Learning the Network Structure of Heterogenerous Data via Pairwise Exponential MRF

Jong Ho Kim, Youngsuk Park
December 17, 2016

1 Introduction

Markov random fields (MRFs) are a fundamental fool on data modeling. And the study on inferring the network structure (or Markov structure) has provoked an interested in many applications in machine learning. 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 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 [13]. In order to faithfully model such 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 a challenging problem for both computational and mathematical reasons, and scalable inference methods have not yet been developed. In this paper, we propose a subset of the multivariate exponential family distribution, which we call the *pairwise exponential Markov random field* (PE-MRF).

Our PE-MRF model explicitly reveals the Markov structure across the variables and can cover many common distributions, such as Ising models, Gaussian MRFs, and mixed (heterogeneous) models. A distinct representation of the joint distribution allows for a compact representation of heterogeneous models and, eventually, a much faster structure/parameter learning method ($O(p^3)$, compared to $O(p^4)$, to learn a p node distribution, which has on the order of p^2 unknown variables).

After formally defining the model, we propose a method of learning 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 only been used for solving Ising models [2, 17]. 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 a convex problem, minimizing the log-determinant divergence [15] plus a group sparsity penalty [7]. We call this the *group graphical lasso* for PE-MRFs, since it turns out to be a generalization of the well-known graphical lasso [6], a popular method of learning Gaussian MRFs. In other words, 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 method is sparsistent [15, 19], meaning that, under some mathematical assumptions, we are 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, typically requiring Newton-type methods [3, 10, 16, 19], we develop an algorithm based on the alternating direction method of multipliers (ADMM) [4], where 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:

- Propose the PE-MRF, a pairwise exponential family distribution capable of revealing the Markov structure across heterogeneous variables.
- Formulate an approximated maximum likelihood problem by deriving a tractable upper bound on the log-partition function.
- Develop a scalable ADMM algorithm with closed-form updates for each subproblem.
- Prove that our method is sparsistent.

1.1 Related Works on Pairwise Models

Note that our PE-MRFs extends previous models of graphical inference [11, 16, 19, 2, 17, 11, 12, 4] We examine several alternative methods in more detail below. Our PE-MRF model is the first to be able to attain three desired conditions: (1) generality of the model, (2) scalablity of the algorithm, and (3) statistical guarantee to recover Markov structure, but previous approaches satisfy up to two of these three desired conditions.

Limited Heterogeneous Distributions. When the distributions at every node are uniparameter, Yang et al. [19] proposed a pairwise model with pseudo-likelihood approach to learn the Markov structure, which is unable to be generalized to multiparameter settings. Similarly, Lee et al. [10] used a pseudo-likelihood approach for a limited heterogeneous model (discrete-Gaussian mixed models). However, they did not provide any sparsistency guarantee, and it was unable to generalize to other exponential family distributions.

Vector-Space MRFs (VS-MRFs). A separate approach, proposed by Tansey et al. [16] and called VS-MRF, 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 for heterogeneous setting. Instead, VS-MRFs rely on the pseudo-likelihood. From an algorithmic perspective, to learn a p-node graphical model in k ADMM iterations [4], our algorithm has a runtime of $O(kp^3)$ whereas VS-MRF takes $O(kp^4)$.

Node-wise Regression. Aside from approximated maximum likelihood, another common approach for scalable inference is node-wise regression[11, 12, 14]. This type of method, however, has only been suggested for a limited class of homogeneous distributions, and not for broader problems in the general heterogeneous domain. As such, there is no known method of using node-wise regression to learn heterogeneous distributions, or to guarantee sparsistency in these cases.

2 Model: 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 $G = (V, E)$, with $|V| = p$ across a set of heterogeneous variables. Here, we denote $\langle A, B\rangle_F = \mathbf{Tr}(AB)$, the Frobenius inner product.

Definition For a p-variate random vector $X = \{X_1, \ldots, X_p\}$ defined over (heterogeneous) domains $\mathcal{X} = \otimes\{\mathcal{X}_r\}_{r=1}^p$, suppose the conditional distribution of each variable X_r given the remaining $p-1$ variables, which we call $X_{\backslash r}$, 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, 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 $\log \int_{\mathcal{X}} \exp\big\{ \sum_{r=1}^p \theta_r^T B_r(x_r) + \sum_{s,t=1}^p \left\langle \Theta_{st}, B_t(x_t) B_s(x_s)^T \right\rangle_F + \sum_{r=1}^p C_r(x_r) \big\} \nu(dx)$.Note that the edge-parameters $\{\Theta_{st}\}_{s,t=1}^p$ explicitly reveal underlying Markov structure, i.e., the $m_s \times m_t$ matrix $\Theta_{st} = 0$ if and only if X_s and X_r are conditionally independent given all others [9].

Examples. This PE-MRFs can model GMRFs by setting $B_r(X_r) = X_r$, $C_r(X_r) = \frac{1}{2}\log(2\pi)$, $\mathcal{X}_r = \mathbf{R}$. And it can also include Ising Models by setting $B_r(X_r) = X_r$, $C_r(X_r) = 0$, and $\mathcal{X}_\nabla = \{0, 1\}$. Not only that, it is able to design other mixed (heterogeneous) models, as long as the $A(\theta)$ is finite, by choosing a associated $B_r(X_r)$ and $C_r(X_r)$ with a desired exponential in the node-wise manner.

3 Estimation: Approximate Maximum Likelihood Approach

For learning the Markov structure, we formulate the maximum likelihood problem with regularization

$$\underset{\theta}{\text{minimize}} - \left\langle \boldsymbol{\theta}, \overline{\boldsymbol{B}(\mathbf{x})} \right\rangle + A(\boldsymbol{\theta}) + R_\lambda(\boldsymbol{\theta}). \tag{2}$$

Here $\overline{\boldsymbol{B}(\mathbf{x})}$ is the averaged sufficient statistic $B(x^{(k)})$ over $\{x^{(1)}, \ldots, x^{(n)}\}$ samples. And, to estimate sparse graphical structure, we use the ℓ_1/ℓ_2 group lasso penalty [8] $R_\lambda(\boldsymbol{\theta}) = \lambda \sum_{s \neq t} w_{st} \left\| \Theta_{st} \right\|_F$, where $\|A\|_F = \sqrt{\sum_{i,j=1}^{m_i,m_j} a_{ij}}$ is Frobenious norm. 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 values, then this becomes a standard lasso penalty. Here, $\{w_{st}\}_{s,t=1}^p$ is a set of scalar parameters balancing the weights on the set of quantities $\{\|\Theta_{st}\|_F\}_{s,t=1}^p$, each of whose value can depend on the size or variance.

3.1 Approximated Maximum Likelihood

One big issue is that $A(\boldsymbol{\theta})$ involves a high-dimensional integral and is typically intractable for the general multivariate cases. In order to overcome this, we use an approximated maximum likelihood [2], where we replace $A(\boldsymbol{\theta})$ in problem (2) with $U(\boldsymbol{\theta})$, a tractable (convex) upper bound,

$$\underset{\boldsymbol{\theta}}{\text{minimize}} - \left\langle \boldsymbol{\theta}, \overline{\boldsymbol{B}(\mathbf{x})} \right\rangle + U(\boldsymbol{\theta}) + R_\lambda(\boldsymbol{\theta}). \tag{3}$$

In the following, we present one theorem deriving convex upper bound on log partition function $A(\boldsymbol{\theta})$, which eventually allows to build a scalable method of learning the Markov structure of such distributions. For notational simplicity, we define

$$M_\nu[\boldsymbol{\theta}] = \begin{bmatrix} \nu & (\theta_1/2)^T & (\theta_2)^T/2 & \ldots & (\theta_p/2)^T \\ \theta_1/2 & \Theta_{11} & \Theta_{12} & \ldots & \Theta_{1p} \\ \theta_2/2 & \Theta_{21} & \Theta_{22} & & \Theta_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_p/2 & \Theta_{p1} & \Theta_{2p} & \ldots & \Theta_{pp} \end{bmatrix}.$$

Then, we have

Theorem 3.1

$$U(\boldsymbol{\theta}) = \frac{1}{2} \min_{\nu \in \mathbf{R}} \left\{ -\frac{1}{2} \log\det\left(-M_{1+\nu}[\boldsymbol{\theta}]\right) - \nu \right\} - \frac{1}{2} \left\langle M_1[\boldsymbol{\theta}], D \right\rangle_F + c.$$

where D is a diagonal matrix and c is constant, both of which only depends on the types of nodes.

Sketch of Proof) First, we express the $A(\boldsymbol{\theta})$ as its entropy function, and derive the simplified one $H(B_1(X_1), \ldots, B_p(X_p))$ (in Lemma). Second, we add independent uniformly distributed variable U_r with sufficient narrow supports to the node-potentials $B_r(X_r)$ so that additive variable $B_r(X_r) + U_r$ can uniquely determine $B_r(X_r)$. And express $H(X)$ as $H(B_1(X_1) + U_1, \ldots, B_p(X_p) + U_p)$, where D appears . Lastly, we derive the upperbound on $H(B_1(X_1) + U_1, \ldots, B_p(X_p) + U_p)$ from the fact that, under the same covariance matrix, the entropy of any continuous random variables is upperbounded by that of Guassian distribution.

Corollary 3.2 *For a PE-MRF, the approximated negative maximum log-likelihood problem becomes*

$$\min_{\boldsymbol{\Theta} \in \mathbf{S}_{++}^{d+1}} \left\{ \left\langle \boldsymbol{\Theta}, \overline{\boldsymbol{B}_{aug}(\mathbf{x})} \right\rangle_F - \log\det \boldsymbol{\Theta} + R_\lambda(\boldsymbol{\Theta}) \right\}$$

where $d = \sum_{r=1}^p m_r$, $\boldsymbol{\Theta} = M_\nu[\boldsymbol{\theta}]$ is the augmented natural parameter of $\boldsymbol{\theta}$, $\overline{\mathbf{B}_{aug}(\mathbf{x})} = M_1[\overline{\mathbf{B}(\mathbf{x})}] + D$ is augmented and shifted version of averaged sufficient statistics, and $R_\lambda(\boldsymbol{\Theta}) \equiv R_\lambda(\boldsymbol{\theta})$.

We can view this problem as a ℓ_1/ℓ_2 regularized log-determinant divergence with respect to the sample average of node-potentials, defined by the Bregman divergence corresponding to the log-determinant function [15]. We call problem (3) the *group graphical lasso* for a PE-MRF, since it is an extension of the classic graphical lasso problem [6, 11, 15] to heterogeneous settings. For a zero-mean Gaussian MRF, for example, we set $[\theta_1, \ldots, \theta_p]^T = \mathbf{0}$. 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.

4 Optimization Algorithm

Here, we propose an algorithm to solve the group graphical lasso (3). Our approach is based on the alternating direction method of multipliers (ADMM) [4] With ADMM, we split problem (3) into two subproblems and use a message-passing algorithm to iteratively converge to the solution. We show the closed-form updates for each of the subproblems, which allows for a scalable algorithm in any setting. We use a stopping criterion based on the primal and dual residual values being below specified thresholds [4]. To solve, we introduce a consensus variable $\boldsymbol{Z} = \boldsymbol{\Theta}$ and rewrite (3) as its equivalent problem,

$$\min_{\boldsymbol{Z}=\boldsymbol{\Theta},\boldsymbol{\Theta}\in\mathbf{S}_{++}^{d+1}} \left\langle \boldsymbol{\Theta}, \overline{\boldsymbol{B}_{aug}(\mathbf{x})} \right\rangle_F - \log\det\boldsymbol{\Theta} + \lambda_n \sum_{i\neq j} w_{ij} \left\| Z_{ij} \right\|_F ,$$

$\boldsymbol{\Theta}$-Update. The $\boldsymbol{\Theta}$ update is $\boldsymbol{\Theta}^{k+1} := \frac{1}{2\eta} Q\big(\Lambda + \sqrt{\Lambda^2 + 4\eta I}\big) Q^T$, where $\eta = \frac{\rho}{n}$ and $Q\Lambda Q^T$ is the eigendecomposition of $\eta(\boldsymbol{Z}^k - \boldsymbol{U}^k) - \overline{\mathbf{B}_{aug}(\mathbf{x})}$ [18].

$\boldsymbol{Z}$-Update. For $i, j \in \{1, \ldots, p\}$, compute $Z_{ij,i\neq j} = \left(1 - \frac{\lambda_n w_{ij}}{\rho \left\| \Theta_{ij}^{k+1} + U_{ij}^k \right\|_F}\right) \left(\Theta_{ij}^{k+1} + U_{ij}^k\right)$ if $\left\| \Theta_{ij}^{k+1} + U_{ij}^k \right\|_F \geq \frac{\lambda_n w_{ij}}{\rho}$ or 0 otherwise. And $\boldsymbol{Z}^{k+1} := \boldsymbol{\Theta}^{k+1} + \boldsymbol{U}^k$ for the rest elements.

$\boldsymbol{U}$-Update. The $\boldsymbol{U}$ update is $\boldsymbol{U}^{k+1} := \boldsymbol{U}^k + \boldsymbol{\Theta}^{k+1} - \boldsymbol{Z}^{k+1}$.

Algorithmic Complexity. Note that the eigendecomposition of the $\boldsymbol{\Theta}$ update is the main computational task in our algorithm, with a runtime of $O(p^3)$. And thus, for k ADMM iterations, the computational cost is $O(kp^3)$, which is very efficient considering the fact that the total number of parameters $O(p^2)$. On the other hand, the pseudo-likelihood approach requiring Newton-type methods [20, 10, 16] 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 statistical and graphical conditions under which we are guaranteed to recover the underlying graphical structure embedded in the edge matrix.

Notation. We denote a (column vector of) node potentials $B_{node}(X) = vec[B_1(X_1), \ldots, B_p(X_p)]$. We define an operator norm $|||A|||_{\infty,2} = \max_i \sum_j \left\| A_{ij} \right\|_2$. For a likelihood function $g(\boldsymbol{\Theta}) = \langle \boldsymbol{\Theta}, \overline{\mathbf{B}_{aug}(\mathbf{x})} \rangle - \log\det\boldsymbol{\Theta}$, we denote its optimal solution as $\boldsymbol{\Theta}^*$. Then its gradient on $\boldsymbol{\Theta}^*$ is $\Sigma^* := \nabla g(\boldsymbol{\Theta}^*) = \overline{\mathbf{B}_{aug}(\mathbf{x})}$ and its Hessian is $\Gamma^* = \nabla^2 g(\boldsymbol{\Theta}^*) := \overline{\mathbf{B}_{aug}(\mathbf{x})} \otimes \overline{\mathbf{B}_{aug}(\mathbf{x})}$ [5].

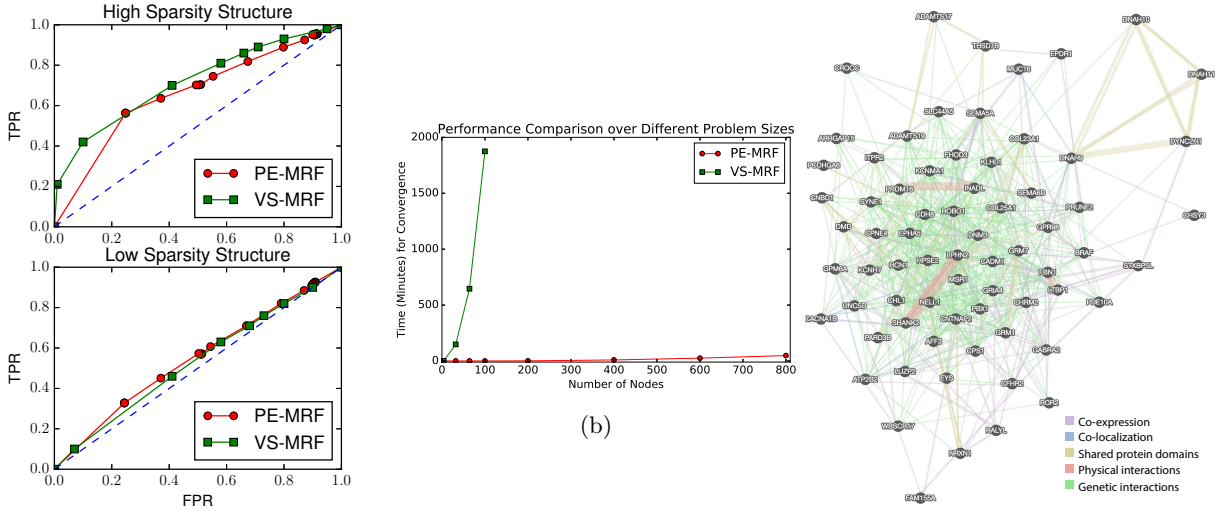
Assumptions. We make assumptions, partly adopted from Loh et al.[11] and Ravikumar et al. [15].

1. The PE-MRF has an underlying graphical structure with singleton separator sets.
2. $\mathbf{E}[B_{node}(X)]$ and $\mathbf{Cov}[B_{node}(X)]$ is bounded.
3. The Hessian satisfies incoherence condition, $\left|\left|\left| \Gamma_{S^cS} \left(\Gamma_{S^cS}\right)^{-1} \right|\right|\right|_{\infty,2} \leq \frac{w_{min}}{w_{max}}(1-\alpha)$ for some $\alpha \in (0, 1]$.

Theorem 5.1 *For $\lambda_n \geq \kappa_1 \sqrt{\frac{\log(m_{max} p)}{n}}$, let $\hat{\boldsymbol{\Theta}}$ be the (unique) solution of group graphical lasso. If the number of samples $n \geq \kappa_2 \log m_{max} p$, then the estimated edge $E(\hat{\boldsymbol{\Theta}}) = \{(s,t) \mid \left\| \hat{\Theta}_{ij} \right\|_2 \geq \kappa_3 \lambda_n\}$ can exactly recover the real edge set E with probability at least $1 - e^{-cn}$.*

Here κ_1, κ_2 and κ_3 depend on the $\{m_r\}, \{w_{st}\}$ and other parameters defined in assumptions, and c is some universal constant.

As a result, assuming a PE-MRF X satisfies the graph condition, for the proper choice of parameter and sample size, our estimator guarantees with high probability to recover the true underlying Markov structure.



(a) (b) (c)

Figure 1: (a) ROC curves on synthetic data yield similar results to VS-MRF, (b) Scalability comparison, (c) Known interactions from biomedical database of all genes in the top k-core of the inferred network.

6 Experiments

Synthetic Data. We analyze performance on this 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 Tansey et al.'s method [16], called VS-MRF, which is the only other approach that can explicitly model such a diverse distribution. 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. Our PE-MRF solver can learn the 100-node distribution in **under 30 seconds** (Figure 1b). In contrast, VS-MRF [16] takes **over 31 hours** to converge.

Heterogeneous Genomic Networks. We use PE-MRF and test its applicability for inference of genomic regulatory networks. We use Level III public data from The Cancer Genome Atlas (TCGA) [13] 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. We infer a PE-MRF of 500-node Gaussian (genes) and 314-node Poisson (miRNAs), choosing the regularization parameter λ minimizing the Akaike information criterion (AIC).

In particular, we consider a 65-core of the gene-gene subnetwork. From the external data GeneMANIA [21], We validate how many gold standard edges exist between genes from the tightly connected 65-core, and confirmed that the core of the inferred gene subnetwork is well supported by many established interactions (Figure 1c). Despite the 290 patient samples, we can still notice the abundance of interaction edges, which match the connectivity of the inferred gene subnetwork.

7 Conclusion

We have developed a method of learning Markov networks. We propose a pairwise exponential Markov random field (PE-MRF), a subclass of the exponential family for heterogeneous distributions. Then we derive an upperbound on the log partition function, formulate an approximated maximum likelihood, and develop an scalable ADMM algorithm with closed-form updates. We then prove that our estimator guarantees the true edge recovery. Our promising results, as well as the widespread applications with heterogeneous data, lead to many potential extensions of this work. For example, it is possible that the Markov structure changes over time. Instead of inferring a single network, we could use the timestamped observations to estimate a time-varying network.

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

[4] 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.

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

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

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

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

[9] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

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

[11] 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.

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

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

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

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

[20] E. Yang, P. Ravikumar, G. Allen, and Z. Liu. On graphical models via univariate exponential family distributions. *arXiv preprint arXiv:1301.4183*, 2013.

[21] 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.