A Discussion of "Nonparametric Bayes Modeling of Populations of Networks" by Durante, Dunson, and Vogelstein

Scott W. Linderman and David M. Blei
Columbia University

August 6, 2017

1 Introduction

We congratulate the authors on their excellent paper. While the modeling of *single* networks has received much attention, we agree with the authors that the modeling of *populations* of networks has been largely overlooked. Given that such data are increasingly common in fields such as neuroscience, this paper makes a timely contribution. In this discussion, we consider a factorial generalization of the proposed mixture of latent space models, and we suggest cases in which factorial models may naturally capture our intuition about the underlying generative process of the data. We compare these two models using the human brain data studied in the main paper, and we suggest some avenues for future work. Code to reproduce the figures in this discussion is available at https://github.com/blei-lab/factorial-network-models.

2 Model

Durante et al. (2016) propose a probabilistic model for populations of networks. These populations may be represented as collections of binary adjacency matrices, $\{\boldsymbol{A}_n\}_{n=1}^N$, where matrix $\boldsymbol{A}_n \in \{0,1\}^{V\times V}$ represents the connectivity of the n-th network. Entry $A_{n,[u,v]}$ is set to one if an edge is observed from vertex u to vertex v in network n; otherwise it is set to zero. We assume the V vertices are the same in all N networks. Moreover, the networks are undirected ($A_{n,[u,v]} \equiv A_{n,[v,u]}$) and without self-loops ($A_{n,[v,v]} \equiv 0$). Thus, it suffices to model only the lower triangular entries.

The authors build on the latent space model (LSM), a canonical model in probabilistic network analysis (Hoff et al., 2002; Hoff, 2008). An LSM is defined by the following parameters and latent variables: a bias $z_{u,v} \in \mathbb{R}$ for each edge; an embedding $\boldsymbol{x}_v \in \mathbb{R}^D$ for each vertex; and a positive-definite "scaling" matrix $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$, $\boldsymbol{\lambda} \in \mathbb{R}_+^D$, that determines the relative importance of the D latent dimensions. For convenience, let $\boldsymbol{Z} = \{\{z_{u,v}\}_{u=1}^V\}_{v=1}^{u-1}$ denote the set of per-connection biases. Given these parameters and latent variables, the edges are conditionally independent Bernoulli random variables, and the likelihood of a network is,

$$p(\boldsymbol{A}_n \,|\, \boldsymbol{Z}, \{\boldsymbol{x}_v\}_{v=1}^V, \boldsymbol{\Lambda}) = \prod_{u=1}^{V}\prod_{v=1}^{u-1} \mathrm{Bern}\left(A_{n,[u,v]} \,|\, \sigma(z_{u,v} + \boldsymbol{x}_u^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{x}_v)\right), \tag{1}$$

where $\text{Bern}(y \mid \rho) = \rho^y (1-\rho)^{1-y}$ is the Bernoulli likelihood function and $\sigma(x) = (1+e^{-x})^{-1}$ is the logistic function. The per-connection bias terms are only warranted when $N > 1$; otherwise, the model is over-parameterized. On top of this bias, the probability of connection between vertices u and v increases with the inner product between their embeddings $\boldsymbol{x}_u$ and $\boldsymbol{x}_v$, weighted by the matrix $\boldsymbol{\Lambda}$. In other words, the LSM is a low-rank model of connection log-odds.

LSM's can model a variety of individual network structures—simple Erdős-Rényi networks (Erdös and Rényi, 1959), small-world networks (Watts and Strogatz, 1998), scale-free networks (Barabási and Albert, 1999), and stochastic block models (Nowicki and Snijders, 2001). But a population of networks may exhibit a diversity of such connectivity patterns. The *mixture of latent space models* (MoLSM), as suggested by Durante et al. (2016), is naturally suited to this type of heterogeneous data.

Now let there be H separate mixture components, each with a unique set of vertex embeddings $\boldsymbol{x}_v^{(h)} \in \mathbb{R}^D$ and its own scaling factor $\boldsymbol{\lambda}^{(h)}$. Furthermore, let $h_n \in \{1, \ldots, H\}$ denote the mixture component to which the n-th network is attributed. The likelihood of a network is,

$$p\left(\boldsymbol{A}_n \mid \boldsymbol{Z}, \{\{\boldsymbol{x}_v^{(h)}\}_{v=1}^V, \boldsymbol{\lambda}^{(h)}\}_{h=1}^H, h_n\right) = \prod_{u=1}^{V} \prod_{v=1}^{u-1} \text{Bern}\left(A_{n,[u,v]} \mid \sigma(z_{u,v} + \boldsymbol{x}_u^{(h_n)\intercal} \boldsymbol{\Lambda}^{(h_n)} \boldsymbol{x}_v^{(h_n)})\right). \quad (2)$$

To regularize the model, the authors use a sparsity-inducing, multiplicative inverse gamma (MIG) prior on $\boldsymbol{\lambda}^{(h_n)}$,

$$\lambda_d^{(h)} = \prod_{d'=1}^{d} \left(\nu_{d'}^{(h)}\right)^{-1}, \qquad \nu_1^{(h)} \sim \text{Gamma}(a_1, 1), \qquad \nu_d^{(h)} \sim \text{Gamma}(a_2, 1). \quad (3)$$

This prior pushes $\lambda_d^{(h)}$ toward zero for larger values of d, incentiving the model to use as few dimensions as possible. Tuning a_1 and a_2 adjusts the strength of this prior.

For posterior inference in the MoLSM, Durante et al. (2016) use Pólya-gamma augmentation to develop a Gibbs sampler when $\boldsymbol{x}_v^{(h)}$ and $z_{u,v}$ have Gaussian priors. While the Bernoulli likelihoods are not conjugate with these Gaussian priors, conditioning on the Pólya-gamma auxiliary variables renders them so. These auxiliary variables have straightforward and naïvely parallelizable updates as well, making the overall algorithm highly efficient.

3 A Factorial Generalization

Mixture models are but one of many types of latent structure that may underlie a population of data. A mixture model attributes each data point (here, each network) to one mixture component, and given this assignment, the likelihood is a function of that component's parameters (here, the latent embeddings). Feature-based models offer an alternative generative story, where each data point possesses a set of features, and the likelihood is a function of the parameters associated with those features. In *factorial* feature models (Ghahramani, 1995; Ghahramani and Jordan, 1996; Meeds et al., 2007; Ghahramani et al., 2007), the features are discrete, and in the simplest case, binary. This offers a intuitive interpretation: each data point possesses a subset of possible features that contribute to its likelihood. Miller et al. (2009) developed factorial *single* network models; we derive a feature based, factorial latent space model for *populations* of networks, using the MoLSM as our starting point.

Consider the embeddings and scalings of the H mixture components. Specifically, let

$$\widetilde{\boldsymbol{x}}_v = \left[\boldsymbol{x}_v^{(1)\intercal}, \ldots, \boldsymbol{x}_v^{(H)\intercal}\right]^\intercal, \tag{4}$$

$$\widetilde{\boldsymbol{\lambda}} = \left[\boldsymbol{\lambda}^{(1)\intercal}, \ldots, \boldsymbol{\lambda}^{(H)\intercal}\right]^\intercal \tag{5}$$

denote column vectors in $\mathbb{R}^{D \cdot H}$ formed by concatenating the embeddings and scaling factors of each mixture component, respectively. Then, introduce a "mask" vector for each network defined by

$$\widetilde{\boldsymbol{m}}_n = \left[\mathbb{I}[h_n = 1] \cdot \mathbf{1}_D^\intercal, \ldots, \mathbb{I}[h_n = H] \cdot \mathbf{1}_D^\intercal\right]^\intercal. \tag{6}$$

Here $\mathbb{I}[\cdot]$ is an indicator that evaluates to one if its argument is true and zero otherwise, and $\mathbf{1}_D$ is a column vector of length D filled with ones. The mask represents the mixture component h_n as a vector with exactly D ones in the coordinates corresponding to $\boldsymbol{x}_v^{(h_n)}$ and $\boldsymbol{\lambda}^{(h_n)}$. The likelihood in Eq. (2) can now be equivalently expressed as

$$p\left(\boldsymbol{A}_n \mid \boldsymbol{Z}, \{\widetilde{\boldsymbol{x}}_v\}_{v=1}^V, \widetilde{\boldsymbol{\lambda}}, h_n\right) = \prod_{u=1}^{V} \prod_{v=1}^{u-1} \mathrm{Bern}\left(A_{n,[u,v]} \mid \sigma\left(z_{u,v} + \widetilde{\boldsymbol{x}}_u^\intercal \operatorname{diag}(\widetilde{\boldsymbol{\lambda}} \odot \widetilde{\boldsymbol{m}}_n) \widetilde{\boldsymbol{x}}_v\right)\right), \tag{7}$$

where $\odot$ denotes element-wise multiplication. The mask effectively turns on or off certain dimensions of the latent space according to the network's mixture assignment h_n.

This suggests an extension: rather than restricting the model to exactly H unique masks, instead allow each network to take on any of the $2^{D \cdot H}$ possible binary masks. More generally, let K denote the total number of latent factors (so far $K = D \cdot H$). Intuitively, the set of networks is characterized by K latent factors; in any given network, only a subset of those factors play a role in determining the edge probabilities. Unlike the mixture model, in which only H different subsets of factors are allowed, here any possible combination of factors can be chosen, making this a strict generalization of the mixture model. Moreover, with more flexibility in the choice of subset, it is likely that $K < D \cdot H$ dimensions will suffice. We call this the *factorial latent space model* (fLSM).

With finite K, a natural prior is

$$p(\widetilde{\boldsymbol{m}}_n \mid \{\rho_k\}_{k=1}^K) = \prod_{k=1}^{K} \mathrm{Bern}(\widetilde{m}_{n,k} \mid \rho_k), \tag{8}$$

$$p(\rho_k; \alpha) = \mathrm{Beta}(\rho_k \mid \tfrac{\alpha}{K}, 1), \tag{9}$$

where α is a hyperparameter that controls the sparsity of the mask matrices. One of the primary contributions of Durante et al. (2016) is a Bayesian nonparametric model that grows in complexity (number of mixture components, number of latent dimensions per component) as the data demands. We achieve similar flexibility here: as K goes to infinity, the prior (8)–(9) converges to a Bayesian nonparametric prior known as the Indian buffet process (Griffiths and Ghahramani, 2005). Intuitively, each new network has nonzero probability of introducing a new latent factor, i.e. of increasing the dimensionality of the embeddings.

Posterior inference in the factorial model requires modifications to the mixture model inference algorithm. Rather than sampling a mixture identity h_n for each network, we now must sample the binary mask $\widetilde{\boldsymbol{m}}_n$. Assuming a large but finite value of K—which is akin to the "weak limit" approximation used by Durante et al. (2016)—we may Gibbs sample each coordinate of the mask vector $\widetilde{m}_{n,k}$ holding the remainder $\widetilde{\boldsymbol{m}}_{n,\neg k}$ fixed. These conditionals are of the same form as the class conditional probabilities in the mixture model,

$$p(\widetilde{m}_{n,k} \mid \widetilde{\boldsymbol{m}}_{n,\neg k}, \boldsymbol{A}_n, \boldsymbol{Z}, \{\widetilde{\boldsymbol{x}}_v\}_{v=1}^V, \widetilde{\boldsymbol{\lambda}}, \rho_k) \propto \mathrm{Bern}(\widetilde{m}_{n,k} \mid \rho_k) \times p\left(\boldsymbol{A}_n \mid \boldsymbol{Z}, \{\widetilde{\boldsymbol{x}}_v\}_{v=1}^V, \widetilde{\boldsymbol{\lambda}}, \widetilde{\boldsymbol{m}}_n\right). \tag{10}$$

In sampling the binary masks, we effectively sample the number of factors active in a network. Sparsity inducing priors on $\widetilde{\lambda}_k$ are no longer necessary; they are superseded by the prior on the binary masks.

Of course, we also pay a computational cost in generalizing from mixtures to factorial models. In the mixture model, the conditional distribution of h_n could be computed exactly, but the conditional distribution of $\widetilde{\boldsymbol{m}}_n$ in the factorial model may take on 2^K values, which will generally be intractable. Another concern is that the coordinate-wise Gibbs sampler proposed above will suffer poor mixing. For example, $\widetilde{m}_{n,k}$ may be highly correlated with $\{\widetilde{x}_{v,k}\}_{v=1}^V$, making it hard to turn off or on a factor while holding the embeddings fixed. Indeed, this is also a concern in mixture models where mixture assignments and parameters may be strongly coupled. In some models it is possible to address this concern by integrating out the embeddings when sampling the binary masks; such a collapsed sampler would also benefit the mixture model. Unfortunately, due to the quadratic form in the edge probabilities, this marginalization does not appear to be straightforward, even after Pólya-gamma augmentation. In the experiments below, however, we find that the simple, uncollapsed Gibbs sampler suffices for these data.

4 Experiments

We compared the performance of the standard latent space model (LSM), the mixture of latent space models (MoLSM), and the proposed factorial generalization (fLSM). We studied the same brain connectivity data as Durante et al. (2016). This dataset contains $N = 42$ networks, each with $V = 68$ vertices. We held out 24062 randomly chosen edges for testing—approximately 25% of the $\frac{1}{2}NV(V-1) = 95676$ total number of edges—and trained the models on the remaining edges. A simple independent Bernoulli model (i.e. $A_{n,[u,v]} \sim \text{Bern}(p_{u,v})$) achieves 77.8% accuracy on the training data and 76.5% accuracy on the test data, indicating that the networks are fairly stereotyped.

We measured performance as a function of number of latent dimensions. For the LSM and fLSM, we varied the number of latent dimensions from 2 to 20. For the MoLSM, we fixed the number of mixture components to $H = 10$ and we varied the number of dimensions per component from $D = 2$ to $D = 20$. The MoLSM has $H \cdot D$ dimensions in total, but each network can only use D latent factors. We also studied the effect of the multiplicative inverse gamma (MIG) prior for the LSM and the MoLSM, as opposed to a simple inverse gamma prior. (As argued above, the fLSM already benefits from a sparsity inducing prior on the number of components used by any network.)

We assessed the convergence of the Gibbs sampler by inspecting the log joint probability of the training data. We found that the samplers typically converged in under 100 iterations for all models. Thus, we fit each model with 500 iterations of Gibbs sampling, and computed training and test likelihood estimates by averaging over the last 250 iterations. For each sample of the latent variables and parameters ($\boldsymbol{Z}$, $\{\widetilde{\boldsymbol{x}}_b\}_{v=1}^V$, etc.), we computed the log likelihood (7) of the training and test data. To facilitate comparison, we standardized these log likelihoods by subtracting the log likelihood under an independent Bernoulli model and normalizing by the number of entries. The MIG hyperparameters were set to $a_1 = 2.5$ and $a_2 = 3.5$, and the mixture models were initialized with k-means clustering, as in Durante et al. (2016). The fLSM hyperparameter was set to $\alpha = \frac{K}{2} + 1$, such that the prior probability of each feature is about $\frac{1}{3}$ in expectation.

Figure 1 shows the log likelihoods of the training and test data relative to the Bernoulli baseline. We will consider each column in turn, working from left to right. (i) The standard LSM shows continued improvement in training log likelihood as the number of dimensions increases (red), though the test likelihood plateaus after about 10 dimensions. (ii) Incorpo-

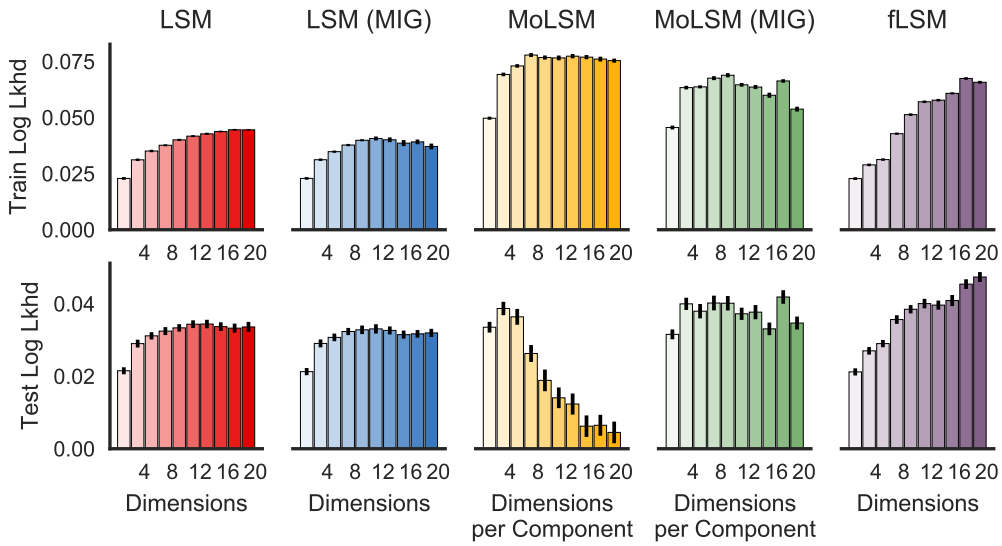

Figure 1: *Train and test likelihood for various latent space models (LSM) on brain network data as a function of latent dimensionality. Without a sparsity-inducing MIG prior, the mixture of LSMs (MoLSM) tends to over-fit the data, and with a prior its performance shows an improvement of* ≈ 0.01 *nats/connection over the LSM. The factorial LSM (fLSM) achieves high test performance by sharing embeddings across the entire population while letting individual networks use only a subset of the dimensions. It provides a further* ≈ 0.01 *nats/connection improvement over the MoLSM. Error bars denote* ± 1 *standard deviation.*

rating an MIG prior biases the model toward using fewer components, but this additional regularization is unnecessary for the standard LSM on this dataset (blue). (iii) By contrast, the mixture model shows severe over-fitting in the absence of regularization (yellow). Recall that here we are varying the number of dimensions per component, and there are ten components total. Some mixture components only have two to five networks assigned to them, which is not sufficient to accurately infer the embedding. (iv) With an MIG prior (green), the MoLSM no longer exhibits this over-fitting, and the test likelihoods exceed those of the standard LSM by about 0.01 nats/connection. (v) Finally, the fLSM (purple) seems to strike a nice balance. It selectively uses factors, permitting network-to-network variability while enforcing parameter sharing across the population. The fLSM achieves highest performance: the baseline performance is 76.5% accuracy on test data, and the 0.05 nat improvement yields 80.5% accuracy.

Figure 2 shows the inferred factors of the fLSM. The left panel is a sample from the posterior distribution of factor usage (i.e. a sample of $\{\tilde{\boldsymbol{m}}_n\}_{n=1}^N$). The factors are sorted in decreasing order of magnitude $\|\bar{\boldsymbol{x}}_k\|_2$, where $\bar{\boldsymbol{x}}_k = \left[\widetilde{x}_{1,k}, \ldots, \widetilde{x}_{V,k}\right]$. The first twelve factors are employed by almost all networks, but the lower magnitude factors are used more variably. While this model was given 20 latent dimensions, only 15 were used in this sample. This variability likely reflects both heterogeneity in the population and posterior uncertainty.

The right panels of Figure 2 show the average network (similar to Figure 9 of Durante et al. (2016)) and then the contributions of a few of the individual factors, i.e. $\bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^\mathsf{T}$. Intuitively, the fLSM models the log-odds of a binary adjacency matrix by summing a subset of these contributions. Note how the highest magnitude factors encode the strong block structure while the lower magnitude factors capture more subtle patterns. One next step would be to compare factor usage to other covariates of the patients. We expect that differences in network connectivity arising from the features of an individual patient (age, disease history, etc.) would be reflected in differences in factor usage as well.

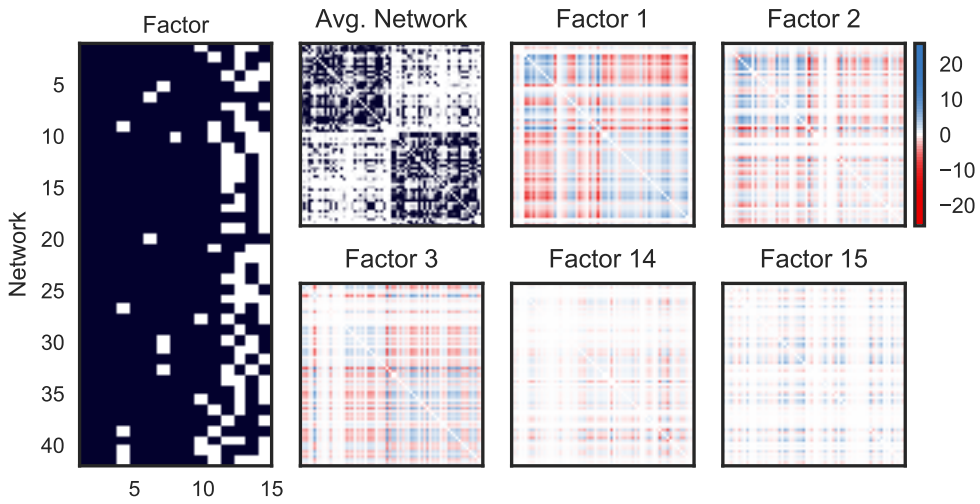

Figure 2: *Inferred factors of the LSM and their usage. Left: a sample from the posterior distribution of factor usage. The n-th row corresponds to the binary vector $\widetilde{\boldsymbol{m}}_n$, where black denotes one and white denotes zero. Top: the average network $\frac{1}{N}\sum_n \boldsymbol{A}_n$ exhibits strong block structure reflecting within-hemisphere connectivity. Right: individual contributions of a subset of factors, as given by $\bar{x}_k \bar{x}_k^{\mathsf{T}}$. Blue denotes increased probability of connection; red denotes decreased. Under the fLSM, each network n sums a subset of these factors.*

5 Conclusion

Durante et al. (2016) present a mixture of latent space models for capturing low-dimensional structure in populations of networks. Mixing over latent embeddings handles network-to-network variability within the population and provides promising results compared to other hierarchical models that assume a single set of edge probabilities across all networks. Their Bayesian nonparametric approach allows flexible inference of the number of mixture components and latent dimensions in a data-driven manner.

We present a generalization from the mixture model to a factorial model in which the population of networks share a set of latent factors, but only a subset of those factors actively determine edge probabilities for any given network. For example, in studying brain connectivity, the factorial model postulates that a set of latent features determines the probability of fiber tracts between two brain regions (e.g. age, diseases, drug use), and each patient may possess a different subset of these features. By contrast, the mixture model postulates a collection of different "types" of patients, each with characteristic patterns of connectivity. These two models are similar, but they may yield different insights into the population of networks under study.

This suggests a number of avenues for future work. In many cases, the network itself is not directly observable; rather, we have access to only indirect measurements, like neural spike trains, which provide statistical information about the presence or absence of an edge between two vertices. In these cases, we would like to perform joint inference of the network *and* its underlying structure, as in Linderman and Adams (2014) and Linderman et al. (2016). In other experiments, like longitudinal studies of brain connectivity, the observed networks are not exchangeable draws from a latent mixture or factor model, but rather a sequence of observations that we expect to be correlated over time, as in other dynamic network models (e.g. Bassett et al., 2011; Linderman et al., 2014). Likewise, while Durante et al. (2016) mix over complete network models, a natural alternative is to mix over per-edge probabilities, as in the mixed-membership stochastic block model (Airoldi

et al., 2008). In such models, each vertex pair has an associated pair of latent discrete "roles", and the probability of connection is a function of these underlying roles. Finally, as the authors mention, scaling inference to larger networks is another area for future work. Variational inference algorithms could be well suited to this challenge (Jordan et al., 1999; Blei et al., 2017).

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., and Grafton, S. T. (2011). Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, 108(18):7641–7646.

Blei, D., Kucukelbir, A., and McAuliffe, J. (2017). Variational inference: A review for statisticians. *Journal of American Statistical Association*, 112(518):859–877.

Durante, D., Dunson, D. B., and Vogelstein, J. T. (2016). Nonparametric Bayes modeling of populations of networks. *Journal of the American Statistical Association*, (just-accepted).

Erdös, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae (Debrecen)*, 6:290–297.

Ghahramani, Z. (1995). Factorial learning and the EM algorithm. In *Advances in Neural Information Processing Systems*, pages 617–624.

Ghahramani, Z., Griffiths, T. L., and Sollich, P. (2007). Bayesian nonparametric latent feature models.

Ghahramani, Z. and Jordan, M. I. (1996). Factorial hidden Markov models. In *Advances in Neural Information Processing Systems*, pages 472–478.

Griffiths, T. L. and Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, volume 18, pages 475–482.

Hoff, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, pages 657–664.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Linderman, S. and Adams, R. (2014). Discovering latent network structure in point process data. In *International Conference on Machine Learning*, pages 1413–1421.

Linderman, S., Adams, R. P., and Pillow, J. W. (2016). Bayesian latent structure discovery from multi-neuron recordings. In *Advances in Neural Information Processing Systems*, pages 2002–2010.

Linderman, S., Stock, C. H., and Adams, R. P. (2014). A framework for studying synaptic plasticity with neural spike train data. In *Advances in Neural Information Processing Systems*, pages 2330–2338.

Meeds, E., Ghahramani, Z., Neal, R. M., and Roweis, S. T. (2007). Modeling dyadic data with binary latent factors. *Advances in Neural Information Processing Systems*, 19:977.

Miller, K., Jordan, M. I., and Griffiths, T. L. (2009). Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems*, pages 1276–1284.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic block structures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.