Learning the Network Structure of Heterogeneous Data via Pairwise Exponential Markov Random Fields

Youngsuk Park
Stanford University
youngsuk@stanford.edu

David Hallac
Stanford University
hallac@stanford.edu

Stephen Boyd
Stanford University
boyd@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

Abstract

Markov random fields (MRFs) are a useful tool for modeling relationships present in large and high-dimensional data. Often, this data comes from various sources and can have diverse distributions, for example a combination of numerical, binary, and categorical variables. Here, we define the *pairwise exponential Markov random field* (PE-MRF), an approach capable of modeling exponential family distributions in heterogeneous domains. We develop a scalable method of learning the graphical structure across the variables by solving a regularized approximated maximum likelihood problem. Specifically, we first derive a tractable upper bound on the log-partition function. We then use this upper bound to derive the *group graphical lasso*, a generalization of the classic graphical lasso problem to heterogeneous domains. To solve this problem, we develop a fast algorithm based on the alternating direction method of multipliers (ADMM). We also prove that our estimator is sparsistent, with guaranteed recovery of the true underlying graphical structure, and that it has a polynomially faster runtime than the current state-of-the-art method for learning such distributions. Experiments on synthetic and real-world examples demonstrate that our approach is both efficient and accurate at uncovering the structure of heterogeneous data.

1 Introduction

Markov random fields (MRFs) are a fundamental tool for many applications in machine learning [5, 20]. Often, it is necessary to model MRFs between heterogeneous entities, where the nodes in the graphical model can refer to different types of objects. For example, modeling medical patients may require joint reasoning about the relationship between categorical and continuous variables (i.e., age, gender and medical history events) [1]. Or, when analyzing protein interactions, data about different types of proteins might be captured using different experimental methods [17]. In order to faithfully model such heterogeneous domains using graphical models, nodes of the model must follow different types of distributions (Gaussian, Poisson, Binary, etc.). While the structure of such heterogeneous graphical models is typically learned through observational data, estimating them is challenging for both computational and mathematical reasons, and scalable inference methods have not yet been developed.

In this paper, we propose a class of multivariate exponential family distributions, which we call the *pairwise exponential Markov random field* (PE-MRF). PE-MRFs explicitly reveal the Markov structure across different variables and can cover many common distributions, such as Ising models, Gaussian MRFs, and mixed (heterogeneous) models. Our approach extends previous methods of graphical inference [15, 21, 25] by using a different representation of the joint distribution. This allows for a compact representation of heterogeneous variables, which eventually leads to a much faster structure/parameter learning method ($O(p^3)$, compared to $O(p^4)$, to learn a p node distribution with $O(p^2)$ unknown parameters).

After formally defining the PE-MRF model, we propose a method of estimating its parameters. Because solving the exact maximum likelihood problem is computationally intractable in general high-dimensional settings, we extend an approach known as the approximated maximum likelihood, which previously has

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

only been used for solving Ising models [2, 22]. This approach relies on deriving a tractable upper bound on the (intractable) log-partition function of the PE-MRF. We then show that the estimator can be simplified into solving a convex problem, minimizing the log-determinant divergence [19] plus a group sparsity penalty [10]. We call this the *group graphical lasso* for PE-MRFs, since it turns out to be a generalization of the well-known graphical lasso [9], a popular method of learning Gaussian MRFs. The graphical lasso is just a special case of our method, which is more generally able to reveal the Markov structure in heterogeneous settings. Furthermore, we prove that our estimator is sparsistent [19, 25], meaning that, under some mathematical assumptions, we are asymptotically guaranteed to recover the true underlying Markov structure of the distribution.

By converting the problem into the *group graphical lasso*, we are able to infer the structure in a scalable way. In contrast to the pseudo-likelihood, a commonly-used alternative approach which in the general case requires Newton-type methods [4, 14, 21, 25], we develop an algorithm based on the alternating direction method of multipliers (ADMM) [6], and we solve for closed-form updates for each of the ADMM subproblems. These fast updates speed up the solver and make ADMM more scalable than other methods. Finally, we test our approach's accuracy and scalability on both synthetic and real datasets.

Summary of Contributions. The main contributions of this paper are as follows:

- We propose a pairwise exponential family distribution (PE-MRF), explicitly revealing the Markov structure across heterogeneous variables.
- We formulate an approximated maximum likelihood problem by deriving a tractable upper bound on the log-partition function.
- We develop a scalable ADMM algorithm with closed-form updates.
- We prove that our estimator is sparsistent.

1.1 Related Work on Pairwise Models

The PE-MRF is related to several recently suggested models for inferring Markov random fields. Our primary contribution is that the PE-MRF model presents the most scalable method to date for learning the Markov structure of heterogeneous distributions. Furthermore, there have been proposed approaches which satisfy up to two of our three desired conditions (generality, scalability, and sparsistency), but PE-MRFs are the first to attain all three. We examine several alternative methods in more detail below.

Limited Heterogeneous Distributions. When the distributions at every node are uniparameter, Yang et al. [25] proposed learning the parameters via a pseudo-likelihood approach, solved by Newton's method, and provided sparsistency guarantees. However, their model is unable to generalize to multiparameter settings. Consider the case of a Gaussian MRF. Here, for example, this approach cannot model the problem unless either the mean or the variance is known at every node beforehand. Similarly, Lee et al. [14] used a pseudo-likelihood approach to learn the Markov structure of discrete-Gaussian mixed models. However, their model did not provide any sparsistency guarantee, nor did it generalize to other exponential family distributions. Our PE-MRFs, on the other hand, provide sparsistency and can also solve for distributions with multiparameter and multivariate variables at each node, a much broader class of problems.

Vector-Space MRFs (VS-MRFs). A separate approach, VS-MRF [21], is capable of learning general heterogeneous distributions. In fact, their approach and ours can cover the same classes of pairwise exponential families. However, there is a significant contrast between VS-MRF and PE-MRF in terms of scalability. By modeling the problem differently, we can derive an approximated maximum likelihood, which allows for a very scalable algorithm (but had previously only been used in homogeneous settings [2, 22]). Instead, VS-MRFs rely on the pseudo-likelihood. From an algorithmic perspective, to learn a p-node graphical model in k ADMM iterations [6], our algorithm has a runtime of $O(kp^3)$ whereas VS-MRF takes $O(kp^4)$. Note that there are on the order of p^2 unknown parameters. We leverage this speedup in Section 6, running our algorithm on large models, where we empirically discover that VS-MRF is several orders of magnitude slower.

Node-wise Regression. In node-wise regression, each node estimates the associated edge structure of its local neighborhood. For Gaussian MRFs and discrete models, this method has been used to learn the Markov structure in a scalable way [15, 16, 18]. In addition, Banerjee et al. [2] developed a block coordinate descent method for solving such homogeneous models, which can be viewed as iterative node-wise regressions with a penalty. This type of method, however, has only been suggested for a limited class of homogeneous distributions, and not for broader problems in general heterogeneous domains. As such, there is no known method of using node-wise regression to learn heterogeneous distributions, or to guarantee sparsistency in these cases.

2 Problem Setup

Consider a set of n independent multivariate observations $\{x^1, x^2, \ldots, x^n\}$. We assume that these are sampled i.i.d. from an exponential family distribution $p(x;\boldsymbol{\theta})$ represented by a p-node graphical model with natural parameter $\boldsymbol{\theta}$. Here, the p nodes may have heterogeneous domains. For example, some elements may be defined over the set of real numbers, while others may have a finite discrete domain (i.e., $\{1, 2, \ldots, m\}$) or a countably infinite one. We use these samples to estimate the underlying distribution. More specifically we infer a Markov Random Field described by $G = (V, E)$ with $|V| = p$. This structure can be encoded in the exponential family parameter $\boldsymbol{\theta}$.

2.1 Pairwise Exponential Markov Random Fields

We define the *pairwise exponential Markov random field* (PE-MRF), a subclass of the multivariate exponential family distribution that can explicitly reveal the Markov structure across heterogeneous variables.

Here, we denote $\langle A, B\rangle_F = \mathbf{Tr}(AB)$ as the Frobenius inner product between two matrices and $\mathbf{vec}[a_1, \ldots, a_k] = [a_1^T, \ldots, a_k^T]^T$ as the concatenation of a set of vectors.

Definition For a random vector $X = \{X_1, \ldots, X_p\}$ defined over (heterogeneous) domains[1] $\mathcal{X} = \otimes\{\mathcal{X}_r\}_{r=1}^p$, suppose the conditional distribution of each variable X_r given the remaining $p-1$ variables $X_{\setminus r}$, follows a (known) exponential family distribution on the domain $\mathcal{X}_r$. This distribution is specified by an m_r-dimensional node-potential $B_r(X_r)$ and scalar base measure $C_r(X_r)$. Then, a random vector X is defined as a PE-MRF if, for $x = \{x_1, \ldots, x_p\} \in \mathcal{X}$, it follows the joint distribution

$$p(x;\boldsymbol{\theta}) = \exp\{\sum_{r=1}^p \theta_r^T B_r(x_r) + \sum_{s,t=1}^p \left\langle \Theta_{st}, B_t(x_t)B_s(x_s)^T\right\rangle_F + \sum_{r=1}^p C_r(x_r) - A(\boldsymbol{\theta})\}. \quad (1)$$

Here $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_p, \Theta_{11}, \Theta_{12}, \ldots, \Theta_{pp}\}$ is the natural parameter, consisting of node-parameter $\theta_r \in \mathbf{R}^{m_r}$ and edge-parameter $\Theta_{st} \in \mathbf{R}^{m_s \times m_t}$. $A(\theta)$ is the (finite-valued) log-partition function[2] $\log \int_{\mathcal{X}} \exp\big\{\sum_{r=1}^p \theta_r^T B_r(x_r) + \sum_{s,t=1}^p \left\langle \Theta_{st}, B_t(x_t)B_s(x_s)^T\right\rangle_F + \sum_{r=1}^p C_r(x_r)\big\}\nu(dx)$.

Note that the edge-parameters $\{\Theta_{st}\}_{s,t=1}^p$ explicitly reveal the Markov structure because $\Theta_{st} \in \mathbf{R}^{m_s \times m_t}$ is a zero matrix if and only if X_s and X_r are conditionally independent given all other variables [12].

Exponential Family Expression. Distribution (1) can be written as

$$p(x;\boldsymbol{\theta}) = \exp\{\langle\boldsymbol{\theta}, \boldsymbol{B}(x)\rangle + C(x) - A(\boldsymbol{\theta})\}, \quad (2)$$

with base measure $C(x) = \sum_{r=1}^p C_r(x_r)$, sufficient statistic $\boldsymbol{B}(x)$, and inner product $\big\langle\boldsymbol{\theta}, \boldsymbol{B}(x)\big\rangle$ given by

$$\boldsymbol{B}(x) = \big\{\{B_r(x_r)\}_{r=1}^p, \{B_s(x_s)B_t(x_t)^T\}_{s,t=1}^p\big\},$$
$$\big\langle\boldsymbol{\theta}, \boldsymbol{B}(x)\big\rangle = \sum_{r=1}^p \theta_r^T B_r(x_r) + \sum_{s,t=1}^p \left\langle \Theta_{st}, B_t(x_t)B_s(x_s)^T\right\rangle_F.$$

Alternative Quadratic Expression. The model in (1) can also be written in quadratic form, as

$$p(x;\boldsymbol{\theta}) = \exp\{b(x)^T\boldsymbol{\Theta}b(x) + C(x) - A(\boldsymbol{\theta})\},$$

where we introduce the (extended) node-potential vector $b(x) = \mathbf{vec}[1, B_1(x_1), \ldots, B_p(x_p)]$ with base measure $C(x) = \sum_{r=1}^p C_r(x_r) - 1$, where

$$\boldsymbol{\Theta} = \begin{bmatrix} 1 & \theta_1^T/2 & \cdots & \theta_p^T/2 \\ \theta_1/2 & \Theta_{11} & \cdots & \Theta_{1p} \\ \vdots & \vdots & \ddots & \\ \theta_p/2 & \Theta_{p1} & & \Theta_{pp} \end{bmatrix}.$$

Node-Conditional Distribution. The conditional distribution of X_r given $X_{\setminus r}$ follows

$$p(x_r|x_{\setminus r};\boldsymbol{\theta}) \propto \exp\Big\{(\theta_r + \sum_{t\neq r}\Theta_{rt}B_t(x_t))^T B_r(x_r) + \left\langle \Theta_{rr}, B_r(x_r)B_r(x_r)^T\right\rangle_F + C_r(x_r)\Big\}. \quad (3)$$

This is just an exponential family with sufficient statistic $\{B_r(x_r), B_r(x_r)B(x_r)^T\}$ and base measure $C(x_r)$.

2.2 Examples of PE-MRFs

PE-MRFs can model many popular distributions, ranging from homogeneous pairwise models (exponential, Poisson, gamma, etc...) to general mixed ones. From the node-conditional distribution in Equation (3), we can design PE-MRFs on a node-by-node basis, by choosing the desired potential B_r and base measure C_r. Note that, in order to get a valid joint distribution, we must consider domain constraints $\mathcal{D}$ on the parameter $\boldsymbol{\theta}$ to guarantee a finite log-partition function $A(\boldsymbol{\theta})$.

Gaussian MRF (GMRF). PE-MRFs can model GMRFs by setting the node-potential $B_r(x_r) = x_r$,

[1] Here, $\otimes$ refers to the Kronecker product.

[2] ν is a proper measure on $\mathcal{X}$. Refer to [23].

so that the corresponding sufficient statistic at each node becomes $\{x_r, x_r^2\}$. This is a valid joint distribution under the (negative definite) domain constraints $\mathcal{D} = \{(\theta_{node}, \Theta_{edge}) \mid \Theta_{edge} \prec 0\}$. For a zero mean GMRF, we put additional the restrictions $\theta_{node} = \mathbf{0}$. If a variable X_r has known variance σ_r^2, then we can additionally assign $\Theta_{rr} = -\frac{1}{2\sigma_r^2}$.

Ising and Discrete Models. PE-MRFs can also model a discrete distribution with domain $\mathcal{X}_r = \{0, 1, \dots, m_r\}$, by choosing the node-potential $B_r(x_r) = [\mathbb{I}(x_r = 1), \dots, \mathbb{I}(x_r, = m_r)]^T$. Moreover, restricting $\theta_r = \mathbf{0}$ and $\Theta_{rr} = \mathbf{diag}([\Theta_{rr:11}, \dots, \Theta_{rr:m_r m_r}])$ gives the minimal representation of a discrete model.

Mixed Models. Likewise, by choosing suitable node potentials, we can design any associated mixed model using a PE-MRF. The exponential family distributions that PE-MRFs can cover include, but are not limited to, Poisson, quadratic Poisson [26] (which can capture both positive and negative correlations), lognormal, gamma, Dirichlet, and any combination thereof, under proper domain constraints.

3 Learning the Structure: Approximate Maximum Likelihood Approach

In order to learn the Markov structure of a PE-MRF distribution, we formulate the following (negative) maximum likelihood problem with regularization,

$$\underset{\theta}{\text{minimize}} - l(\boldsymbol{\theta}) + R_\lambda(\boldsymbol{\theta}). \tag{4}$$

Here, $l(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \hat{\boldsymbol{\mu}} \rangle - A(\boldsymbol{\theta})$ is the log likelihood of $\boldsymbol{\theta}$ (up to scale and constant), where $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^n \boldsymbol{B}(x^k)$ is the empirical *mean parameter*, or averaged sufficient statistic $B(x^k)$ over the samples $\{x^k\}_{k=1}^n$,

$$\hat{\boldsymbol{\mu}} = \{\{\frac{1}{n}\sum_{k=1}^n B_r(x_r^k)\}_{r=1}^p, \{\frac{1}{n}\sum_{k=1}^n B(x_s^k)B(x_t^k)^T\}_{s,t=1}^p\}.$$

Regularization. $R_\lambda(\boldsymbol{\theta})$ is a regularization function with tuning parameter λ that encourages structural sparsity. We use the ℓ_1/ℓ_2 group lasso penalty [11],

$$R_\lambda(\boldsymbol{\theta}) = \lambda \sum_{s \neq t} w_{st} \left\| \Theta_{st} \right\|_F, \tag{5}$$

where $\left\|\cdot\right\|_F$ is the Frobenius norm, i.e., $\left\|A\right\|_F = \sqrt{\sum_{i,j=1}^{m_i,m_j} a_{ij}}$. This encourages the st-th block, for every $s \neq t$, to be a zero matrix. Note that if the Θ_{st}'s are all scalar parameters, then this becomes a standard lasso penalty. Here, $\{w_{st}\}_{s,t=1}^p$ is a set of scalar values typically depending on the size and variance of an associated $B_s(X_s)B_t(X_t)^T$, in order to balance the weights on $\{\|\Theta_{st}\|_F\}_{s,t=1}^p$ [14].

3.1 Approximated Maximum Likelihood

When the exponential family has a tractable $A(\boldsymbol{\theta})$, for example with a Gaussian MRF, we can attempt to exactly solve the maximum likelihood in Problem (4). However, in the general case, $A(\boldsymbol{\theta})$ involves a high-dimensional integral and is typically intractable to compute. In order to overcome this, we use an approximated maximum likelihood approach [2], where we replace $A(\boldsymbol{\theta})$ in Problem (4) with a tractable (convex) upper bound $U(\boldsymbol{\theta})$,

$$\underset{\boldsymbol{\theta}}{\text{minimize}} - \langle \boldsymbol{\theta}, \hat{\boldsymbol{\mu}} \rangle + U(\boldsymbol{\theta}) + R_\lambda(\boldsymbol{\theta}). \tag{6}$$

3.1.1 Upper Bound on Log-Partition Function

In this section, we use the notion of *mean parameter* and variational analysis [23]. These allow us to have an alternative expression for $A(\boldsymbol{\theta})$, which we use to eventually derive the upper bound $U(\boldsymbol{\theta})$.

Notation. Recall that the natural parameter $\boldsymbol{\theta}$ and sufficient statistic $\boldsymbol{B}(x)$ of PE-MRFs are defined over the domain $\mathcal{Y} = \mathbf{R}^{m_1} \times \cdots \mathbf{R}^{m_p} \times \mathbf{R}^{m_1 m_1} \times \mathbf{R}^{m_1 m_2} \times \cdots \mathbf{R}^{m_p m_p}$. Here, we define the inner product over the domain as $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{r=1}^p a_r^T b_r + \sum_{s,t=1}^p \langle A_{st}, B_{st} \rangle_F$ for elements $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{Y}$, which is consistent with the definition in (2).

We define a set of realizable expected sufficient statistics $\boldsymbol{B}(\cdot)$ over all valid distributions, the *mean parameters* [23], as

$$\mathcal{M}(\boldsymbol{B}) = \Big\{ \boldsymbol{\mu} := \{\{\mu_r\}_{r=1}^p, \{\mu_{st}\}_{s,t=1}^p\} \in \mathcal{Y} \mid \exists p(\cdot) \in \Delta_d \\ \text{s.t. } \mathbf{E}[B_r(X_r)] = \mu_r, \mathbf{E}[B_s(x_s)B_t(x_t)^T] = \mu_{st} \Big\},$$

where $d = \sum_{r=1}^p m_r$ is the sum of dimensions of the node potentials and Δ_d is a probability simplex in $\mathbf{R}^d$.

Let $\nu \in \mathbf{R}$ be a fixed scalar. We define a map $M_\nu : \mathcal{Y} \to \mathbf{R}^{(d+1)\times(d+1)}$, for $\boldsymbol{\mu} =\in \mathcal{Y}$, as

$$M_\nu[\boldsymbol{\mu}] = \begin{bmatrix} \nu & \mu_1^T & \mu_2^T & \cdots & \mu_p^T \\ \mu_1 & \mu_{11} & \mu_{12} & \cdots & \mu_{1p} \\ \mu_2 & \mu_{21} & \mu_{22} & & \mu_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_p & \mu_{p1} & \mu_{2p} & \cdots & \mu_{pp} \end{bmatrix}.$$

With these notations, we now derive an upper bound. Assume that a PE-MRF is regular [23] and the conditional distribution on each node comes from a known exponential family. Let each node r have the distance $c_r = \inf_{a\neq b \in \mathcal{X}_r} \|B_r(a) - B_r(b)\|_\infty$ in the domain of its sufficient statistic $B_r(\mathcal{X}_r)$. We assume $c_r = 0$ for continuous nodes and that discrete nodes $\mathcal{I}_D$ are separable with respect to sufficient statistic, meaning $c_r > 0$.

Theorem 3.1 *For a PE-MRF, the log partition function $A(\theta)$ has the following upper bound,*

$$A(\theta) \le \max_{\boldsymbol{\mu} \in \mathcal{M}(\boldsymbol{B})} \left\{ \langle \boldsymbol{\mu}, \boldsymbol{\theta} \rangle + \frac{1}{2} \log\det \Big(M_1[\boldsymbol{\mu}] + D \Big) \right\} + f_1,$$

where $D = \mathbf{diag}([0, l_1, \dots, l_p])$, $l_r = \frac{c_r^3}{12}\overbrace{[1, \dots, 1]}^{m_r}$, and $f_1 = \frac{d}{2}\log(2\pi e) - \sum_{r \in \mathcal{I}_D} (m_r \log c_r)$.

The proof follows from Wainwright et al. [22], where the problem is alternatively represented as the *Shannon entropy* $H(X)$ over the mean parameter. In particular, we attain the upper bound from the relationship between the entropy $H(X)$ and the entropy of node-potentials $H(\{B_r(X_r)\}_{r=1}^p)$, in addition to a different choice of $\{l_r\}_{r=1}^p$ for heterogeneous domains.

By taking the relaxation of the dual, we can convert the high-dimensional problem from Theorem 3.1 into the following tractable form.

Corollary 3.2 *The log partition function $A(\theta)$ has the following upper bound*

$$\begin{aligned} A(\boldsymbol{\theta}) \le \frac{1}{2} \min_{\nu \in \mathbf{R}} & \left\{ -\frac{1}{2} \log\det\big(-M_{1+\nu}[\boldsymbol{\theta}']\big) - \nu \right\} \\ & - \frac{1}{2} \left\langle M_1[\boldsymbol{\theta}'], D \right\rangle_F + f_2, \end{aligned}$$

where $\boldsymbol{\theta}' = \{\theta_1/2, \dots, \theta_p/2, \Theta_{11}, \Theta_{12}, \dots, \Theta_{pp}\} \in \mathcal{Y}$, and $f_2 = \frac{3}{2} + d + \frac{d}{2}\log(2\pi e) - \sum_{r \in \mathcal{I}_D} (m_r \log c_r)$.

3.1.2 Approximated Maximum Likelihood: Graphical Group Lasso

By plugging the upper bound from Corollary 3.2 into the approximated maximum likelihood from Equation (6), we attain the following optimization problem.

Theorem 3.3 *For a PE-MRF, the approximated negative maximum log-likelihood problem is equivalent to*

$$\min_{\boldsymbol{\Theta} \in \mathbf{S}_{++}^{d+1}} \left\{ \langle \boldsymbol{\Theta}, M_1[\hat{\boldsymbol{\mu}}] + D \rangle_F - \log\det \boldsymbol{\Theta} + R_\lambda(\boldsymbol{\Theta}) \right\}, \quad (7)$$

with parameter matrix $\boldsymbol{\Theta} = -M_{\nu_{\boldsymbol{\theta}}}[\boldsymbol{\theta}']$, where $\nu_{\boldsymbol{\theta}} \in \mathbf{R}$ is a dummy parameter, and $R_\lambda(\boldsymbol{\Theta}) = R_\lambda(\boldsymbol{\theta})$.

Note that there may be additional parameter constraints for some PE-MRF distributions, which can easily be incorporated into (7). We can view this problem as a ℓ_1/ℓ_2 regularized log-determinant divergence with respect to the empirical average of sufficient statistics, defined by the Bregman divergence corresponding to the log-determinant function[3] [19]. We call problem (7) the *group graphical lasso* for a PE-MRF, since it is an extension of the classic graphical lasso problem [9, 15, 19] to heterogeneous settings.

3.2 Gaussianization of the Group Graphical Lasso

For a zero-mean Gaussian MRF, we set additional constraints $[\theta_1, \dots, \theta_p]^T = \mathbf{0}$. In fact, in this case our problem becomes equivalent to the graphical lasso. In the general case, however, the naive graphical lasso optimizes with respect to the empirical covariance matrix, whereas our approach optimizes by using the sample average of sufficient statistics of a PE-MRF. Still, the following Corollary shows how the group graphical lasso in (7) can be interpreted as a generalization of the classic graphical lasso model to heterogeneous domains.

Corollary 3.4 *The group graphical lasso (7) is equivalent to the (exact) maximum likelihood problem with a group lasso penalty for a GMRF $Z \sim \mathcal{N}(\mu^*, \Sigma^*)$,*

$$\begin{aligned} \Sigma^* &= \mathbf{Cov}[b_{node}(X)] + \mathbf{diag}([l_1, \dots, l_p]), \\ \mu^* &= [\Sigma^*]^{-1}\, \mathbf{E}[b_{node}(X)], \end{aligned}$$

where $b_{node}(X) = \mathbf{vec}[B_1(X_1), \dots, B_p(X_p)]$.

The proof is immediate from the fact that the (log) likelihood function of $\mathcal{N}(\mu^*, \Sigma^*)$ can be expressed as $g(\boldsymbol{\Theta}) = \langle \boldsymbol{\Theta}, M_1[\mathbf{E}[\boldsymbol{B}(X)]] + D \rangle - \log\det \boldsymbol{\Theta}$ by using the Schur complement.

Therefore, Corollary 3.4 implies that the estimator of the group graphical lasso for a PE-MRF is equivalent to solving a distinct graphical lasso problem for $b_{node}(X)$ (instead of X), where the lasso regularization term is replaced by a group lasso penalty. Here, we treat $b_{node}(X)$ as following a Gaussian distribution with specified mean and covariance.

4 Optimization Algorithm

Here, we propose an algorithm to solve the group graphical lasso (7) for PE-MRFs. Our approach is based on the alternating direction method of multipliers (ADMM) [6]. To solve, we introduce a consensus variable $\boldsymbol{Z} = \boldsymbol{\Theta}$ and rewrite (7) as its equivalent problem, for $A = \big(M_1[\hat{\boldsymbol{\mu}}] + D\big)$

$$\min_{\boldsymbol{Z} = \boldsymbol{\Theta}, \boldsymbol{\Theta} \in \mathbf{S}_{++}^{d+1}} \langle \boldsymbol{\Theta}, A \rangle_F - \log\det \boldsymbol{\Theta} + \lambda_n \sum_{i \neq j} w_{ij} \|Z_{ij}\|_F\,.$$

[3] $D_{\log\det(\cdot)}(\boldsymbol{\Theta} || \bar{\boldsymbol{B}}) = -\log\det \boldsymbol{\Theta} + \log\det \bar{\boldsymbol{B}} + \left\langle \bar{\boldsymbol{B}}^{-1}, (\boldsymbol{\Theta} - \bar{B}) \right\rangle_F$

ADMM solves the corresponding augmented Lagrangian in an iterative manner with respect to the primal variable $\mathbf{\Theta}$, the consensus variable $\boldsymbol{Z}$, and the (scaled) dual variable $\boldsymbol{U}$. For iteration $k+1$, closed-form updates for each of three subproblems are provided below. ADMM is guaranteed to converge to the global optimal for convex problems, with a standard (primal and dual residual) stopping criterion [6].

$\mathbf{\Theta}$ Update. The $\mathbf{\Theta}$ update is

$$\mathbf{\Theta}^{k+1} := \frac{1}{2\eta} Q\big(\Lambda + \sqrt{\Lambda^2 + 4\eta I}\big)Q^T,$$

where $\eta = \frac{\rho}{n}$ and $Q\Lambda Q^T$ is the eigendecomposition of $\eta(\boldsymbol{Z}^k - \boldsymbol{U}^k) - \big(M_1[\hat{\boldsymbol{\mu}}] + D\big)$ [24].

$\boldsymbol{Z}$-Update. The $\boldsymbol{Z}^{k+1}$ update is $M_{\nu_{\boldsymbol{z}}^{k+1}}[\{z_1^{k+1}, \ldots, z_p^{k+1}, \quad Z_{11}^{k+1}, Z_{12}^{k+1}, \ldots, Z_{pp}^{k+1}\}]$ where for $i, j \in \{1, \ldots, p\}$

$$\begin{aligned} &\nu_{\boldsymbol{z}}^{k+1} = \nu_{\theta}^{k+1} + \nu_{\boldsymbol{u}}^{k}, \\ &z_i = \theta_i^{k+1} + u_i^k, \quad Z_{ii} = \Theta_{ii}^{k+1} + U_{ii}^k, \\ &Z_{ij, i\neq j} = \begin{cases} \left(1 - \frac{\eta_{ij}}{\gamma_{ij}}\right)\left(\Theta_{ij}^{k+1} + U_{ij}^k\right) & \gamma_{ij} \geq \eta_{ij} \\ 0 & \text{otherwise} \end{cases}, \end{aligned}$$

with $\eta_{ij} = \frac{\lambda_n w_{ij}}{\rho}$ and $\gamma_{ij} = \left\|\Theta_{ij}^{k+1} + U_{ij}^k\right\|_F$ [6].

$\boldsymbol{U}$-Update. The $\boldsymbol{U}$ update is

$$\boldsymbol{U}^{k+1} := \boldsymbol{U}^k + \mathbf{\Theta}^{k+1} - \boldsymbol{Z}^{k+1}.$$

Algorithmic Complexity. Note that the eigendecomposition of the $\mathbf{\Theta}$ update is the main computational task in our algorithm, with a runtime of $O(p^3)$. For k ADMM iterations, the computational cost is $O(kp^3)$, which is very efficient considering the fact that the total number of parameters to estimate is $O(p^2)$. On the other hand, the pseudo-likelihood approach requiring Newton-type methods [14, 21] needs to compute $O(p^3)$ operations every ADMM iteration at each of the p nodes, requiring $O(kp^4)$ in total. We will see in Section 6 how this leads to a significant difference in scalability on large problems.

5 Sparsistency

In this section, we present the conditions under which we are guaranteed to recover the underlying graphical structure embedded in the true parameter matrix $\mathbf{\Theta}^{true}$. From Corollary 3.4, recall that the solution of the group graphical lasso (7) is equivalent to the estimator of a (distinct) Gaussian MRF regularized by a group lasso penalty. This implies that our estimate $\hat{\mathbf{\Theta}}$ of (7) may have a different value from $\mathbf{\Theta}^{true}$, unless the node potential $b_{node}(X)$ follows a proper normal distribution, which happens for example in GMRFs and lognormal MRFs. Nonetheless, under some mathematical assumptions for general PE-MRFs, we can demonstrate that the estimator of group graphical lasso (7) is sufficient to recover the underlying Markov structure represented by $E(\mathbf{\Theta}^{true}) = \{(s,t) \mid \left\|\Theta_{ij}^{true}\right\|_2 > 0 \text{ for } 1 \leq s \neq t \leq p\}$.

Notation. We introduce several norms. We let $\|M\|_\infty = \max_{i,j}|M_{ij}|$. For a group of vectors $\{g_i\}_{i=1}^k$ and its concatenation $g = \mathbf{vec}[g_1, \ldots, g_k]$, we define a (group) norm $\|g\|_{\infty,2} = \|[\|g_1\|_2, \ldots, \|g_k\|_2]\|_\infty$. Then, this norm induces an operator norm $|||A|||_{\infty,2}$.

Additionally, we introduce some parameters related to the group graphical lasso. For the log likelihood function $g(\mathbf{\Theta}) = \langle \mathbf{\Theta}, M_1[\hat{\boldsymbol{\mu}}] + D\rangle - \log\det\mathbf{\Theta}$ in (7), we define its (asymptotic) estimator as $\mathbf{\Theta}^* = [\mathbf{E}[M_1[\boldsymbol{B}(X)]] + D]^{-1}$. The corresponding gradient and Hessian are denoted as $\Sigma^* := \nabla g(\mathbf{\Theta}^*) = (\mathbf{\Theta}^*)^{-1}$, $\Gamma^* = \nabla^2 g(\mathbf{\Theta}^*) := (\mathbf{\Theta}^*)^{-1} \otimes (\mathbf{\Theta}^*)^{-1}$ respectively [7]. Here, Σ^* and Γ^* are not necessarily the covariance and Fischer information.

We denote S as the edge set of $\mathbf{\Theta}^{true}$ (including all self-edges), as S^c as its complement. Then, $\Gamma^*_{SS^c} \in \mathbf{R}^{|S|\times(p^2-|S|)}$ is defined as its the submatrix indexed by S and S^c. We define $\{\kappa\} = \{\kappa_{\Sigma^*}, \kappa_{\Gamma^*}, \kappa_{\mathbf{Cov}[\boldsymbol{B}]}, \kappa_{\mathbf{Cov}_{\min}}\}$ with $\kappa_{\Sigma^*} := |||\Sigma^*|||_{\infty,2}$, $\kappa_{\Gamma^*} = |||\Gamma^*|||_{\infty,2}$, $\kappa_{\mathbf{Cov}[\boldsymbol{B}]} = \|\mathbf{Cov}[\boldsymbol{B}(X)]\|_\infty$, and $\kappa_{\mathbf{Cov}_{\min}}$ as the minimum value among non-zero elements of $(\mathbf{Cov}[b_{node}(X)])^{-1}$. We denote $m_{max} = \max_r m_r$, $w_{max} = \max_{s,t} w_{st}$, and $w_{min} = \min_{s,t} w_{st}$.

Assumptions. We make three assumptions

1. The PE-MRF has an underlying graphical structure with singleton separator sets with the maximum degree d.
2. Incoherence condition: $\left|\left|\left|\Gamma_{S^cS}\left(\Gamma_{SS}\right)^{-1}\right|\right|\right|_{\infty,2} \leq \frac{w_{min}}{w_{max}}(1-\alpha)$ for some $\alpha \in (0,1]$.
3. Boundedness condition: $\mathbf{E}[\boldsymbol{B}(X)]$ and $\mathbf{Cov}[\boldsymbol{B}(X)]$ are bounded.

Intuitively, the incoherence condition limits the correlation between edges variables and and non-edge variables; see Loh et al. [15] and Ravikumar et al. [19].

Lemma 5.1 *Suppose the boundedness condition holds. Then, for $\delta_n \geq 2m_{max}\sqrt{\kappa_{\mathbf{Cov}[B]}}\sqrt{\frac{\log(m_{max}p)}{n}}$, there exists an universal constant $c_1 > 0$ satisfying*

$$\mathbf{Pr}\left[\|M_1[\hat{\boldsymbol{\mu}}] - (M_1[\mathbf{E}[\boldsymbol{B}(X)]])\|_{\infty,2} < \delta_n\right] \geq 1 - e^{-c_1 n}.$$

Theorem 5.2 *Suppose a PE-MRF satisfies all three assumptions. For a regularization parameter* $\lambda_n > \frac{8(w_{max}+w_{min})}{\alpha w_{max} w_{min}} \sqrt{\kappa_{\mathbf{Cov}[B]}} \sqrt{\frac{\log(m_{max}p)}{n}}$, *let* $\hat{\mathbf{\Theta}}$ *be the unique solution of the group graphical lasso (7). If the number of samples is given by* $n > c_2(\{\kappa\}, \{w_{st}\}, m_{max}, d, \alpha) \log(m_{max}p)$, *then the following two statements about* $\hat{\mathbf{\Theta}}$ *hold with probability at least* $1 - e^{-c_1 n}$*:*

1. $\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|_{\infty,2} < 2\kappa_{\Sigma^*}\left(\frac{w_{max}w_{max}\alpha}{4(w_{max}+w_{min})} + w_{max}\right)\lambda_n$, *where* $\mathbf{\Theta}^* = (M_1[\mathbf{E}[B(X)]] + D)^{-1}$.
2. *The recovered edge* $E(\hat{\mathbf{\Theta}}) = \{(s,t) \mid \left\|\hat{\Theta}_{ij}\right\|_2 > 2\kappa_{\Sigma^*}\left(\frac{w_{max}w_{max}\alpha}{4(w_{max}+w_{min})} + w_{max}\right)\lambda_n\}$ *becomes the same as the real edge set* $E(\mathbf{\Theta}^{true})$.

Here $c_2(\{\kappa\}, \{w_{st}\}, m_{max}, d, \alpha) = 16\kappa_{\Gamma^*}^2(1+\frac{4w_{max}+w_{min}}{w_{min}\alpha})^2 \kappa_{\mathbf{Cov}[B]} \max\{9\kappa_{\Sigma^*}^2 d^2 m_{max}, 9\kappa_{\Sigma^*}^6 \kappa_{\Gamma^*}^2 d^2 m_{max}^2, \kappa_{\mathbf{Cov}_{min}}^2/4\}$.

The first statement about an error bound can be derived from Lemma 5.1 and the primal-dual witness approach [19], albeit with different inequalities due to the group lasso penalty. The second is based on the relationship between the graphical structure with singleton separator sets and the generalized inverse covariance matrix [15], with an additional constraint that the error is less than the threshold $\kappa_{Cov_{min}}/2$ for a large n.

From Theorem 5.2, for any PE-MRF satisfying our three assumptions, we are asymptotically guaranteed to recover the true underlying Markov structure. Note that the first two assumptions are likely to hold for sparse graphs with well-balanced weights $\{w_{st}\}$, and the last one holds for PE-MRFs consisting of known exponential family distributions at each node. Moreover, even if all parts of a graph are not separated by singleton sets, we can still recover the graphical structure of the sub-parts that are separated [15].

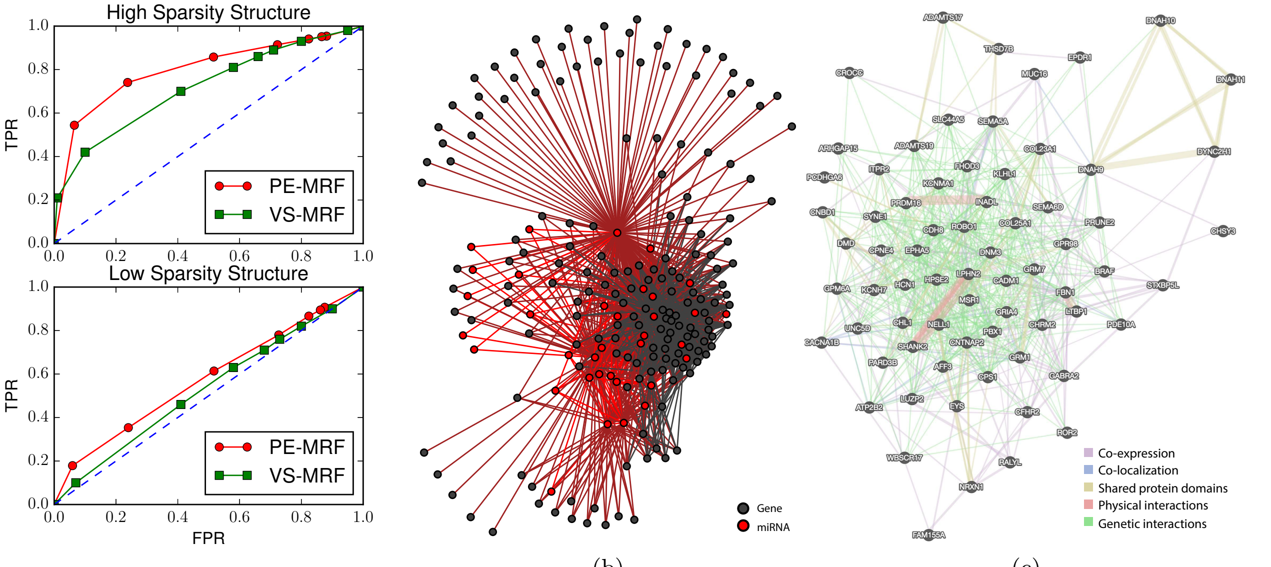
6 Experiments

Synthetic Data. We analyze performance[4] on a heterogeneous synthetic network containing 32 nodes: eight Bernoulli, eight gamma, eight Gaussian, and eight Dirichlet (k=3). We consider two cases: sparse (10% of all potential edges exist) and dense (50% exist). For both cases, we run 30 independent trials, with random edge weights, each time taking 1000 samples from the distribution via Gibbs sampling. We compare with VS-MRF [21], which is the only other approach that can explicitly model such a diverse distribution. In our algorithm, we choose $w_{st} = \sqrt{m_s \times m_t}$ and normalize the raw data so that the rows are numerically well-balanced. We plot the ROC curves for edge recovery percentage (since we care more about capturing the *structure* than the precise weights of each edge) in Figure 1a. As shown, both are better able to recover sparse graphs than dense ones. Overall, the two methods attain similar accuracies, and as shown in Figure 2, our PE-MRF model scales much better to large datasets. In this experiment, we keep the same proportion of Bernoulli, gamma, Gaussian, and Dirichlet nodes, but vary the total problem size. We compare the runtimes of PE-MRF and VS-MRF on the same machine. As shown, PE-MRF is several orders of magnitude faster, opening up new applications that previously could not be modeled in a heterogeneous way. Our PE-MRF solver can learn a 100-node distribution in just **5 minutes**. In contrast, VS-MRF [21] takes **over 31 hours** to converge.

Heterogeneous Genomic Networks. Inferring the structure of genomic regulatory networks from experimental data is a fundamental task in computational biology. Gaussian graphical models are particularly popular and well-suited for this task, as gene expression measurements generated by microarray technology approximately follows a Gaussian distribution. However, new technologies for DNA sequencing can produce data with heterogeneous distributions that violate the Gaussianity assumption. Here, we use PE-MRF to infer a genomic regulatory network. We use Level III public data from The Cancer Genome Atlas (TCGA) [17] for 290 breast cancer patients. The data consists of miRNA sequencing counts mapped back to a reference genome, which follow a Poisson distribution, and microarray gene expression profiles, which are Gaussian. We employ three common steps to process the data: adjust for sequencing depth, remove genes whose mutations are known to have low functional impact on cancer progression, and filter out miRNAs with low variance across samples. In total, the dataset contains expression profiles for 500 genes and 314 miRNAs. Our goal is to infer a heterogeneous MRF with a total of 814 nodes, 500 of which are Gaussian (genes) and 314 of which are Poisson (miRNAs).

We infer a PE-MRF, choosing the regularization parameter λ that minimizes the Akaike information criterion (AIC), and plot the resulting network in Figure 1b. The network contains three types of edges: between two genes, between two miRNAs, and between one of each. All three edge types are interesting and potentially worth exploring further, but we demonstrate here the utility of the gene-gene subnetwork due to the availability of comprehensive gene data.

[4] All code is available at `https://github.com/youngsuk0723/PE-MRF-Code`.



(a) (b) (c)

Figure 1: (a) ROC curves comparing our PE-MRF method with VS-MRF, (b) Inferred multi-modal network of miRNAs and genes, (c) Known interactions from biomedical database of the k-core of the inferred network.

In particular, we consider a k-core of the gene-gene subnetwork, the subnetwork of genes whose inferred degree within the selected subnetwork is at least k (we choose $k = 65$). To validate our model, we use external data and observe how many gold standard edges exist between genes from the tightly connected 65-core. Using GeneMANIA [27], we find the core of the inferred gene subnetwork is well supported by many established interactions as shown in Figure 1c. We infer this network from only the 290 patient samples, yet we notice the abundance of interaction edges in Figure 1c, which match the connectivity of the inferred gene subnetwork. The many genetic interactions (in green) between genes from the core are particularly encouraging for our model, since genetic interactions correspond closely with partial correlations [3, 13], which is precisely what our PE-MRF attempts to model.

7 Conclusion

In this paper, we have proposed a method of learning Markov networks from observational data. Our approach models an underlying distribution as a pairwise exponential Markov random field (PE-MRF), a class of multivariate exponential families that is well-suited for heterogeneous distributions. To estimate the parameters in a scalable way, we derive the approximated maximum likelihood problem and develop an ADMM algorithm with closed-form updates. We then prove sparsistency, or guaranteed recovery of the true Markov structure, of our method. Our promising results, as well as the widespread applications with heterogeneous data sources, lead to many potential extensions of this work. For example, if the the Markov structure changes over time, we could use the timestamped observations to estimate a time-varying network, instead of inferring a single network. Similarly, one could learn multiple heterogeneous distributions (separate yet coupled) at once, similar to the joint graphical lasso [8]. Moreover, PE-MRFs can be extended to include higher-order interactions, which open up new applications and can help increase the potential impact of our work.

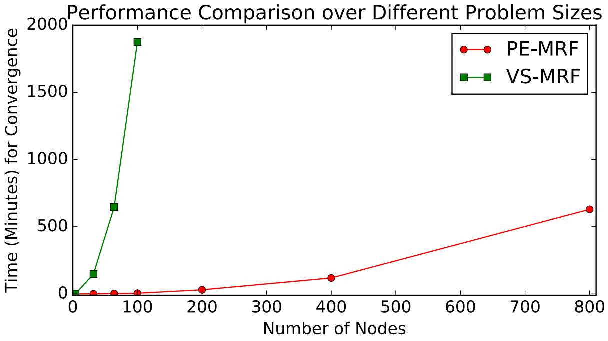


Figure 2: Scalability comparison between our PE-MRF model and VS-MRF [21] on heterogeneous data (Bernoulli, gamma, Gaussian, and Dirichlet nodes).

Acknowledgements. This work was supported by NSF IIS-1149837, NIH BD2K, DARPA XDATA, DARPA SIMPLEX, Stanford Data Science Initiative, Boeing, Bosch, Lightspeed, SAP, and Volkswagen.

References

[1] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, and K. Ohe. Extraction of adverse drug effects from clinical records. *Proceedings of the 13th World Congress on Medical Informatics*, 160(Pt 1):739–43, 2010.

[2] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *JMLR*, 9:485–516, 2008.

[3] B. Barzel and A.-L. Barabási. Network link prediction by global silencing of indirect correlations. *Nature Biotechnology*, 31(8):720–725, 2013.

[4] J. Besag. Statistical analysis of non-lattice data. *The Statistician*, pages 179–195, 1975.

[5] C. Bishop. Pattern recognition. *Machine Learning*, 128, 2006.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[7] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[8] P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

[9] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[10] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *ICML*, pages 433–440. ACM, 2009.

[11] A. Jalali, P. Ravikumar, V. Vasuki, and S. Sanghavi. On learning discrete graphical models using group-sparse regularization. In *AISTATS*, pages 378–387, 2011.

[12] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques.* MIT press, 2009.

[13] T. Le, L. Liu, A. Tsykin, G. J. Goodall, B. Liu, B.-Y. Sun, and J. Li. Inferring microRNA–mRNA causal regulatory relationships from expression data. *Bioinformatics*, 2013.

[14] J. Lee and T. Hastie. Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24(1):230–253, 2015.

[15] P.-L. Loh, M. Wainwright, et al. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics*, 41(6):3022–3049, 2013.

[16] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

[17] C. G. A. Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[18] P. Ravikumar, M. Wainwright, J. Lafferty, et al. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[19] P. Ravikumar, M. Wainwright, G. Raskutti, B. Yu, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[20] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications.* CRC Press, 2005.

[21] W. Tansey, O. H. M. Padilla, A. S. Suggala, and P. Ravikumar. Vector-space Markov random fields via exponential families. *ICML*, 2015.

[22] M. Wainwright and M. Jordan. Log-determinant relaxation for approximate inference in discrete Markov random fields. *Signal Processing, IEEE Transactions on*, 54(6):2099–2109, 2006.

[23] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[24] D. M. Witten and R. Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636, 2009.

[25] E. Yang, Y. Baker, P. Ravikumar, G. Allen, and Z. Liu. Mixed graphical models via exponential families. In *AISTATS*, pages 1042–1050, 2014.

[26] E. Yang, P. Ravikumar, G. Allen, and Z. Liu. On Poisson graphical models. In *NIPS*, pages 1718–1726, 2013.

[27] K. Zuberi, M. Franz, H. Rodriguez, J. Montojo, C. T. Lopes, G. D. Bader, and Q. Morris. GeneMANIA prediction server 2013 update. *Nucleic Acids Research*, 41(W1):W115–W122, 2013.