Bayesian Analysis (2008) **3**, Number 4, pp. 699–706

Comment on article by Rydén

Padhraic Smyth* and Sergey Kirshner†

1 Introductory Remarks

This paper is an informative and useful addition to the literature on estimation of hidden Markov models. The author has done an admirable job of covering many of the key ideas that are relevant to statistical inference of HMMs. Empirical analysis of the computational aspects of HMMs can be quite difficult to perform. As a simple example, elementary diagnostic tools such as scatter-plots are not so useful in analyzing HMMs, unlike (say) the case of low-dimensional finite mixture models.

The detailed comparison of EM and MCMC in this paper is particularly useful. Most prior work on the topic of HMMs tends to rely on one technique or the other, with little practical advice provided on how to compare the two approaches.

2 Background Comments on HMMs in Computer Science and Engineering

As mentioned in the introduction of the paper, HMMs are now a standard tool in applied statistics. We completely agree and would like to reinforce this point from the viewpoint of computer science and engineering, where HMMs and their various cousins are widely used as a relatively simple (but very useful) framework for modeling processes with sequential dependence. Most of the early applications of HMMs were in speech modeling, but in the past 10 or 15 years their use has extended to a much broader set of application areas such as robotics (Thrun et al. 1998), computer vision (Wilson and Bobick 1999; Oliver et al. 2000), information extraction from text (McCallum et al. 2000), language modeling (Griffiths et al. 2005), and sensor data analysis (Ihler et al. 2007).

Indeed, as pointed out in the Introduction of the paper, there are numerous variations and extensions of the "basic HMM." Examples from machine learning include factorial HMMs (Ghahramani and Jordan 1997), various forms of dynamic Bayesian networks (Murphy 2002), conditional random fields (Lafferty and Pereira 2001), and so forth. A useful unification of these many different models is to view them within the more general framework of graphical models (Smyth et al. 1997; Bilmes 2004). For example, in this context, the Baum-Welch algorithm can be seen as a special case of the more general tree-based inference algorithms for graphical models proposed by Lauritzen and Spiegelhalter (1988) and by Pearl (1988).

From an inference viewpoint, the EM algorithm has been the most widely used

*Department of Computer Science, University of California, Irvine, CA, mailto:smyth@ics.uci.edu
†Department of Statistics, Purdue University, West Lafayette, IN, mailto:skirshne@stat.purdue.edu

 DOI:10.1214/08-BA326B

methodology since the early work by Baum et al. (1970) and the influential review paper by Rabiner (1989). However, as in statistics, in recent years there has a been a growing and substantial body of work on using Bayesian inference techniques in the context of HMMs. Playing an important role here is the growing awareness of Bayesian thinking in computer science, with early seeds planted by the likes of Geman and Geman (1984), followed by popularization of Bayesian methodologies due to the work of Mackay (1992), Buntine (1994), Heckerman et al. (1995), Neal (1996), Jordan (1998), and many others. As computational power continues to increase, and as Bayesian methodologies become more widely known in computer science and engineering, we can expect this trend to continue and work on HMMs will increasingly have a Bayesian flavor. This present paper is likely to be a valuable reference for such work.

3 Comments on EM and Gibbs Sampler Convergence

The illustration of convergence of EM and Gibbs sampling, in Figures 2 and 3 respectively, is quite informative. For example, it is clear in both cases that the amount of overlap in the two Gaussian densities (or equivalently the amount of missing information in the model c) has a significant effect on the resulting inference, e.g., on the rate of convergence of EM (Figure 2, right) and on the posterior parameter variances in Gibbs sampling (Figure 3).

Figure 2 (left) plots the trajectories of the estimated state means as a function of iteration number using EM. We noticed that for $\sigma = 1.5$ the estimate of μ_3 appears to be diverging away from its true value $\mu_3 = -2$ after EM goes beyond about 30 iterations. There is also some indication that the estimates for the other two mean values may also be diverging, again for $\sigma = 1.5$. Presumably, after more iterations, EM does converge to a point in parameter space that is a maximum of the likelihood surface. It would be worthwhile to show that convergence does indeed occur (by showing more iterations of EM on the x-axis). It would also be interesting to know if the value of the log-likelihood at the maximum found by EM is higher than the value of the log-likelihood at the true parameter values (a likely situation of course for relatively small amounts of data). Finally, it would be interesting to show a plot of the posterior densities of the means as estimated by the Gibbs sampler, with the location of the EM point estimate also marked.

In our experience with HMMs, the transition probabilities are sometimes more difficult to estimate accurately (compared to the means) when using relatively small amounts of training data. Perhaps the author can provide some observations on this based on the experiments in this paper.

4 Credibility and Confidence Intervals

The paper makes a clear case in Section 2 that when interval estimates for parameters are required, MCMC methods can be as computationally cheap to implement as frequentist alternatives such as the parametric bootstrap—and MCMC inference can

provide additional benefits such as parameter averaging for prediction. But what if we only want point estimates of our HMM parameters to begin with? Are there advantages in using MCMC to seek the posterior mode in parameter space, as a stochastic search alternative to EM (which can get trapped at local maxima)?

Section 2.3 provides an interesting comparison of both non-Bayesian and Bayesian approaches to determining credibility intervals for HMM parameter estimates. The Gibbs sampler provides a credibility estimate directly from the estimated (sample-based) posterior density for the parameter of interest. The parametric bootstrap provides a confidence interval using K MLE parameter estimates computed from K bootstrap samples of the original data (here $K = 661$). It is not entirely clear that the credibility and confidence intervals from the Gibbs sampler and from the parametric bootstrap, respectively, should agree in principle. Gibbs provides us with an empirical estimate of our posterior uncertainty; the parametric bootstrap tells us about the uncertainty in our MLE point estimates from different bootstrap data sets of the same size. In practice the resulting intervals from both approaches will likely be similar (as they are in the example in Section 2.3), but fundamentally these intervals appear to capture two different notions of uncertainty. Can the author comment further on this difference?

At the end of Section 2.3 the author compares computation times for the Gibbs sampler and the parametric bootstrap, pointing out that "bootstrapping is more automatic," but also noting hat the computation time of the bootstrap method will depend on the stopping procedure for EM that is being used. It may be worth noting that, in addition to the identifiability issues discussed in Section 2.3, EM is also prone to finding local maxima (particularly for small data sets). Thus, an additional complicating factor in using the bootstrap in this context is the presence of local maxima during EM optimization—these local maxima could result in significant (and artificial) inflation of the bootstrap confidence interval estimates.

As final comment on the bootstrap method, we wondered how the identifiability constraints are implemented if the means are multi-dimensional, rather than one-dimensional as in the example in the paper?

5 Inferring the Appropriate Number of States

We found the comparison of the bootstrapped GLRT and reversible jump MCMC for model selection in Section 3 to be interesting and informative. As a general comment on selecting the number of states in an HMM, it is worth noting that in the case where the HMM is an approximation to a true data-generating process that is not itself an HMM (as is usually the case in practice), there is no real notion of the "correct" number of states. As we get more data, the "best" number of states is the number that leads to the best predictions on new data, and this number will increase as we get more training data. So "best" in practice may be a function of how much data we have, rather than some data-independent notion of truth. This is particularly important in situations where the inferred states are accorded specific interpretations (e.g., for data analysis in the sciences), versus the case where the model is being used as a black-box predictor as

is often the case in speech processing for example.

In Section 3.1, the economic data seems like it will by definition have high overlap between the components given that the means are the same for both components. Perhaps the author could comment on how this may influence model selection, in the light of Section 2 where we saw that significant overlap adds considerable variance into the posterior on parameters. Does the relatively large variance due to component overlap also translate into uncertainty or instability in model selection?

It would be interesting to know the results of applying the model selection methods from Section 3 to the simple simulated data from Section 2 (since we know ground truth for this data). If the author has run this analysis, perhaps he could provide a few sentences or comments on the results? Of particular interest would be what happens when the amount of training data is reduced—is there any indication that reversible jump MCMC or bootstrapped GLRT methods have any particular advantage over each other in the small training data regime?

As the author points out, penalized likelihood criteria for model selection (e.g., BIC) are problematic when applied to models such as HMMs. However, these criteria are widely used (particularly in the sciences) and relatively easy and fast to compute. In this context, for completeness, it would be worthwhile to report such scores for the data set used in Section 3. It would also be useful to see how non-parametric Bayesian methods compare when used for model selection, such as Dirichlet processes which place an implicit prior over the number of hidden states (e.g., Xing and Sohn 2007). We realize of course that this could entail significant additional work on the part of the author—but hopefully future comparative studies by the author or by others could include such approaches.

The author mentions several methods in Section 3 for determining the number of states in an HMM. We would like to point out yet another option: use cross-validation (or simply test set validation if one has a very large amount of data). For example, from an MLE viewpoint, one can compute the log-probability of unseen data for different numbers of states using EM for estimation, and then (for example) select the model with the highest out-of-sample log-probability. While acknowledging that the resulting cross-validation estimates can be quite noisy for small data sets, we have nonetheless found this approach to be simple to implement and it has the advantage that its predictive nature makes it easy for scientific collaborators to trust and understand (e.g., see an application to climate data in Robertson et al. (2004)).

6 "MC within EM" versus MCMC

As described in the paper, the model in Section 4 no longer allows an exact and tractable E-step for the EM. The solution used in the paper is to average over samples from the posterior distribution of the latent variables, instead of computing the exact expectation. This approach can be viewed as a hybrid between EM and Gibbs sampler MCMC, with the difference of choosing a mode of the posterior distribution over model param-

eters (MCEM) rather than sampling from that distribution (MCMC). Given that the computational cost of sampling values of the latent variable would often dominate, we should not expect MCEM to offer a computational advantage over MCMC. Without such an advantage, is there a reason to bother with MCEM at all? It may be worth considering an alternative to sampling within EM such as variational methods which can be faster than sampling (e.g., Ghahramani and Jordan 1997).

7 Computational Costs and Human Costs

The paper provides a number of useful practical estimates of computation time, e.g., comparing the time taken for Gibbs sampling and bootstrap methods in Sections 2.3 and 3.2. There is a current trend in computer systems towards making parallel computing readily available on the desktop, whether in the form of multiple processors in your desktop machine (already a reality) or easy access to distributed (e.g., grid) computing resources. While parallelization at best only provides a linear speed-up of P, where P is the number of processors, this trend towards parallelization is a boon for simulation-based inference. As examples of "trivial parallelizations," for MCMC one can run multiple independent chains on different processors, or for bootstrap methods one can allocate different bootstrapped data sets to different processors. In consequence, as computation times decrease, we can anticipate even more interest and activity in using the types of inference methods from this paper in applied problems.

The paper discusses briefly (e.g., in the Summary of Section 3, in Section 3.5) the point that in addition to computation time, another factor to consider in choosing an inference method is the complexity of the software code required to implement a particular method. We would go a little further and argue that the "human cost" involved correctly implementing and testing an inference procedure in software can actually be a much more significant cost in many applied projects compared to the computational cost of running the inference. In particular, for relatively complex models and relatively complex inference procedures (such as reversible jump MCMC—as pointed out in Section 3.5) the human cost of implementation can be significant, particularly for practitioners who do not have prior experience with working with complex inference methods. Debugging the software associated with complex MCMC samplers is nontrivial, e.g., differentiating between slow mixing and a bug in the code can be quite difficult. Thus, human cost is an important factor (in addition to computational cost) when deciding what inference methodology to choose in practical situations.

8 Concluding Comments

This paper is a welcome addition to the HMM literature for researchers interested in the relative advantages and disadvantages of EM and MCMC inference. The detailed and specific examples comparing EM and MCMC should be particularly valuable for example for students and for practitioners, who may not have much prior experience with these techniques. In collaborative work involving HMMs it is difficult to provide a

short answer to the question "is it better to use an EM algorithm or a Bayesian approach when modeling data with HMMs?" This paper will be a very useful starting point for answering such questions.

References

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). "A Maximization Technique Occurring in Statistical Analysis of Probabilistic Functions of Markov Chains." *Annals of Mathematic Statistics*, 41(1): 164–171. 700

Bilmes, J. A. (2004). "Graphical models and automatic speech recognition." In Johnson, M., Khudanpur, S., Ostendorf, M., and Rosenfeld, R. (eds.), *Mathematical Foundations of Speech and Language Processing*, volume 138 of *The IMA Volumes in Mathematics and its Applications*, 191–246. New York: Springer. 699

Buntine, W. (1994). "Operations for learning with graphical models." *Journal of Artificial Intelligence Research*, 2: 159–225. 700

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6): 721–741. 700

Ghahramani, Z. and Jordan, M. I. (1997). "Factorial hidden Markov models." *Machine Learning*, 29: 245–273. 699, 703

Griffiths, T. L., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. (2005). "Integrating Topics and Syntax." In Saul, L. K., Weiss, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems 17*, 537–544. Cambridge, MA: MIT Press. 699

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). "Learning Bayesian networks: The combination of knowledge and statistical data." *Machine Learning*, 20(3): 197–243. 700

Ihler, A., Hutchins, J., and Smyth, P. (2007). "Learning to detect events with Markov-modulated Poisson processes." *ACM Transactions on Knowledge Discovery from Data*, 1(3). 699

Jordan, M. I. (ed.) (1998). *Learning in Graphical Models*. Cambridge, MA: MIT Press. 700

Lafferty, J. and Pereira, F. (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." In *International Conference on Machine Learning (ICML 2001)*, 282–289. Morgan Kaufmann. 699

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). "Local computations with probabilities on graphical structures and their application to expert systems (with discussion)." *Journal of the Royal Statistical Society, Series B*, 50(2): 157–224. 699

Mackay, D. J. C. (1992). "A practical Bayesian framework for backpropagation networks." *Neural Computation*, 4: 448–472. 700

McCallum, A., Nigam, K., Rennie, J., and Seymore, K. (2000). "Automating the construction of internet portals with machine learning." *Information Retrieval*, 3: 127–163. 699

Murphy, K. P. (2002). "Dynamic Bayesian Networks: Representation, Inference and Learning." Ph.D. thesis, Department of Electrical Engineering and Computer Science, UC Berkeley. 699

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. New York: Springer. 700

Oliver, N., Rosario, B., and Pentland, A. (2000). "A Bayesian computer vision system for modeling human interactions." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22: 831–843. 699

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann. 699

Rabiner, L. R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." *Proceedings of the IEEE*, 77: 257–286. 700

Robertson, A. W., Kirshner, S., and Smyth, P. (2004). "Downscaling of daily rainfall occurrence over northeast Brazil using a hidden Markov model." *Journal of Climate*, 17(22): 4407–4424. 702

Smyth, P., Heckerman, D., and Jordan, M. (1997). "Probabilistic independence networks for hidden Markov probability models." *Neural Computation*, 9: 227–269. 699

Thrun, S., Burgard, W., and Fox, D. (1998). "A probabilistic approach to concurrent mapping and localization for mobile robots." *Machine Learning*, 31(1–3): 29–53. 699

Wilson, A. D. and Bobick, A. F. (1999). "Parametric hidden Markov models for gesture recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21: 884–900. 699

Xing, E. P. and Sohn, K.-A. (2007). "Hidden Markov Dirichlet process: Modeling genetic inference in open ancestral space." *Bayesian Analysis*, 2(3): 501–528. 702