Rejoinder for "Bayesian Nonparametric Latent Feature Models"

Z. Ghahramani, P. Sollich, and T. L. Griffiths

February 16, 2007

We thank Dr. Dunson for a stimulating discussion of our paper. In his discussion, Dunson makes several comments about our paper, and then proposes an alternative approach to sparse latent feature modelling. We first address his comments, and then turn to his suggested approach.

The first comment is that although the utility of sparse latent factor models has been illustrated by West and colleagues, it is not clear whether there are practical advantages to allowing the number of latent factors to be unbounded, as in our approach, as opposed to defining a model with a finite but unknown number of latent factors.

There are two advantages, we believe, one philosophical and one practical. The philosophical advantage is what motivates the use of nonparametric Bayesian methods in the first place: If we don't really believe that the data was actually generated from a finite number of latent factors, then we should not put much or any of our prior mass on such hypotheses. It is hard to think of many real-world generative processes for data in which one can be confident that there are some small number of latent factors. On the practical side, a finite model with an unknown number of latent factors may be preferable to an infinite model if there were significant computational advantages to assuming the finite model. However, inference in finite models of unknown dimension is in fact more computationally demanding, due to the variable dimensionality of the parameter space. Our experience comparing sampling from the infinite model and using Reversible Jump MCMC to sample from an analogous finite but variable-dimension model suggests that the sampler for the infinite model is both easier to implement and faster to mix (Wood et al, 2006).

Dunson also states that for West and colleagues "performance is best when the number of latent features represented in the sample is much less than the sample size". However, West's (2003) model is substantially different from ours; it is essentially a linear Gaussian factor analysis model with a sparse prior on the factor loading matrix, while our infinite latent feature models

can be used in many different contexts and allow the factors themselves to be sparse. We do not feel that the results that West reports on a particular application and choice of model specification can be generalized to Bayesian inference in all sparse models with latent features.

A second comment is that the assumption of feature exchangeability makes inference in the latent feature space awkward. This is a similar problem to the one suffered by Dirichlet process mixture (DPM) models where feature indices can change across samples in an MCMC run. We agree that questions such as "what does latent feature k represent" are meaningless in models with exhangeable features. We would never really be interested in such questions. However, there are plenty of meaningful inferences that can be derived from such a model, such as asking how many latent features two data points share. Rather than looking at averages of $\mathbf{Z}$ across MCMC runs, which makes no sense in an model with exchangeable features, one can look at averages of the $N \times N$ matrix $\mathbf{Z}\mathbf{Z}^T$, whose elements measure the number of latent features two data points share. Dunson's proposed solution, a prior that explicitly orders features by their frequency of occurence, is interesting but probably not enough to ensure that meaningful inferences can be made about $\mathbf{Z}$. For example, if two latent features have approximately the same frequency across the data, then any reasonably well-mixing sampler will frequently permute their labels, again muddling inferences about $\mathbf{Z}$ and the parameters associated with the two latent features.

A third comment made by Dunson is that one can define a more flexible model by having a non-parametric model for the features scores v_{ik}, rather than a parametric model. We entirely agree with this last point, and we did not intend to imply that v needs to come from a parametric model. A non-parametric model for v_{ik}, for example based on the Dirichlet process, is potentially very desirable in certain contexts. One possible disadvantage of such a model is that it requires additional bookkeeping and computation in an MCMC implementation. For certain parametric models for v_{ik}, one can analytically integrate out the $\mathbf{V}$ matrix, making the MCMC sampler over other variables mix faster.

We now turn to the proposed exponentiated gamma Dirichlet process (EGDP). This is an interesting model, well worth further study and elaboration.

Our first comment on this model is that the γ_h random variables defined in equation (1) of the discussion are rather unnecessary. Pushing through the transformation of variables, we can compute the distribution on π_h implied by assuming that γ_h follows a particular distribution. In the case of the exponentiated gamma model, this gives $\pi_h \sim \text{Beta}(1, \beta_h)$. This leads us to the question of why this way around and not, e.g., $\text{Beta}(\alpha_h, 1)$? The latter would be a more natural way to generalize our $\pi_h \sim \text{Beta}(\alpha/K, 1)$ to have non-

exchangeable latent features. In this proposal, the α_h would get smaller for $h \to \infty$, with the mean frequency for feature h being $\frac{\alpha_h}{\alpha_h+1}$. Writing both models in terms of their Beta distributions over feature frequencies highlights the similarities and differences between the two proposals. The choice Beta$(\alpha_h, 1)$ provides an alternative method for producing sparseness. Of course one could also look at Beta$(\alpha_h\beta, \beta)$, to generalize our two-parameter model.

Making the features inequivalent is attractive in some respects, but on the other hand may reduce flexibility. With exponentially decreasing β's, the higher index features will be so strongly suppressed that they will be hard to "activate" even with large amounts of data.

For the factor model in equation (5) of the discussion, we disagree that making the f's all positive is necessarily a good thing—one then models data that lie in a (suitably affinely transformed) octant of the space spanned by the columns of $\mathbf{L}$, rather than in the whole space. This is not merely a method for fixing a sign indeterminacy, but makes quite a different assumption about the data than in an ordinary factor analysis model. This model with positive factors is similar to a large body of work on non-negative matrix factorization models (e.g. Paatero and Tapper, 1994; Lee and Seung 1999).

To summarize, we thank Dr. Dunson for his interesting discussion and we hope that our work, his discussion, and this rejoinder will stimulate further work on sparse latent feature models.

REFERENCES

Lee, D. D. and Seung, H. S. (1999) Learning the parts of objects by non-negative matrix factorization, *Nature*, **401**(6755):788–791.

Paatero, P. and U. Tapper (1994) Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics* **5**:111–126.