A Latent Mixed Membership Model for Relational Data

Edoardo Airoldi, David Blei, Eric Xing
School of Computer Science
Carnegie Mellon University
{eairoldi,blei,xing}@cs.cmu.edu

Stephen Fienberg
Department of Statistics, and
School of Computer Science
Carnegie Mellon University
fienberg@stat.cmu.edu

ABSTRACT

Modeling relational data is an important problem for modern data analysis and machine learning. In this paper we propose a Bayesian model that uses a hierarchy of probabilistic assumptions about the way objects interact with one another in order to learn latent groups, their typical interaction patterns, and the degree of membership of objects to groups. Our model explains the data using a small set of parameters that can be reliably estimated with an efficient inference algorithm. In our approach, the set of probabilistic assumptions may be tailored to a specific application domain in order to incorporate intuitions and/or semantics of interest. We demonstrate our methods on simulated data and we successfully apply our model to a data set of protein-to-protein interactions.

Keywords

latent mixed-membership, hierarchical mixture model, variational inference, relational data, protein-protein interactions.

1. INTRODUCTION

Modeling relational data is an important problem for modern data analysis and machine learning. Many data sets contain interrelated observations. For example, scientific literature connects papers by citation, web graphs connect pages by links, and protein-protein interaction data connect proteins by physical interaction records. Such data violate the classical exchangeability assumptions made in machine learning and statistics; moreover, the relationships between data are often of interest as observations themselves. One may try to predict citations of newly written papers, predict the likely links of a web-page, or cluster proteins based on patterns of interaction between them.

There is a history of probabilistic models for relational data analysis in Statistics. Part of this literature is rooted in the stochastic block modeling ideas from psychometrics and sociology. These ideas are due primarily to Holland and Leinhardt, e.g., [12], and later elaborated upon by others, e.g., see [8, 23, 20, 11]. In machine learning, Markov random networks have been used for link prediction [21] and the traditional block models from Statistics have been extended with nonparametric Bayesian priors [13].

In this paper, we develop a mixed membership model for analyzing patters of interaction between data. Mixed membership models for soft classification have emerged as a powerful and popular analytical tool for analyzing large databases involving text [2], text and references [4, 7], text and images [1], multiple disability measures [6, 15], and genetics information [19, 18, 24]. These models use a simple generative model, such as bag-of-words or naive Bayes, embedded in a hierarchical Bayesian framework involving a latent variable structure; this induces dependencies and introduces statistical control over the estimation of what might otherwise be an extremely large set of parameters.

We propose a Bayesian model that uses a hierarchy of probabilistic assumptions about how objects interact with one another in order to learn latent groups, their typical interaction patterns, and the degree of membership of objects to groups. Given data, we find an approximate posterior distribution with an efficient variational inference algorithm. In our approach, the set of probabilistic assumptions may be tailored to a specific application domain in order to incorporate semantics of interest. We demonstrate our methods on simulated data, and we successfully apply the model to a data set of protein-protein interactions.

2. THE MODEL

In this section, we describe a probabilistic model of interaction patterns in a group of objects. Each object can exhibit several patterns that determine its relationships to the others. We will use protein-protein interaction modeling as a working example; however, the model can be used for any relational data where the primary goal of the analysis is to learn latent group interaction patterns and mixed group membership of a set of objects.

Suppose we have observed the physical interactions between N proteins [1]. We represent the interaction data by an $N \times N$ binary adjacency matrix $\boldsymbol{r}$ where $r_{i,j} = 1$ if the ith protein interacts with the jth protein. Usually, an interaction between a pair of proteins is indicative of a unique

[1]Such information can be obtained experimentally with "yeast two-hybrid" tests and others means, and in practice the data may be noisy. For simplicity, we defer explicit treatment of observation noise, although plugging in appropriate error processes is possible.

Link-KDD '05, August 21, 2005, Chicago, Illinois, USA.

biological function they both involve; it may be possible to infer the functional classes of the study proteins from the protein interactions.

In a complex biological system, many proteins are functionally versatile and can participate in multiple functions or processes at different times or under different biological conditions. Thus, when modeling functional classes of the proteins, it is natural to adopt a flexible model which allows multiple scenarios under which a protein can interact with its partners. For example, a signal transduction protein may sometimes interact with a cellular membrane protein as part of a signal receptor; at another time, it may interact with the transcription complex as an auxiliary transcription factor. By assessing the similarity of observed protein-to-protein interaction patterns, we aim to recover the latent function groups and the degree with which the proteins take part in them.

In the generative process, we model the observed adjacency matrix as a collection of Bernoulli random variables. For each pair of objects, the presence or absence of an interaction is drawn by (1) choosing a latent class for each protein from a protein-specific distribution and (2) drawing from a Bernoulli distribution with parameter associated with the pair of latent classes. A protein represents several functional groups through its distribution of latent classes; however, each protein participates in one function when determining its relationship to another.

For a model with K groups, the parameters are K-dimensional Dirichlet parameters α, a $K \times K$ matrix of Bernoulli parameters η, and $\rho \in [0, 1]$ which is described below. Each θ_i is a Dirichlet random variable (i.e., a point on the $K-1$ simplex) and each z_{ij1} and z_{ij2} are indicators into the K groups. The generative process of the observations, $r_{(N \times N)}$, is as follows:

1. For each object $i = 1, \ldots, N$:
   1.1. Sample $\boldsymbol{\theta}_i \sim \mathit{Dirichlet}\,(\boldsymbol{\alpha})$.
2. For each pair of objects $(i, j) \in [1, N] \times [1, N]$:
   2.1. Sample group $\boldsymbol{z}_{i,j,1} \sim \mathit{Multinomial}\,(\boldsymbol{\theta}_i, 1)$
   2.2. Sample group $\boldsymbol{z}_{i,j,2} \sim \mathit{Multinomial}\,(\boldsymbol{\theta}_j, 1)$
   2.3. Sample $r_{i,j} \sim \mathit{Bernoulli}\ \left(\rho\, \eta_{z_{i,j,1}, z_{i,j,2}} + (1 - \rho)\, \delta_0 \right)$

The parameter ρ controls how often a zero is due to noise and how often it occurs as a function of the constituent proteins' latent class memberships in the generative process. In turn, this leads to ones in the matrix being weighted more as ρ decreases, and allows for the model to pick up sparsely interconnected clusters. For the rest, the model uses three sets of latent variables. The $\boldsymbol{\theta}_i$s are sampled once for the entire collection of observations; the $\boldsymbol{z}_{i,j,1}$s and $\boldsymbol{z}_{i,j,2}$s are sampled once for each protein-protein interaction variable $r_{i,j}$.

The generative process described above leads to a joint probability distribution over the observations and the latent variables,

$$p(\boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{z}_1, \boldsymbol{z}_2 | \boldsymbol{\alpha}, \boldsymbol{\eta}) = \prod_{i=1}^{N} p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}) \prod_{j=1}^{N} p(\boldsymbol{z}_{i,j,1}|\boldsymbol{\theta}_i) \times \times p(\boldsymbol{z}_{i,j,2}|\boldsymbol{\theta}_j)\, p(r_{i,j}|\boldsymbol{z}_{i,j}, \boldsymbol{\eta}).$$

The marginal probability of the observations is not tractable to compute,

$$p(\boldsymbol{r}|\boldsymbol{\alpha}, \boldsymbol{\eta}) = \int_{\Theta} \int_{Z} \prod_{i=1}^{N} p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}) \prod_{j=1}^{N} p(\boldsymbol{z}_{i,j,1}|\boldsymbol{\theta}_i) \times \times p(\boldsymbol{z}_{i,j,2}|\boldsymbol{\theta}_j)\, p(r_{i,j}|\boldsymbol{z}_{i,j}, \boldsymbol{\eta})\, d\boldsymbol{z}\, d\boldsymbol{\theta}.$$

We carry out *approximate* inference and parameter estimation to deal with this issue.

The only input to this model is the number of groups. The goal is to learn the posterior distribution of the membership proportions of each protein and the group interaction probabilities. We will focus on the interpretability of these quantities, e.g., consistent functional annotations of the proteins within groups. Note that there are several ways to select the number of groups. For example, [13] uses a nonparametric Bayesian prior for a single-membership block model.

In our fully generative approach, it is possible to integrate outside information about the objects into the hierarchy of probabilistic assumptions. For example, we can include outside information about the proteins into the generative process that includes the linkage. In citation data, document words can be modeled along with how the documents cite each other.

3. INFERENCE AND ESTIMATION

In this section we present the elements of approximate inference essential for learning the hyper-parameters of the model and inferring the posterior distribution of the degrees of membership for each object.

In order to learn the hyper-parameters we need to be able to evaluate the likelihood, which involves a non-tractable integral as we stated above—see equation. In order to infer the degrees of membership corresponding to each object, we need to compute the posterior degrees of membership given the hyper-parameters and the observations

$$p(\boldsymbol{\theta}|\boldsymbol{r}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{r}|\boldsymbol{\alpha}, \boldsymbol{\eta})}{p(\boldsymbol{r}|\boldsymbol{\alpha}, \boldsymbol{\eta})}, \tag{1}$$

Using variational methods, we can find a lower bound of the likelihood and approximate posterior distributions for each object's membership vector.

The basic idea behind variational methods is to posit a variational distribution on the latent variables $q(\boldsymbol{\theta}, \boldsymbol{z})$, which is fit to be close to the true posterior in Kullback-Leibler (KL) divergence. This corresponds to maximizing a lower bound, $\mathbb{L}\left[\, \boldsymbol{\gamma}, \boldsymbol{\phi}\, ;\, \boldsymbol{\alpha}, \boldsymbol{\eta}\, \right]$, on the log probability of the observations given by Jensen's inequality:

$$\begin{aligned} \log p(\boldsymbol{r}|\boldsymbol{\alpha}, \boldsymbol{\eta}) \geq & \sum_{i=1}^{N} \mathbb{E}_q \left[\, \log p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}_i) \,\right] + \\ & + \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}_q \left[\, \log p(\boldsymbol{z}_{i,j,1}|\boldsymbol{\theta}_i) \,\right] + \\ & + \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}_q \left[\, \log p(\boldsymbol{r}_{i,j}|\boldsymbol{z}_{i,j}, \boldsymbol{\eta}) \,\right] + \\ & + \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}_q \left[\, \log p(\boldsymbol{z}_{i,j,2}|\boldsymbol{\theta}_j) \,\right] - \\ & - \mathbb{E}_q \left[\, \log q(\boldsymbol{\theta}, \boldsymbol{z}) \,\right]. \end{aligned}$$

1. initialize $\boldsymbol{\gamma}_{ig}^0 = \frac{2N}{K}$ for all i, g
2. **repeat**
3. **for** $i = 1$ to N
4. **for** $j = 1$ to N
5. get **variational** $\boldsymbol{\phi}_{ij1}^{t+1}$ and $\boldsymbol{\phi}_{ij2}^{t+1} = f(r_{ij}, \boldsymbol{\gamma}_i^t, \boldsymbol{\gamma}_j^t, \boldsymbol{\eta}^t)$
6. partially update $\boldsymbol{\gamma}_{\boldsymbol{i}}^{\boldsymbol{t+1}}$, $\boldsymbol{\gamma}_{\boldsymbol{j}}^{\boldsymbol{t+1}}$ and $\boldsymbol{\eta}^{\boldsymbol{t+1}}$
7. **until** convergence

1. initialize $\phi_{ij1g}^0 = \phi_{ij2h}^0 = \frac{1}{K}$ for all g, h
2. **repeat**
3. **for** $g = 1$ to K
4. update $\phi_{ij1g}^{s+1} \propto f_1(\boldsymbol{\phi}_{ij2}^s, \boldsymbol{\gamma}, \boldsymbol{\eta})$
5. normalize $\boldsymbol{\phi}_{ij1}^{s+1}$ to sum to 1
6. **for** $h = 1$ to K
7. update $\phi_{ij2h}^{s+1} \propto f_2(\boldsymbol{\phi}_{ij1}^s, \boldsymbol{\gamma}, \boldsymbol{\eta})$
8. normalize $\boldsymbol{\phi}_{ij2}^{s+1}$ to sum to 1
9. **until** convergence

Figure 1: Left: The two-layered variational inference for γ and ϕ. The inner layer consists of Step 5. The function f is described in details in the right panel. Right: Inference for the variational parameters $(\boldsymbol{\phi}_{ij1}, \boldsymbol{\phi}_{ij2})$ corresponding to the basic observation $r_{i,j}$. This is the detailed description of Step 5. in the left panel. The functions f_1 and f_2 are updates for ϕ_{ij1g} and ϕ_{ij2h} described in the text of Section 3.1.

where the expectations are taken with respect to $q(\boldsymbol{\theta}, \boldsymbol{z})$. We choose a fully factorized variational distribution such that this optimization is tractable.

3.1 Variational Inference

The fully factorized variational distribution q is as follows.

$$\begin{aligned} q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) &= \prod_{i=1}^{N} q(\boldsymbol{\theta}_i|\boldsymbol{\gamma}_i) \prod_{j=1}^{N} q(\boldsymbol{z}_{i,j,1}|\boldsymbol{\phi}_{i,j,1})\, q(\boldsymbol{z}_{i,j,2}|\boldsymbol{\phi}_{i,j,2}) \\ &= \prod_{i=1}^{N} Dirichlet\,(\boldsymbol{\theta}_i|\boldsymbol{\gamma}_i) \;\times \\ &\quad \times \prod_{j=1}^{N} \Big(Mult\,(\boldsymbol{z}_{i,j,1}|\boldsymbol{\phi}_{i,j,1})\, Mult\,(\boldsymbol{z}_{i,j,2}|\boldsymbol{\phi}_{i,j,2}) \Big) \end{aligned}$$

The lower bound for the log likelihood $\mathbb{L}[\boldsymbol{\gamma}, \boldsymbol{\phi}\,;\, \boldsymbol{\alpha}, \boldsymbol{\eta}]$ can be maximized using exponential family arguments and coordinate ascent [22]; this leads to the following updates for the variational parameters $(\boldsymbol{\phi}_{i,j,1}, \boldsymbol{\phi}_{i,j,2})$, for each pair (i, j):

$$\begin{aligned} \phi_{i,j,1,g}^* &\propto \exp\Big\{\, \psi(\gamma_{i,g}) - \psi(\sum_{g=1}^{K} \gamma_{i,g}) \,\Big\} \times \\ &\quad \times \prod_{h=1}^{K} \eta_{g,h}^{r_{i,j}\,\phi_{i,j,2,h}} \prod_{h=1}^{K} (1 - \eta_{g,h})^{(1 - r_{i,j})\,\phi_{i,j,2,h}} \\ \phi_{i,j,2,h}^* &\propto \exp\Big\{\, \psi(\gamma_{j,h}) - \psi(\sum_{h=1}^{K} \gamma_{j,h}) \,\Big\} \times \\ &\quad \times \prod_{g=1}^{K} \eta_{g,h}^{r_{i,j}\,\phi_{i,j,1,g}} \prod_{g=1}^{K} (1 - \eta_{g,h})^{(1 - r_{i,j})\,\phi_{i,j,1,g}} \end{aligned}$$

for $g, h = 1, \ldots, K$, and to the following updates for the variational parameters $\boldsymbol{\gamma}_i$, for each i:

$$\gamma_{i,g}^* = \alpha_t + \sum_{j=1}^{N} \phi_{i,j,1,g} + \sum_{j=1}^{N} \phi_{j,i,2,g}.$$

The vectors $\boldsymbol{\phi}_{i,j,1}$ and $\boldsymbol{\phi}_{i,j,2}$ are normalized to sum to one. The complete algorithm to perform variational inference in the model is described in detail in Figure 1. Variational inference is carried out for fixed values of $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$, in order to maximize the lower bound for the likelihood. Then we maximize the lower bound with respect to $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$. We iterate these two steps (variational inference and maximization) until convergence. The overall procedure is a variational expectation-maximization (EM) algorithm.

3.2 Remarks

The variational inference algorithm presented in Figure 1 is not the naïve variational inference algorithm. In the naïve version of the algorithm, we initialize the variational Dirichlet parameters $\boldsymbol{\gamma}_i$ and the variational Multinomial parameters $\boldsymbol{\phi}_{ij}$ to non-informative values, then we iterate until convergence the following two steps: (i) update $\boldsymbol{\phi}_{ij1}$ and $\boldsymbol{\phi}_{ij2}$ for all pairs (i, j), and (ii) update $\boldsymbol{\gamma}_i$ for all objects i. In such algorithm, at each variational inference cycle we need to allocate $NK + 2N^2K$ numbers.

The nested variational inference algorithm trades time for space thus allowing us to deal with large graphs; at each variational cycle we need to allocate $NK+2K$ numbers. The increased running time is partially offset by the fact that the algorithm can be parallelized and leads to empirically observed faster convergence rates, as we show in Figure 3. This algorithm is also better than MCMC variations (i.e., blocked and collapsed Gibbs samplers) in terms of memory requirements and/or convergence rates.

It is also important to note that the variational Dirichlet parameters $\boldsymbol{\gamma}$ and the Bernoulli parameters $\boldsymbol{\eta}$ are closely related in this model: it is necessary to keep the $\boldsymbol{\gamma}$s across variational-EM iterations in order to better inform the M-step estimates of $\boldsymbol{\eta}$. Thus, we smooth the $\boldsymbol{\gamma}$ parameters in between EM iterations instead of resetting them to a non-informative value, $2N/K$ in our model. Using a damping parameter ϵ we obtain: $\tilde{\gamma}_{i,g} = (1 - \epsilon)\,\gamma_{i,g}^* + \epsilon\,\frac{2N}{K}$.

3.3 Parameter Estimation

Using the optimal lower bound $\mathbb{L}[\boldsymbol{\gamma}^*, \boldsymbol{\phi}^*\,;\, \boldsymbol{\alpha}, \boldsymbol{\eta}]$ as a tractable surrogate for the likelihood we here look for (pseudo) empirical Bayes estimates for the hyper-parameters. [3]

Such maximization amounts to maximum likelihood estimation of the Dirichlet parameters α and Bernoulli parameter matrix η using expected sufficient statistics, where the expectation is taken with respect to the variational distribution. Finding the MLE of a Dirichlet requires numerical optimization. [17] For each Bernoulli parameter, the ap-

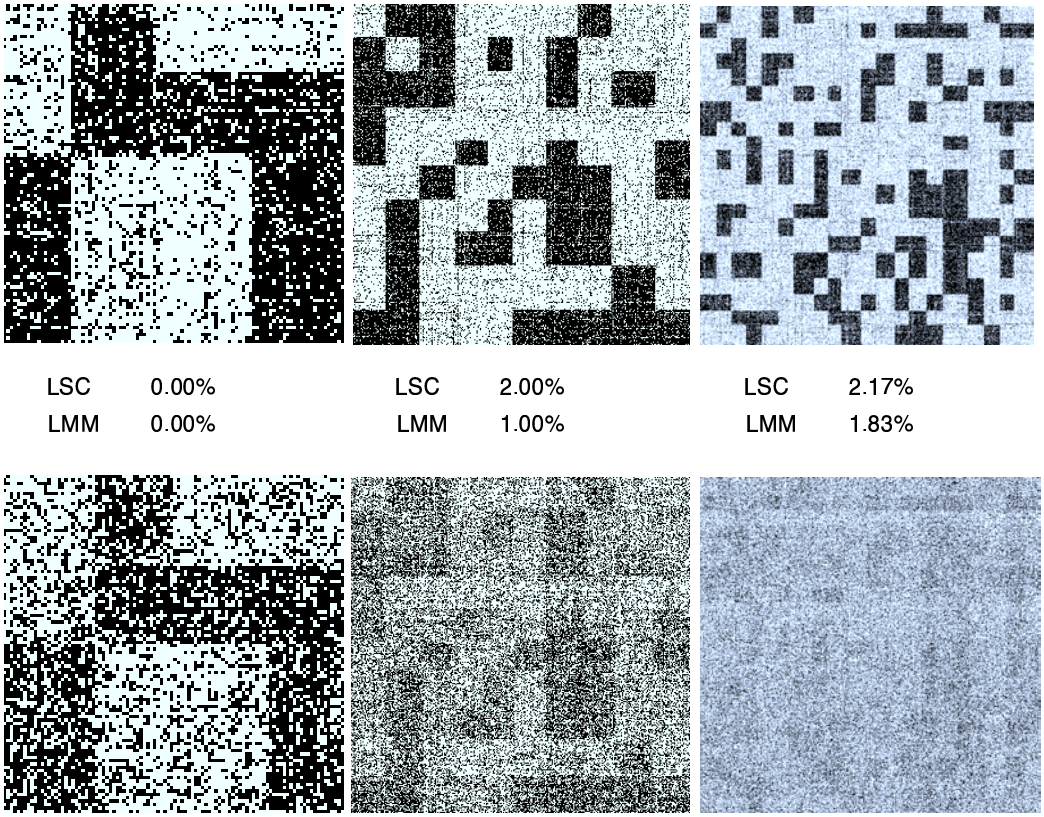

Figure 2: Error rates on simulated protein-protein interaction networks, the lower the better, for spectral clustering with local scaling (LSC) versus latent mixed-membership (LMM). From left to right: the adjacency matrices contain 100, 300 and 600 proteins and 4, 10 and 20 latent functional groups, respectively. From top to bottom: the matrices were generated using Dirichlet parameter $\alpha = 0.05$(stringent membership)$, 0.25$(more diffused membership), respectively. The proteins are re-ordered to make explicit the structure of the group interactions. The number of proteins per cluster averages 30 over all matrices. The Bernoulli probabilities in η are either 0.9 or 0.1. Random guesses about single-membership of proteins to clusters correspond to error rates of $0.75, 0.9$ and 0.95, respectively.

proximate MLE is:

$$\eta^*_{g,h} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \phi_{i,j,1,g}\, \phi_{i,j,2,h}\, r_{i,j}}{\sum_{i=1}^{N}\sum_{j=1}^{N} \phi_{i,j,1,g}\, \phi_{i,j,2,h}},$$

for every index pair $(g,h) \in [1,K] \times [1,K]$.

We also smooth the probabilities of interactions between any member of group a and any member of group b, that is $\eta_{a,b}$, by assuming $\eta_{a,b} \sim Beta\,(\beta_1, \beta_2)$ for each pair of groups $(a,b) \in [1,K] \times [1,K]$. Variational inference is modified appropriately.

4. EXAMPLES AND EXPERIMENTS

We first tested our model in a controlled setting. We simulated non-contrived adjacency matrices mimicking protein-protein interactions with 100 proteins and four groups, 300 proteins and 10 groups, and 600 proteins and 20 groups. In our experiment, the signal-to-noise ratio is decreasing with the size of the problem, for a fixed Dirichlet parameter $\alpha < 1$. [2] The data are display in Figure 4, where the S/N ratio is roughly 0.5, 0.4 and 0.3 for the both the top and bottom rows, from left to right.

In Figure 4 we compare our model to spectral clustering with local scaling [25] that is particularly suited for recovering the structure of the interactions in the case when proteins take part in a single function. Note that spectral clustering (or normalized cuts) minimizes the total transition probability due to 1-step random walk of objects between clusters. Each object is assumed to have a unique cluster membership. Our model, however, is more flexible. It allows object to have different cluster membership while interacting with different objects. The simulations with the

[2]That is, a fixed $\alpha < 1$ leads to a number of active functions for each protein that increases linearly with the total number of latent functions, but the number of interactions sampled among functional groups decreases with the square of the total number of latent function, and causes an overall decrease of the informative part of the observed matrix $\boldsymbol{r}$.

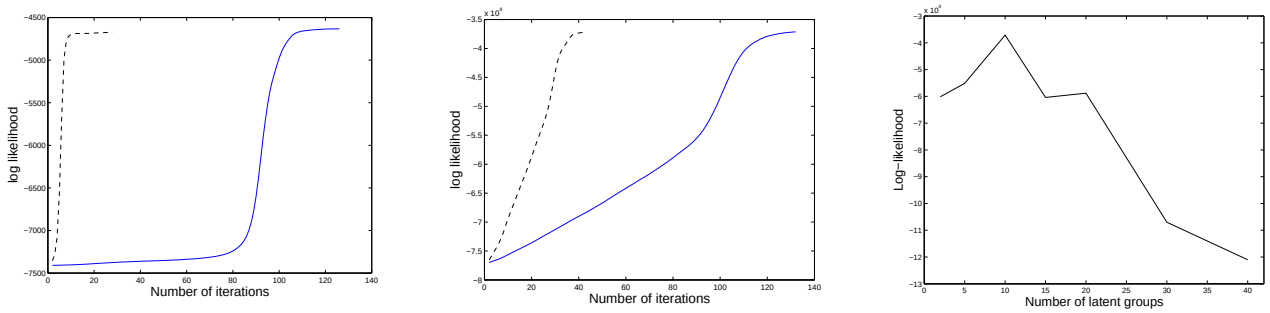

Figure 3: In the first two panels (left and center) we compare the running time of the naïve variational inference (solid line) against the running time of our enhanced (nested) variational inference algorithm (dashed line). The rightmost panel shows how the log likelihood is indicative of the latent number of functions; in the example shown the peak corresponds to the correct number of functions.

Dirichlet parameter $\alpha = 0.05$ are meant to provide mostly unique membership; spectral clustering performs well and our model has a slightly better performance. As proteins participate to more functions, that is, α increases to 0.25 in our simulations, spectral clustering is not an adequate solution anymore. Our model, on the other hand, is able to recover the mixed membership to a large degree, and performs better than spectral clustering.

In a more general formulation of our model we accommodate a collection of observations, e.g., protein-protein interaction patterns measured by different laboratories and under possible different conditions, or daily summaries of email exchanges. We used this general model to understand how the model takes advantage of the information available. Empirical results show that it is better to have a larger adjacency matrix rather than having a collection of small matrices, in order to overcome a fixed signal-to-noise ratio.

In Figure 3 compare the running time of our enhanced variational-EM algorithm to the naïve implementation. Our algorithm is more efficient in terms of space and converges faster. Further, it can be parallelized given that the updates for each interaction (i, j) are independent of one another.

4.1 Case Study: Protein-Protein Interactions

Protein-protein interactions (PPI) form the physical basis for formation of complexes and pathways which carry out different biological processes. A number of high-throughput experimental approaches have been applied to determine the set of interacting proteins on a proteome-wide scale in yeast. These include the two-hybrid (Y2H) screens and mass spectrometry methods. For example, mass spectrometry is used to identify components of protein complexes [9, 10]. High-throughput methods, though, may miss complexes that are not present under the given conditions, for example, tagging may disturb complex formation and weakly associated components may dissociate and escape detection.

The MIPS [16] database was created in 1998 based on evidence derived from a variety of experimental techniques and does not include information from high-throughput datasets. It contains about 8000 protein complex associations in yeast. We analyze a subset of this collection containing 871 proteins, the interactions amongst which were hand-curated. In Table 1 we summarize the main functions of the protein in our sub-collection, where we retained the function names in [14] where possible. Note that, since most proteins participate in more than one function, Table 1 contains more counts (2119) than proteins (871), for an average of ≈ 2.4 functions per protein. Note that the relative importance of each functional category in our sub-collection, in terms of the number of proteins involved, is different from the relative importance of the functional categories over the entire MIPS collection, as reported in [14].

Table 1: Functional Categories. In the table we report the functions proteins in the MIPS collection participate in. Most proteins participate in more than one function (≈ 2.4 on average) and, in the table, we added one count for each function each protein participates in.

#	Category	Size
1	Metabolism	125
2	Energy	56
3	Cell cycle & DNA processing	162
4	Transcription (tRNA)	258
5	Protein synthesis	220
6	Protein fate	170
7	Cellular transportation	122
8	Cell rescue, defence & virulence	6
9	Interaction w/ cell. environment	18
10	Cellular regulation	37
11	Cellular other	78
12	Control of cell organization	36
13	Sub-cellular activities	789
14	Protein regulators	1
15	Transport facilitation	41

4.1.1 *Recovering the Ground Truth*

Our data consists of 871 proteins participating in 255 functions. The functions are organized into a hierarchy, and the 15 functions in Table 1 are those at the top level of the hierarchy. In order to recover what we consider are the true mixed-membership vectors $\boldsymbol{\theta}_i$ corresponding to each protein, we simply normalized the number of times each protein participated in any sub-function of one of the 15 primary functions. The Dirichlet parameter α corresponding to the true mixed-membership is ≈ 0.0667. Most of the proteins in our data participate in two to four functions. In Figure 4 we show the true mixed-membership probabilities for 841 pro-

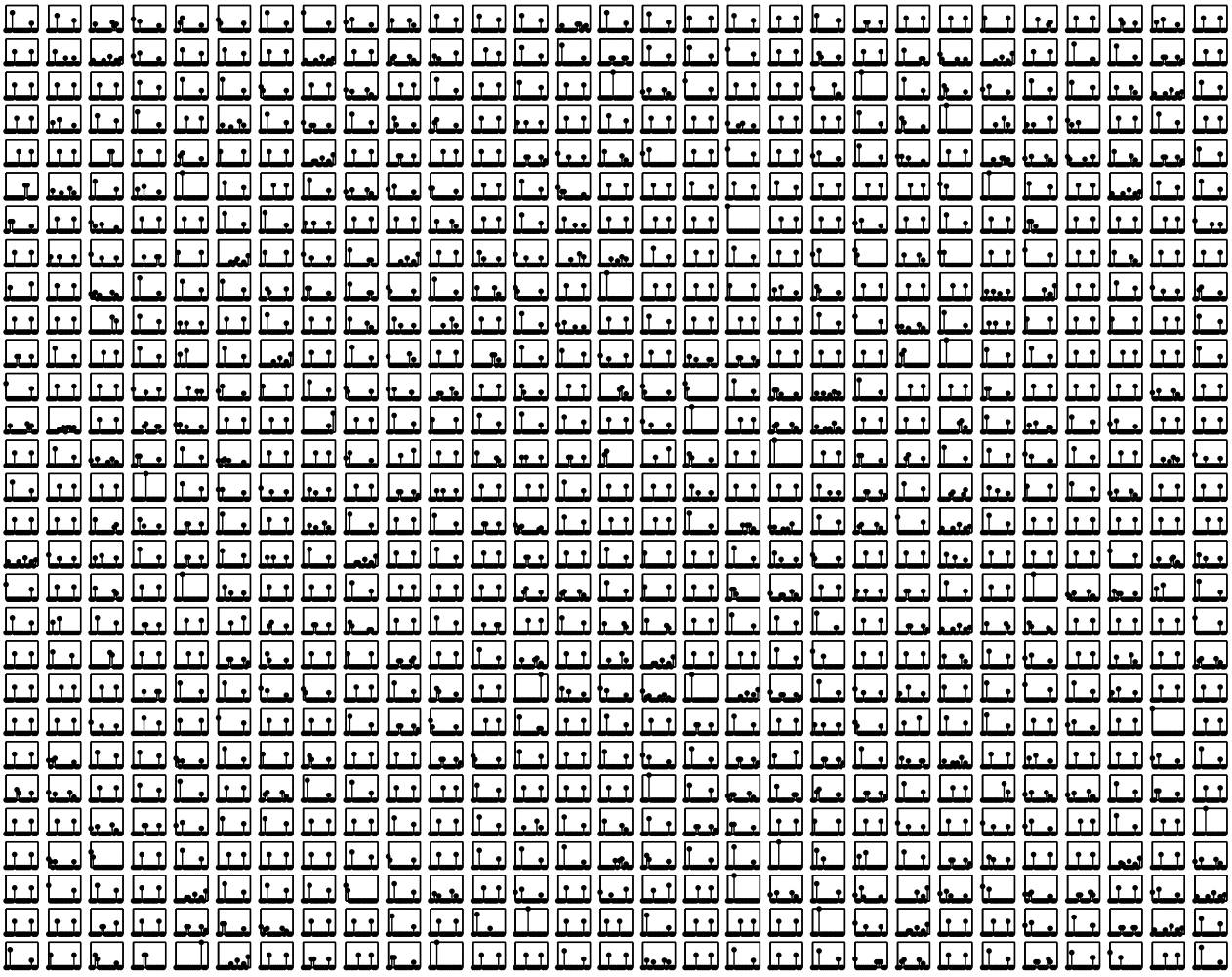

Figure 4: Mixed-membership scores estimated from hand-curated protein-protein interactions: most proteins participate in at least two functions. The figure shows 841 panels arranged in a 29 by 29 grid. Each panel plots the θ_i of the corresponding protein in the MIPS collection. We measure probability on the Y axis and the functional group on the X axis. The functions are numbered from 1 to 15 as in Table 1.

teins.

4.1.2 Evaluating the Performance

In order to evaluate the performance of the competing methods in predicting the (possibly) multiple functional annotations of proteins we devised a very simple measure of accuracy. Briefly, we added the number of functional annotations correctly predicted for each proteins, divided by the total number of functional annotations.

Note that, given their exchangeable nature, the latent functional groups are not identifiable in our model. On the other hand, in order to compute the accuracy above we need to decide which latent cluster correspond to which functional class. We resolved the ambiguity by finding the one permutation that maximized the accuracy on the training data. We then used that permutation in order to compare predicted functional annotations to the ground truth, for all proteins.

In order to compute the accuracy of spectral clustering with local scaling, we implemented softened a soft version of it; we used the cluster predictions and the relative distances between proteins and the centroids of the clusters to obtain normalized scores (probabilities) of membership of each protein to each cluster. These mixed-membership scores enabled us to compute the accuracy measure.

4.1.3 Testing Functional Interaction Hypothesis

In order to compute the accuracy measure proposed above we need to decide which functional annotations are significantly different from zero. We used a simple statistical test to find significant functional associations: we pool all mixed-membership probabilities $\boldsymbol{\theta}_i$ together and we select the 10% most likely protein-function pairs, (i, θ_{ij}), as being significant. That is, under the assumption that most protein-function pairs are not significant, we choose the 10% most likely functional annotations to be the significant ones.

On a different note, the latent mixed-membership model is a useful tool to explore hypothesis about the nexus between latent protein interaction patterns and the functions they are able to express. For example, it is reasonable to assume

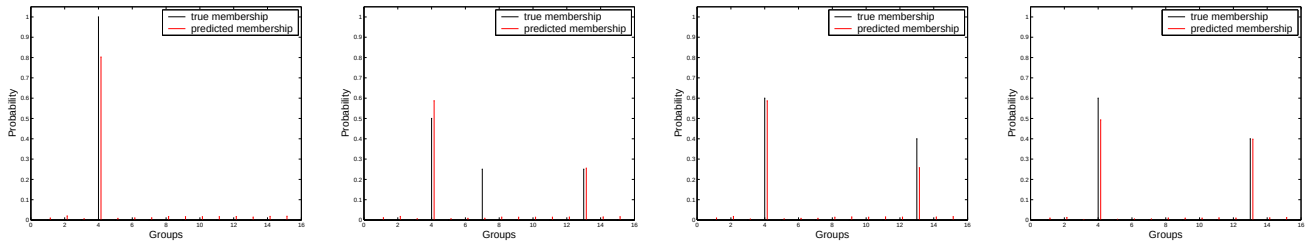

Figure 5: Predicted (red) versus true (black) mixed-membership probabilities for four example proteins.

that proteins that share a common functional annotation tend to interact with one another more often than with proteins with no functional annotations in common. In order to test this hypothesis we can fix the function interaction matrix $\boldsymbol{\eta}$ to be the identity matrix. This leads to accuracies of 43.49% for the latent mixed-membership model and of 41.67% for spectral clustering. That is, we were able to recover 504 and 483 protein-function pairs correctly out of 1159 significant, true functional annotations, for the latent mixed-membership model and (softened) spectral clustering with local scaling respectively.

4.1.4 Unsupervised Learning Experiment

In order to test the behavior of our model on real data in a situation where no information about PPI is available, we tried to recover the mixed-membership probabilities $\boldsymbol{\theta}_i$, and the function interaction matrix $\boldsymbol{\eta}$ corresponding to the hand-curated data, in a completely unsupervised fashion. We were able to recover 502 functional annotations out of 1159, which corresponds to an accuracy of 43.31%, better than spectral clustering at 41.67%.

The number of correctly identified functional annotations is comparable to the number obtained in the previous paragraph (43.49, corresponding to 504 correct annotations) under the assumption that proteins that share a common function are more likely to interact than those which do not. There is also a difference in the estimated interaction patterns between paris of latent groups, $\boldsymbol{\eta}$, which is not diagonal but rather has few positive entries arranged around two positive elements of the diagonal.

4.1.5 PPI Prediction Experiment

It is reasonable to assume that a collection of PPI may inform us on the functions protein are able to express. [5]

In order to get a feel for the prediction error associated with our model, we split the proteins into a training set and a testing set of about the same size. We then slightly modify our model in order to predict the functional mixed-membership probabilities of new proteins, i.e., those in the testing set. In particular, we use available information to learn the function interaction matrix $\boldsymbol{\eta}$, which encodes the interaction patterns between pairs of proteins as they express a corresponding pair of functions. We also consider known the functional annotations of the proteins in the training data in terms of their corresponding mixed membership probabilities $\boldsymbol{\theta}_i$. In order to estimate $\boldsymbol{\eta}$ we considered all protein pairs in the training set, and estimated the strength of the interactions between pairs of expressed functions by composing the corresponding membership probabilities of the proteins involved, under assumption of independence. In the testing phase, we fixed $\boldsymbol{\eta}$, and the $\boldsymbol{\theta}_i$ for the proteins in the training set and fit our model in order to infer the mixed- membership probability vectors of the proteins in the testing set. Alternatives are possible, where the information available is used to calibrate priors for the elements of $\boldsymbol{\eta}$, rather than fixing its values.

We were able to recover 523 functional annotations out of 1159, which for an accuracy of 45.12%. For examples of predicted mixed membership probabilities see Figure 5.

4.1.6 High-Throughput Experimental PPI

In the future we plan to explore PPI generated with high-throughput experimental methods: the tandem-affinity purification (TAP) and high-throughput mass spectrometry (HMS) complex data. [10, 9] We will use all MIPS hand-curated PPI to learn the parameters of our model, in order to provide more reliable (predicted) functional annotations for the proteins in both the TAP and HMS collections. The TAP collection contains 1363 proteins, 469 of which are contained in the MIPS hand-curated collection, whereas the HMS collection contains 1578 proteins, and shares 330 of them with the MIPS hand-curated collection.

5. DISCUSSION AND CONCLUSIONS

We have presented the latent mixed-membership model (LMM) for relational data with stochastic and heterogeneous interactions among objects. In particular, the mixed-membership assumption is very desirable for modeling real data. Given a collection of interaction patterns, our model yields posterior estimation of the multiple group membership of objects, which align closely to real world scenarios (e.g., multi-functionality of proteins). Further, our model estimates interaction probabilities between pairs of latent groups.

In simulations, our model out-performs spectral clustering both in cases when objects have single membership and in cases when objects have mixed-membership. In this latter case, the differential performance of latent mixed-membership model over spectral clustering (with local scaling) is remarkable, since spectral clustering lacks a device for capturing mixed membership. The parameter ρ of LMM enables to recover clusters whose objects are sparsely interconnected, by assigning more weight to the observed edges, i.e., the ones in the observed adjacency matrix $\boldsymbol{r}$. On the contrary, spectral clustering methods assign equal weight to both ones and zeros in the adjacency matrix $\boldsymbol{r}$, so that the classification is driven by the zeros in cases where the number of zeros is overwhelming—this may be a not desirable effect, thus it is important to be able to modulate it, e.g., with ρ.

In the case study we applied our model to the task of predicting the functional annotation of proteins by leveraging protein-protein interaction patterns. We showed how our

model provides a valuable tool to test hypothesis about the nexus between PPI and functionality. We showed how completely unsupervised inference leads to results (in terms of accuracy of the functional annotation of proteins) that are comparable to those of reasonable assumptions about how PPI leads to functionality. We also showed a way to perform cross-validation experiments in this setting, to demonstrate how it is possible to partially learn our model and make use of reliable information (about PPI) in order to infer the functionality of unlabeled proteins. Our results confirm previous findings that information about PPI alone does not lead to accurate functional annotation (in absolute terms) of unlabeled proteins. More information is needed. We plan to integrate high dimensional representation of proteins (static, non-relational) in order to boost the accuracy of functional annotation in future research.

Overall, recovering latent mixed-membership of proteins to clusters that relate to functionality provides a promising approach to learn the generative/mechanistic aspects underlying such data, which can be valuable for seeking deeper insight of the data, as well as for serving as informative priors for future learning tasks.

6. ACKNOWLEDGMENTS

The authors wish to thank Yanjun Qi, at the School of Computer Science of Carnegie Mellon University, for providing the polished version of the MIPS hand-curated data used in the PPI case study.

7. REFERENCES

[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[2] D. Blei, A. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, 2005.

[4] D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *Advances in Neural Information Processing Systems 13*, 2001.

[5] M. H. Deng, K. Zhang, S. Mehta, T. Chen, and F. Z. Sun. Prediction of protein function using protein-protein interaction data. In *IEEE Computer Society Bioinformatics Conference*, 2002.

[6] E. Erosheva and S. E. Fienberg. *Classification—The Ubiquitous Challenge*, chapter Bayesian Mixed Membership Models for Soft Classification, pages 11–26. Springer-Verlag, 2005.

[7] 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.

[8] 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.

[9] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, and et. al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.

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

[11] 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.

[12] P. W. Holland and S. Leinhardt. *Sociological Methodology*, chapter Local structure in social networks, pages 1–45. Jossey-Bass, 1975.

[13] C. Kemp, T. L. Griffiths, and J. B. Tenenbaum. Discovering latent classes in relational data. Technical Report AI Memo 2004-019, MIT, 2004.

[14] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. Kernel-based data fusion and its application to protein function prediction in yeast. In *Proceedings of the Pacific Symposium on Biocomputing*, 2004.

[15] K. G. Manton, M. A. Woodbury, and H. D. Tolley. *Statistical Applications Using Fuzzy Sets*. Wiley, 1994.

[16] H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Guldener, and et. al. Mips: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research*, 32:D41–44, 2004.

[17] T. Minka. Estimating a Dirichlet distribution. Technical report, M.I.T., 2000.

[18] J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

[19] N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. Genetic structure of human populations. *Science*, 298:2381–2385, 2002.

[20] T. A. B. Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 2002.

[21] B. Taskar, M. F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Neural Information Processing Systems 15*, 2003.

[22] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley, 2003.

[23] S. Wasserman and P. Pattison. Logit models and logistic regression for social networks: I. an introduction to markov graphs and p^*. *Psychometrika*, 61:401–425, 1996.

[24] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russel. Distance metric learning with applications to clustering with side information. In *Advances in Neural Information Processing Systems, 15*, 2003.

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