Latent Mixed-Membership Allocation Models of Relational and Multivariate Attribute Data

Edoardo M. Airoldi[1], David M. Blei[1], Stephen E. Fienberg[1,2], Eric P. Xing[1]
[1] School of Computer Science, [2] Department of Statistics
Carnegie Mellon University, Pittsburgh, PA 15213

We consider the the statistical analysis of a matrix of relations between objects of study. Our goal is to make predictions about these objects from the matrix, or to situate them in a low-dimensional latent space. Such data arise in biological settings, collections of author-recipient email, and social networks. Our models [7] combine the features of mixed-membership models [9, 10] and block models for relational data [5].

A fundamental problem of modern systems biology is to predict the functional annotations of proteins [3]. While experiment-based functional annotations are hard to obtain, large amounts of noisy relational information, e.g., interactions between proteins, are readily available from high-throughput experiments. One approach to inferring proteins' functional annotations from such data is to posit latent variables that indicate in which functional processes each protein takes part. If the way proteins interact is indicative of the functionality they are synthesized to perform, then the latent variables will correlate with the proteins' functional annotations which may be partially observable. Unlike in typical latent variable models, we infer such variables from a matrix of interactions rather than the attributes of the proteins themselves.

We develop a hierarchical Bayesian clustering model of relational data for such applications. Each object is associated with a latent vector of membership proportions; relationships between objects arise from interaction between their constituent cluster memberships. With a fixed number of clusters K, we assume that the relational matrix is drawn from the following generative process:

- For each object i:
  - Sample the mixed membership weights $\boldsymbol{\theta}_i \sim \text{Dirichlet}\,(\boldsymbol{\alpha})$.
- For each pair of objects i and j:
  - Sample the latent cluster indicator $\boldsymbol{z}_{i\to j} \sim \text{Multinomial}\,(\boldsymbol{\theta}_i, 1)$.
  - Sample the latent cluster indicator $\boldsymbol{z}_{i\leftarrow j} \sim \text{Multinomial}\,(\boldsymbol{\theta}_j, 1)$.
  - Sample the interaction $r_{ij} \sim p\,(\,\cdot\,|\,\boldsymbol{\eta}, \boldsymbol{z}_{i\to j}, \boldsymbol{z}_{i\leftarrow j})$.

This is a *mixed-membership stochastic block model* (MMSB). Each object is associated with a set of membership proportions $\boldsymbol{\theta}$. To determine the relationship between objects, we pick a cluster from the "sender" vector and one from the "receiver" vector; then, we look up the corresponding parameter in $\boldsymbol{\eta}$, a $K \times K$ collection of parameters, to determine the distribution from which to draw the relationship. This model stands in contrast to the stochastic block model [5], where objects are associated with a single cluster.

There is a relationship between the MMSB and the latent space model of relational data [4]. In the latent space model, the latent vectors are drawn from Gaussian distributions and the interaction data is drawn from a Gaussian with mean $\boldsymbol{\theta}_i' I \boldsymbol{\theta}_j$. In the MMSB, the marginal probability of an interaction takes a similar form, $p(r_{i,j}|\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, \boldsymbol{\eta}) = \boldsymbol{\theta}_i' \boldsymbol{M} \boldsymbol{\theta}_j$, where $M_{ij} = p(r_{i,j}|\boldsymbol{\eta})$ is a matrix of probabilities for each pair of latent functional states in the collection. In contrast to the latent space model, the interaction data can be modeled by an arbitrary distribution, in our model. With binary relationships, i.e., a graph, we can use a collection of Bernoulli parameters; with continuous

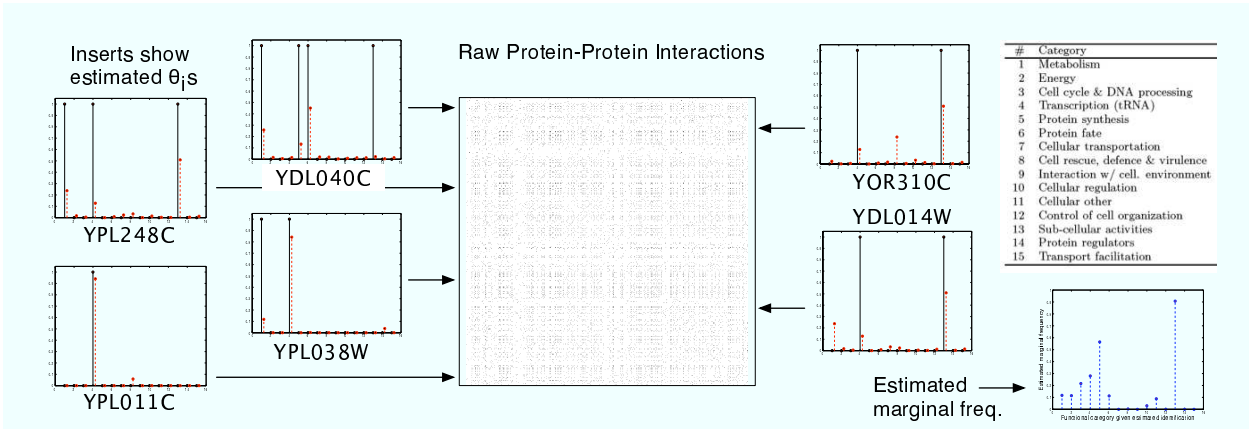

Figure 1: The central panel displays the raw protein-protein interactions. The inserts show manually curated functional annotations (black, solid bars) for six example proteins, versus the estimated mixed-membership scores (red, dashed bars). The 15 bars in each insert correspond to the functional categories described in the top right panel. The bottom right panel shows the estimated frequencies of membership, given the estimated mapping of latent clusters to functional categories.

relationships, we can use a collection of Gaussian parameters. While more flexible, the MMSB does not subsume the latent space model; they make different assumptions about the data.

As with many hierarchical Bayesian models, the posterior distribution of the latent variables is intractable to compute. We approximate this posterior using variational methods, a fast and deterministic alternative to MCMC. We posit a simple distribution of the latent variables with free parameters known as *variational parameters*, chosen to minimize the KL divergence between the variational distribution and the true posterior. The fitted variational distribution can be used as a substitute for the posterior. Compared to MCMC sampling schemes, variational methods trade off sampling until equilibrium for a deterministic optimization procedure based on analytically tractable fixed-point iterations. Thus they avoid issues such as complicated convergence tests and memory requirements for large data sets, significantly simplifying the use of Bayesian techniques; however, they do not come with the same theoretical guarantees as MCMC samplers. The approximate posterior may not become arbitrarily close to the true posterior.

In the MMSB, the latent variables are the membership vectors $\boldsymbol{\theta}_i$ and cluster indicators $\boldsymbol{z}_{i\to j}$ and $\boldsymbol{z}_{i\leftarrow j}$, which we take to be independent in a fully factorized variational distribution:

$$q(\boldsymbol{\theta}, \boldsymbol{z}|\gamma, \phi) \quad = \quad \prod_{i=1}^{N} \Big[\; q(\boldsymbol{\theta}_i|\gamma_i) \prod_{j=1}^{N} \Big(q(\boldsymbol{z}_{i\to j}|\boldsymbol{\phi}_{i\to j})\, q(\boldsymbol{z}_{i\leftarrow j}|\boldsymbol{\phi}_{i\leftarrow j}) \Big) \; \Big],$$

where $\boldsymbol{\gamma}$ are Dirichlet parameters and $\boldsymbol{\phi}_{i\to j}$ and $\boldsymbol{\phi}_{i\leftarrow j}$ are multinomial parameters. We minimize the KL divergence to the true posterior by coordinate ascent [11]:

$$\boldsymbol{\phi}^*_{i\to j} \propto \exp\left\{ \; \mathbb{E}_q[\log \boldsymbol{\theta}_i|\gamma_i] \; \right\} \cdot \boldsymbol{p}\left(\boldsymbol{r}_{ij}|\boldsymbol{\eta}, \boldsymbol{z}_{i\to j}, \boldsymbol{z}_{i\leftarrow j}\right) \qquad \boldsymbol{\phi}^*_{i\leftarrow j} \propto \exp\left\{ \; \mathbb{E}_q[\log \boldsymbol{\theta}_j|\gamma_j] \; \right\} \cdot \boldsymbol{p}\left(\boldsymbol{r}_{ij}|\boldsymbol{\eta}, \boldsymbol{z}_{i\to j}, \boldsymbol{z}_{i\leftarrow j}\right)$$

$$\boldsymbol{\gamma}^*_i \quad = \quad \boldsymbol{\alpha} + \textstyle\sum_{j=1}^{N} \boldsymbol{\phi}_{i\to j} + \sum_{j=1}^{N} \boldsymbol{\phi}_{i\leftarrow j}.$$

We can use the fitted variational distribution to form a lower bound on the log likelihood of the observations, $\mathbb{L}[\boldsymbol{\gamma}^*, \boldsymbol{\phi}^*; \boldsymbol{\alpha}, \boldsymbol{\eta}]$. Using this bound as a tractable surrogate for the likelihood, we find (pseudo) empirical Bayes estimates for the hyper-parameters.

A collapsed Gibbs sampler would require a state-space of N^2 variables which, for applications such as protein interaction modeling, could be in the tens or hundreds of millions. In contrast, the variational algorithms require storage of $(N+2)K$ variables, and note that $K << N$ in most applications. Moreoever, the variational method is parallelizable. In simulations, we show that our method outperforms spectral clustering [6], both when observed interactions are primarily the outcome of single membership, and when they are the outcome of mixed membership.

We report on the analysis of a set of about 750,000 interactions among 871 proteins in Yeast.

The data does not support the hypothesis that proteins interact only when they take part in the same latent process, which correlates to a single function. Rather, the best unsupervised analysis suggests that most proteins interact as they take part in multiple latent processes, which correlate to different functional categories, e.g., transcription and protein synthesis. We demonstrate how to leverage a small subset of functional annotations, e.g., manually curated thus more reliable, to inform the inference process and ultimately de-noise a larger set of protein-protein interactions.

There is a generalization of our model with the following generative process to integrate relational data, $\boldsymbol{r}$, with multivariate attributes, $\boldsymbol{x}$, and partially observed labels, $\boldsymbol{y}$.

$\boldsymbol{\theta} \sim p\,(\boldsymbol{\alpha})$	sample the mixed-membership scores
$(\boldsymbol{z}_x, \boldsymbol{z}_r) \sim p\,(\boldsymbol{z} \,|\, \boldsymbol{\theta})$	sample the latent indicators
$(\boldsymbol{x}, \boldsymbol{r}) \sim p\,(\boldsymbol{x}, \boldsymbol{r} \,|\, \boldsymbol{z}_x, \boldsymbol{z}_r, \boldsymbol{\beta}, \boldsymbol{\eta})$	sample the observations
$\boldsymbol{\zeta} \sim p\,(\boldsymbol{z}_x, \boldsymbol{z}_r)$	sample the predictive indicators given the latent indicators
$\boldsymbol{y} \sim glm\,(\boldsymbol{\zeta}, \boldsymbol{\delta})$	sample the attributes to be predicted

The corresponding variational inference algorithms are readily derived. The number of mixture components, K in the basic model above, need not be set a-priori [1]. Explicit temporal dependence can be introduced, for example, by means of a generalized state-space model on the parameters, $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\delta})$. Further, it is possible to capture systematic variations in local dynamics by positing a probabilistic model for the transition matrix [8]. We explore theoretical and computational issues associated with these models.

We consider two additional applications of our general formulation. In *dynamic networks analysis* we count the number of emails pairs of individuals exchange at each epoch, $\boldsymbol{r}$, and we have access to the content of those emails, $\boldsymbol{x}$. The individuals' attributes we want to predict, $\boldsymbol{y}$, may be whether or not they participate in certain activities, their corporate roles, and so on. The latent variables, $\boldsymbol{z}$, indicate the group memberships of pairs of individuals, which in turn explain both the topics discussed, and whether or not a communication occurs. The underlying parameters, $\boldsymbol{\theta}$, capture the mixed-membership of individuals to groups, and can be regarded as fixed through time. The dynamical behavior can be modeled as linear dynamical system on the $\boldsymbol{\alpha}$. In *multivariate sociometric analysis* we observe multiple sociometric relations between pairs of individuals, $\boldsymbol{r}$. The critical attribute we want to predict, $\boldsymbol{y}$ is, e.g., the individual propensity to gregariousness as measured by the number of positive social contacts [2]. The latent variables, $\boldsymbol{z}$, indicate the memberships of individuals as they decide whether to engage or not in a specific sociometric relation. The corresponding set of parameters, $\boldsymbol{\theta}$, capture the mixed-membership of individuals to groups, independently of the relation.

References

[1] M. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

[2] S. E. Fienberg, M. M. Meyer, and S. Wasserman. Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80:51–67, 1985.

[3] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, and K. Boutilier et. al. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415:180–183, 2002.

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

[5] K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96:1077–1087, 2001.

[6] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pages 1601–1608, 2004.

Additional References

[7] E. M. Airoldi, D. M. Blei, E. P. Xing, and S. E. Fienberg. A latent mixed-membership model for relational data. In *Workshop on Link Discovery: Issues, Approaches and Applications, in conjunction with the 10th International ACM SIGKDD Conference*, 2005.

[8] E. M. Airoldi and C. Faloutsos. Recovering latent time-series from their observed sums: network tomography with particle filters. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 30–39, 2004.

[9] D. M. Blei, M. I. Jordan, and A. Y. Ng. Hierarchical bayesian models for applications in information retrieval. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 7*, pages 25–44. Oxford University Press, 2003.

[10] E. A. Erosheva, S. E. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 97(22):11885–11892, 2004.

[11] E. P. Xing, M. I. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of UAI*, 2003.