Discussion of "Bayesian Nonparametric Latent Feature Models" by Zoubin Ghahramani

David B. Dunson

Biostatistics Branch, MD A3-03, National Institute of Environmental Health Sciences
Research Triangle Park, NC 27709, USA

1 Brief Comments

Ghahramani and colleagues have proposed an interesting class of infinite latent feature (ILF) models. The basic premise of ILF models is that there are infinitely many latent predictors represented in the population, with any particular subject having a finite selection. This is presented as an important advance over models that allow a finite number of latent variables. ILF models are most useful when all but a few of the features are very rare, so that one obtains a *sparse* representation. Otherwise, one cannot realistically hope to learn about the latent feature structure from the available data. The utility of sparse latent factor models has been compellingly illustrated in large p, small n problems by West (2003) and Carvalho et al. (2006). Given that performance is best when the number of latent features represented in the sample is much less than the sample size, it is not clear whether there are practical advantages to the ILF formulation over finite latent variable models that allow uncertainty in the dimension. For example, Lopes and West (2004) and Dunson (2006) allow the number of latent factors to be unknown using Bayesian methods.

That said, it is conceptually appealing to allow additional features to be represented in the data set as additional subjects are added, and it is also appealing to allow partial clustering of subjects. In particular, under an ILF model, subjects can have some features in common, leading to a degree of similarity based on the number of shared features and the values of these features.

Following the notation of Ghahramani et al., the $K \times 1$ latent feature vector for subject i is denoted $\mathbf{f}_i = (f_{i1}, \dots, f_{iK})'$, with $f_{ik} = z_{ik} v_{ik}$, where $z_{ik} = 1$ if subject i has feature k and $z_{ik} = 0$ otherwise, and v_{ik} is the value of the feature. There are then two important aspects of the specification for an infinite latent feature model: (1) the prior on the $N \times K$ binary matrix $\mathbf{Z} = \{z_{ik}\}$, with $K \to \infty$; and (2) the prior on the $N \times K$ matrix $\mathbf{V} = \{v_{ik}\}$.

The focus of Ghahramani et al. is on the prior for $\mathbf{Z}$, proposing an Indian Buffet Process (IBP) specification. The IBP follows in a straightforward but elegant manner from the following assumptions: (i) the elements of $\mathbf{Z}$ are independent and Bernoulli distributed given π_k, the probability of occurrence of the kth feature; and (ii) $\pi_k \sim \text{beta}(\alpha/K, 1)$. Because the features are treated as exchangeable in this specification, it is necessary to introduce a left ordering function, so that it is possible to base inference on a finite approximation focusing only on the more common features.

In this discussion, I briefly consider the more general problem of nonparametric modeling of both $\mathbf{Z}$ and $\mathbf{V}$, proposing an exponentiated gamma Dirichlet process (EGDP) prior. The exponentiated gamma (EG) is used as an alternative to the IBP, with some advantages, while the Dirichlet process (DP) (Ferguson, 1973; 1974) is used for nonparametric modeling of the feature scores among subjects possessing a feature.

2 Exponentiated Gamma Dirichlet Process

To provide motivation, I focus on an epidemiologic application in which an ILF model is clearly warranted. In the Agricultural Health Study (Kamel et al., 2005), interest focused on studying factors contributing to neurological symptom (headaches, dizziness, etc) occurrence in farm workers. Individual i is asked through a questionnaire to record the frequency of symptom occurrence for p different symptom types, resulting in the response vector,

$\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$. It is natural to suppose that the symptom frequencies, $\mathbf{y}_i$, provide measurements of latent features, $\mathbf{f}_i = (f_{i1}, \ldots, f_{iK})'$. Here, $f_{ik} = z_{ik}v_{ik}$, with $z_{ik} = 1$ if individual i has latent risk factor k and 0 otherwise, while v_{ik} denotes the severity of risk factor k for individual i. For example, feature k may represent the occurrence of an undiagnosed mild stroke, while v_{ik} represents how severe the stroke is, with more severe stroke resulting in more frequent neurological problems.

Such data would not be well characterized with a typical latent class model, which requires individuals to be grouped into a single set of classes. However, the approach of Ghahramani et al. is also not ideal in this case, as there are two important drawbacks. First, the assumption of feature exchangeability makes inferences on the latent features awkward. Thus, across posterior samples collected using an MCMC algorithm, the feature index changes meaning. This label ambiguity also also occurs in DPM models. A solution in the setting of ILF models is to choose a prior that explicitly orders the features by their frequency of occurrence, with feature one being the most common. Second, one can potentially characterize the data using fewer features by modeling the feature scores $\{v_{ik}\}$ nonparametrically. This also provides a more realistic characterization of the data. By assuming a parametric model, one artificially inflates the number of features needed to fit the data, making the latent features less likely to characterize a true unobserved risk factor.

An exponentiated gamma Dirichlet process (EGDP) prior can address both of these issues. I first define the exponentiated gamma (EG) component of the prior, which provides a probability model for the random matrix, $\mathbf{Z}$. Without loss of generality, the features are ordered, so that the first trait tends to be more common in the population, and the features decrease stochastically in population frequency with increasing index h. This is accomplished by letting

$$\pi_h = 1 - \exp(-\gamma_h), \quad \gamma_h \overset{ind}{\sim} \mathcal{G}(1, \beta_h), \quad \text{for } h = 1, \ldots, \infty, \tag{1}$$

where $\boldsymbol{\gamma} = \{\gamma_h, h = 1, \ldots, \infty\}$ is a stochastically decreasing infinite sequence of independent gamma random variables, with the stochastic decreasing constraint ensured by letting $\beta_1 < \beta_2 < \ldots < \beta_\infty$. Marginalizing over the prior for $\boldsymbol{\gamma}$, we obtain

$$\begin{aligned}\Pr(Z_{ih} = 1 \,|\, \boldsymbol{\beta}) &= 1 - \int_0^\infty \exp(-\gamma_h)\, \beta_h \, \exp(-\gamma_h \beta_h)\, d\gamma_h \\ &= \frac{1}{1+\beta_h}, \end{aligned} \tag{2}$$

which is decreasing in h for increasing $\boldsymbol{\beta} = \{\beta_h, h = 1, \ldots, \infty\}$.

Note that, unlike for the IBP, the exponentiated gamma (EG) process defined in (1) does not result in a Poisson distribution for $S_i = \sum_{h=1}^\infty Z_{ih}$, the number of traits per subject. Instead S_i is defined as the convolution of independent but not identically distributed Bernoulli random variables. A convenient special case corresponds to

$$\beta_h = \exp\{\psi_1 + \psi_2(h-1)\}, \quad h = 1, 2, \ldots, \infty, \tag{3}$$

which results in a logistic regression model for the frequency of trait occurrence upon marginalizing out $\boldsymbol{\gamma}$. In this case, two hyperparameters, ψ_1 and ψ_2, characterize the EG process, with ψ_1 controlling the frequency of trait one and ψ_2 controlling how rapidly traits decrease in frequency with the index h. The restriction $\psi_2 > 0$ ensures that $\beta_1 < \beta_2 < \ldots < \beta_\infty$. Assuming (1) and (3), it is straightforward to show that the distribution of S_i can be accurately approximated by the distribution of $S_{iT} = \sum_{i=1}^T Z_{ih}$ for sufficiently large T. In most applications, a sparse representation with few dominant features (expressed by choosing $\psi \geq 1$) may be preferred. In such cases, an accurate truncation approximation can be produced by replacing $\mathbf{F} = \mathbf{Z} \bigotimes \mathbf{V}$ with $\mathbf{F}_T = \mathbf{Z}_T \bigotimes \mathbf{V}_T$, with $\bigotimes$ denoting the element-wise product, and $\mathbf{A}_T$ denoting the submatrix of $\mathbf{A}$ consisting of the first T columns. Here, T is a finite integer, e.g., $T = 20$ or $T = 50$.

Expressions (1) and (3) provide a prior for the random binary matrix, $\mathbf{Z}$, allocating features to subjects. In order to complete the EGDP specification, we let $v_{ih} = 0$ if $z_{ih} = 0$

and otherwise

$$(v_{ih} \,|\, z_{ih} = 1) \sim G_h, \quad G_h \sim DP(\alpha G_0). \tag{4}$$

Here, G_h represents a random probability measure characterizing the distribution of the hth latent feature score among those individuals with the feature. This probability measure is drawn from a Dirichlet process (DP) with base measure G_0 and precision α.

3 Nonparametric Latent Factor Models

To illustrate the EGDP, we focus on a nonparametric extension of factor analysis. For subjects $i = 1, \ldots, n$, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})'$ denote a multivariate response vector. Then, a typical factor analytic model can be expressed as:

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}), \tag{5}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ is a mean vector, $\boldsymbol{\Lambda}$ is a $p \times K$ factor loadings matrix, $\mathbf{f}_i = (f_{i1}, \ldots, f_{iK})'$ is a $K \times 1$ vector of latent factors, and $\boldsymbol{\epsilon}_i$ is a normal residual with diagonal covariance $\boldsymbol{\Sigma}$ (see, for example, Lopes and West, 2004). In a parametric specification, one typically assumes $f_{ih} \sim N(0, 1)$, while constraining the factor loadings matrix $\boldsymbol{\Lambda}$ to ensure identifiability.

Instead we let $\mathbf{f}_i \sim F$, with $F \sim EGDP(\boldsymbol{\psi}, \alpha, G_0)$, where F denotes the unknown distribution of $\mathbf{f}_i$ and $EGDP(\boldsymbol{\psi}, \alpha, G_0)$ is shorthand notation for the exponentiated gamma Dirichlet process prior with hyperparameters $\boldsymbol{\psi} = (\psi_1, \psi_2)'$, α and G_0. Due to the constraint that the higher numbered factors correspond to rarer features that are less frequent in the population, we avoid the need to constrain $\boldsymbol{\Lambda}$. To remove sign ambiguity, we instead restrict G_0 to have strictly positive support, ensuring that $f_{ih} \geq 0$ for all i, h.

Note that this characterization has several appealing properties. First, the distributions of the factor scores are modelled nonparametrically, with subjects automatically clustered into groups separately for each factor. One of these groups corresponds to the cluster of subjects not having the factor, while the others are formed through the discreteness property of the DP. Second, the formulation automatically allows an unknown number of factors represented among the subjects in the sample. Thus, uncertainty in the number of factors is accommodated in a very different manner from Lopes and West (2004). Third, for G_0 chosen to be truncated normal, posterior computation can proceed efficiency via a data augmentation MCMC algorithm. Using a truncation approximation (say with $T = 20$), the algorithm alternately updates: (i) $\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}$ conditionally on $\mathbf{F}$ using Gibbs sampling steps; (ii) the elements of $\mathbf{Z}$ by sampling from the Bernoulli full conditional posterior distributions; (iii) $\{\gamma_h, h = 1, \ldots, T\}$ with a data augmentation step (relying on an approach similar to Dunson and Stanford, 2005 and Holmes and Held, 2006); (iv) $\mathbf{V}$ using standard algorithms for DPMs (MacEachern and Müller, 1998). Details are excluded given space considerations.

References

Carvalho, C.M., Lucas, J., Wang, Q., Nevins, J. and West, M. (2005) High-dimensional sparse factor models and latent factor regression. *ISDS Discussion Paper*, Duke University.

Dunson, D.B. (2006) Efficient Bayesian model averaging in factor analysis. *ISDS Discussion Paper*, Duke University.

Dunson, D.B. and Stanford, J.B. (2005) Bayesian inferences on predictors of conception probabilities. *Biometrics*, 61, 126-133.

Ferguson, T.S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209-230.

Ferguson, T.S. (1974) Prior distributions on spaces of probability measures. *The Annals of Statistics*, **2**, 615-629.

Holmes, C.C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**, 145-168.

Lopes, H.F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41-67.

MacEachern, S.N. and Müller, P. (1998) Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223-238.

West, M. (2003) Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics* **7**, 723-732.