wants to mimic the behaviour of standard EM, but using MCEM, that is possible. Of course one could make a strong argument that there would be more sensible ways to use available computing resources, as Smyth and Kirshner point out, for instance by turning to MCMC. A different response to the original question is that one may also run just a single Monte Carlo replication in each E-step. The result is then a counterpart to an MCMC algorithm, but with resampling of parameters replaced by the M-step. Such algorithms are sometimes called stochastic EM (SEM) algorithms. The output of such an algorithm will never be a convergent sequence, but a sequence that can be analysed in a fashion similar the analysis of MCMC output. Some results on the characteristics of SEM algorithms are available (see e.g. Nielsen 2000), but it is fair to say that they are not yet fully understood.

3 Estimating the number of states

The careful analysis carried out by Frühwirth-Schnatter is illuminating, and it is also striking how much the prior actually influences the posterior distribution of the number of states. It is clear that determining the 'best' number of states is a problem that will continue to request attention for time to come, and it is far from certain that there will ever be a definite conclusion on this complex problem.

Let me first comment on the computing time aspect 57 CPU minutes for marginal likelihoods vs. 19 hours for RJMCMC. Although these numbers to quite some extent depend on computer and implementation details, the marginal likelihood approach requires less computations of the log-likelihood $\log p(\mathbf{y}_{1:n}|\theta, d)$. In each sweep of the RJMCMC algorithm employed in the main paper, this log-likelihood was evaluated four times (twice for updating the ω_{ij} and twice in the split/combine move). These evaluations account for the main part of the computation time, and I believe that the smaller number of such evaluations required for computing marginal likelihoods is the main reason for its lower computational cost; this is an advantage of the marginal likelihood approach. On the other hand, in the RJMCMC runs the sampler did not get stuck at states with some σ_i being (very close to) zero, as reported for the Gibbs sampler by Frühwirth-Schnatter. This could be because the RJMCMC sampler is able to escape such states by moving to a model of different dimensionality. Let me also remark that in this model, with conditional distributions $\mathrm{N}(0, \sigma_i^2)$ for the data, the likelihood

Table A: *Maximal log-likelihoods and BIC scores for the Normal variance mixture model and S&P 500 data from Case II. The maximal log-likelihoods were computed by starting the EM algorithm from 50 randomly chosen initial points, and then terminating each EM sequence when the improvement in log-likelihood fell below* 10^{-3}. *The dimensionality of a model with* d *states is* $d(d-1)+d = d^2$ *and the number of observations was* n =1,700, *hence the penalty for a* d*-state model is* $(1/2)d^2 \log n$.

	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$
$\max \log p_\theta(\mathbf{y}_{1:n})$	6115.2	6213.1	6223.2	6230.6	6235.1
BIC	6111.5	6198.3	6189.7	6171.1	6142.1

function is not unbounded unless one observation is exactly zero (which is not the case for the S&P 500 data). Had a common mean been included, i.e. stipulating conditional distributions $\mathrm{N}(\mu, \sigma_i^2)$, the likelihood had however been unbounded. To complement the results obtained by MCMC methods and likelihood ratio tests, the BIC scores for this model are shown in Table A. We find that with BIC one would clearly prefer a model with two states, with a three-state model in second place.

From now on I will use the model and simulated data of Case I to try to shed some more light on the influence of the sample size, as well as the influence on the Dirichlet prior for the rows of the transition probability matrix, on the posterior distribution of the number of states.

For this model I implemented an RJMCMC sampler very similar to the one described in Section 3 of the main paper; in part (i) of a sweep the μ_i and σ^2 were updated with Gibbs steps, part (ii) was the same, part (iii) was void and in part (iv) there was a split/combine move that either attempted to split a mean μ_{i_0} into $\mu_{i_1} = \mu_{i_0} - \xi_\mu \sigma$ and $\mu_{i_2} = \mu_{i_0} - \xi_\mu \sigma$ with $\xi_\nu \sim \mathrm{N}(0, \tau'_\mu)$ and $\tau'_\mu = 0.9$. Further details of this move will not be given here, but it is similar to the split/combine move in Cappé et al. (2005, Example 13.2.2). The priors for the μ_i, σ^2 and A were as in Case I, and the prior for d was uniform on $\{1, 2, \ldots, 8\}$. The estimated posterior for d is shown in Table B. In addition to the same data as in Case I, results are also shown for data being the subsequence consisting of the first 100 observations; this is to address the question by Smyth and Kirshner about performance for smaller data sets. We see that for n =1,000, the model size $d = 3$ comes out as the most probable, but with $d = 4$ and even larger d not ruled out; for $\sigma = 1.5$ also $d = 2$ has a notable posterior probability. We also see that the decision on $d = 3$ as the best model size becomes less clear-cut as σ increases, since more states get non-negligible posterior probabilities.

I did not compute estimates for the Dirichlet prior advocated by Frühwirth-Schnatter, i.e. parameters 4 and $1/(d-1)$ for diagonal and off-diagonal elements respectively of A, but the marginal likelihoods for $d = 3$ and $d = 4$ that she reports for $\sigma = 1$ supports $d = 3$ more strongly than the posterior probabilities reported here. In order to understand better why this is so, we can study Figure D, which shows kernel densities for the

Table B: *Estimated posterior probabilities for the model order d, for the model with conditional densities $\mathrm{N}(\mu_i, \sigma^2)$, priors for μ_i, σ^2 and A as in Case I, and a uniform prior on $\{1, 2, \ldots, 8\}$ for d. Data was as in Case I (n =1,000), or a subset thereof ($n = 100$). The estimates were obtained with a reversible jump MCMC algorithm similar to that in Section 3 of the main paper.*

	$\sigma = 0.5$ n =1,000	$\sigma = 0.5$ n =100	$\sigma = 1$ n =1,000	$\sigma = 1$ n =100	$\sigma = 1.5$ n =1,000	$\sigma = 1.5$ n =100
$d = 1$	0.000	0.000	0.000	0.003	0.000	0.029
$d = 2$	0.000	0.000	0.000	0.123	0.230	0.341
$d = 3$	0.807	0.258	0.487	0.344	0.552	0.291
$d = 4$	0.172	0.364	0.386	0.282	0.176	0.171
$d = 5$	0.020	0.229	0.102	0.142	0.035	0.091
$d = 6$	0.002	0.100	0.019	0.069	0.007	0.043
$d = 7$	0.000	0.038	0.005	0.028	0.000	0.020
$d = 8$	0.000	0.011	0.001	0.011	0.000	0.013

diagonal entries a_{ii}, sorted in ascending order, in all sweeps of the RJMCMC sampler for which the model order was $d = 4$. These plots reveal that the smallest diagonal entry tends to be considerably smaller than the other ones. This effect is particularly pronounced for $\sigma = 0.5$ and $\sigma = 1$. Hence we see that for $d = 4$, a model size which cannot be ruled out according to Table B, most models have one diagonal entry a_{ii} that does not dominate that row of A. The Dirichlet prior for A used by Frühwirth-Schnatter punishes such models, hence giving a smaller posterior probability to $d = 4$. A closer examination of the models of size $d = 4$ showed that, in particular for $\sigma = 0.5$, in such sweeps there are often two means $\mu_i^{[t]}$ that are close together (and close to one of the three true means), thus effectively creating a 'macro state' in the model consisting of two Markov states with essentially a single common mean.

We now discuss the results for the smaller sample size $n = 100$. Table B shows that the posterior probabilities for d, obtained with the RJMCMC sampler, are such that for $\sigma = 0.5$ the model size $d = 4$ is most probable, with $d = 3$ and $d = 5$ close to being tied in second place; for $\sigma = 1$ the model size $d = 3$ is most probable, with $d = 4$ in second place; and for $\sigma = 1.5$ the model size $d = 2$ is most probable, with $d = 3$ in second place. Thus there is a trend to favour smaller models as σ increases. In all cases the posterior distribution is quite spread out however, i.e. giving considerable probability to multiple values of d, so that the decision about d is never done with great certainty. As a comparison we computed Monte Carlo p-values of the generalised likelihood ratio test, using bootstrap as in Section 3.2 of the main paper; results are found in Table C. This test firmly rejects $d = 2$ vs. $d = 3$ for $\sigma = 0.5$ and $\sigma = 1$. For $\sigma = 1.5$ the p-value 0.144 is above commonly used thresholds for rejection however. For all σ the test of $d = 3$ vs. $d = 4$ was not rejected. We thus see that the GLRT would lead us (at the 5% level) to the correct d for $\sigma = 0.5$ and $\sigma = 1$, and to $d = 2$ for $\sigma = 1.5$, although with a

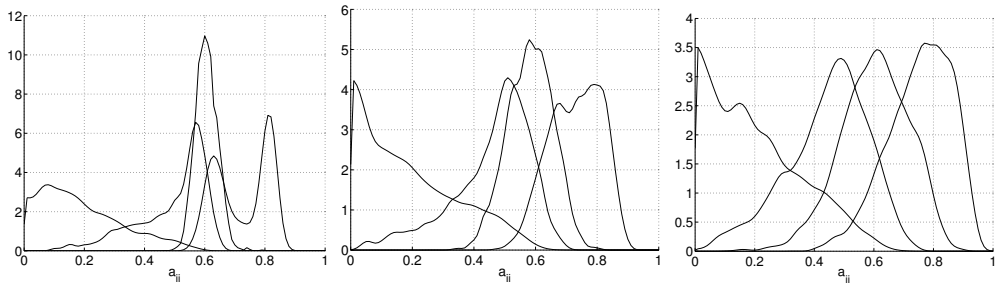

Figure D: *Kernel density estimates computed from the diagonal entries $a_{ii}^{[t]}$, sorted in ascending order for each sweep t, of the transition probability matrix in all sweeps of the reversible jump MCMC sampler for which the model size was $d = 4$. Model and data were as in Case I; $\sigma = 0.5$ (left panel), $\sigma = 1$ (middle panel) and $\sigma = 1.5$ (right panel). The estimates were computed using the Matlab function* `ksdensity` *with default settings, and then mirrored around $x = 0$ to ensure support on the positive real line.*

Table C: *Monte Carlo p-values for testing model order d vs. $d + 1$, for the model with conditional densities $\mathrm{N}(\mu_i, \sigma^2)$. Data was the first 100 samples of the data in Case I. The p-values were obtained by parametric bootstrap, as in Section 3.2 of the main paper, with 200 replications. For each replication as well as for the original data, maximal likelihoods were computed by running EM initialised from 50 randomly chosen starting points, terminating each sequence when the increase in log-likelihood fell below 10^{-3}.*

	$\sigma = 0.5$ n =100	$\sigma = 1$ n =100	$\sigma = 1.5$ n =100
$d = 2$ vs. $d = 3$	0.005	0.040	0.144
$d = 3$ vs. $d = 4$	0.602	0.736	0.890

hint that $d = 3$ might be a better choice. These test results are certainly more distinct than the conclusions obtainable from the quite diffuse posterior on d in all cases. This is a definite advantage of the Monte Carlo GLRT. Its drawback are computation times that are much longer than for the RJMCMC sampler, as in Section 3 of the main paper. Of course, with a prior such as the one proposed by Frühwirth-Schnatter, the posterior for d would look different.

4 Costs of implementation

Smyth and Kirshner suggest that the cost of implementation, i.e. coding, debugging, etc., is often significantly larger than the computational costs of actually running the code to make the desired inference. I couldn't agree more on this view! For precisely this reason, software packages for making inference in hidden Markov models are

highly valuable. Today there are some such packages available, but in the future there will hopefully be more and more versatile ones. The `bayessf` package described by Frühwirth-Schnatter is a valuable addition in this direction.

References

Basawa, I. V. and Prakasa Rao, B. L. S. (1980). *Statistical Inference for Stochastic Processes*. London: Academic Press. 709

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. New York: Springer. 712

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman & Hall. 710

Fort, G. and Moulines, E. (2003). "Convergence of the Monte Carlo expectation maximization for curved exponential families." *Annals of Statistics*, 31: 1220–1259. 711

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling." *Statistical Science*, 20: 50–67. 711

Mitrophanov, A. Y., Lomsadze, A., and Borodovsky, M. (2005). "Sensitivity of hidden Markov models." *Journal of Applied Probability*, 42: 632–642. 709

Nielsen, S. F. (2000). "The stochastic EM algorithm: estimation and asymptotic results." *Bernoulli*, 6: 457–489. 711

Stephens, M. (2000). "Dealing with label switching in mixture models." *Journal of the Rouyal Statistical Society - Series B*, 62: 795–809. 711