Ordinal Graphical Models: A Tale of Two Approaches

Arun Sai Suggala [1] **Eunho Yang** [2 3] **Pradeep Ravikumar** [1]

Abstract

Undirected graphical models or Markov random fields (MRFs) are widely used for modeling multivariate probability distributions. Much of the work on MRFs has focused on continuous variables, and nominal variables (that is, unordered categorical variables). However, data from many real world applications involve *ordered* categorical variables also known as *ordinal* variables, *e.g.,* movie ratings on Netflix which can be ordered from 1 to 5 stars. With respect to *univariate* ordinal distributions, as we detail in the paper, there are two main categories of distributions; while there have been efforts to extend these to multivariate ordinal distributions, the resulting distributions are typically very complex, with either a large number of parameters, or with non-convex likelihoods. While there have been some work on tractable approximations, these do not come with strong statistical guarantees, and moreover are relatively computationally expensive. In this paper, we theoretically investigate two classes of graphical models for ordinal data, corresponding to the two main categories of univariate ordinal distributions. In contrast to previous work, our theoretical developments allow us to provide correspondingly two classes of estimators that are not only computationally efficient but also have strong statistical guarantees.

1. Introduction

Undirected graphical models, also know as Markov random fields (MRF), are very popular for modeling multivariate random variables. They use undirected graphs to model the conditional independence structure among the variables. This conditional independence structure provides us with useful insights about how different variables interact with each other. As a result MRFs are extensively used in a variety of fields, including Natural Language Processing (Manning & Schütze, 1999), Biology (Friedman, 2004) and Medicine (Allen & Liu, 2012).

Popular parametric families in the general class of MRFs are Gaussian graphical models (Speed & Kiiveri, 1986; Rue & Held, 2005) for continuous (and bell-shaped) data, Ising and discrete graphical models (Ising, 1925; Jalali et al., 2011) for nominal data, and mixed cases of these two instances (Lauritzen & Wermuth, 1989; Yang et al., 2014). However, variables that occur in many real world applications have ordered categorical scales. For example, in medical data, diseases are graded from *mild* to *fatal*, severity of an injury is rated from *mild injury* to *death*, stages of a disease is rated from *I* to *III*. Ordinal variables also occur very commonly in data collected from surveys. For example, each subject taking a survey could be asked to respond to a question using categories such as *strongly disagree, disagree, undecided, agree, strongly agree*. These examples clearly show that ordinal data is pervasive in many real world applications.

It will be instructive at this juncture to review *univariate* distributions for ordinal data, which fall into two main categories. The first category of distributions consists of parameterizing various odds ratios involving the ordinal variable, and include the cumulative ratio, continuation ratio, and consecutive ratio logit models (Armstrong & Sloan, 1989; Agresti, 2010). In practice, cumulative ratio and continuation ratio models are know to work better than the consecutive ratio model (McCullagh, 1984), and as a consequence, the consecutive ratio model is relatively less well considered among these three models. Some generalizations of the above models have been considered in literature (e.g., Partial Proportional Odds model (Peterson & Harrell Jr., 1990)), which fall in between the simpler class of models listed earlier and the multinomial logistic regression model in terms of their parametric complexity, but are nonetheless not as popular as the simpler class of models. The second category of univariate ordinal distributions are based on the natural generative assumption that the ordinal variable is a quantization of a real-valued latent variable. Common distributions imposed on the latent variable

[*]Equal contribution [1]Carnegie Mellon University, Pittsburgh, USA [2]School of Computing, KAIST, Daejeon, South Korea [3]AITrics, Seoul, South Korea. Correspondence to: Arun Sai Suggala <asuggala@andrew.cmu.edu>, Eunho Yang <eunhoy@kaist.ac.kr>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

include the logistic distribution, in which case it reduces to the classical cumulative ratio model (Agresti, 2010), as well as the more popular standard normal distribution, in which case it is called the ordered probit model (Becker & Kennedy, 1992).

There has been some effort to construct multivariate ordinal distributions, and they again fall into two categories, namely the multivariate extensions of the corresponding two categories of univariate ordinal distributions. In the first category, there has been a line of work (see Bartolucci et al., 2007, and references therein) on designing parametrized odds ratio based models. Bartolucci et al. (2007) provide a general framework for designing multivariate ordinal models by parameterizing marginal distributions using various odds ratios. However, the parametric form of these models are complex, and have a large number of parameters, and so do not necessarily scale to high-dimensional settings. Moreover, these models cannot be readily expressed as graphical model distributions.

In the second category of multivariate ordinal distributions, the ordinal random vector is modeled as a quantization of a real-valued latent random vector. Here, the efforts have focused on the use of the multivariate normal distribution for the latent random vector, due plausibly to its more convenient mathematical nature; the resulting model is also known as the multivariate probit model (Ashford & Sowden, 1970; Amemiya, 1974). But even with this modeling assumption, the likelihood of the observed ordinal random vector is not available in closed-form, is considerably complex due to the presence a multi-dimensional integral, and in particular is non-convex, so that learning the model given just the ordinal observations is typically computationally intractable. Consequently Expectation Maximization (EM) and approximate EM based approaches (Chib & Greenberg, 1998; Guo et al., 2015) have been proposed to compute the Maximum Likelihood Estimate (MLE). A caveat with these is that they do not come with statistical guarantees, and moreover, as our experiments show, are still computationally expensive. Because of the computational intractability of the MLE, alternative approaches have been proposed which maximize the pairwise, composite likelihood of the data, which involve only the univariate and bivariate marginals (Lindsay, 1988; De Leon, 2005; Han & Pan, 2012). Another class of methods have been proposed which estimate the model parameters in multiple stages (Muthén, 1984; Jöreskog, 1994), unlike the previous approaches which estimate all the parameters using a single estimating equation. Although these approaches are computationally very efficient they do not provide consistent estimates in high dimensional settings.

Thus, in spite of the ubiquity of multivariate ordinal data, there have been limited popular applications of these multivariate ordinal distributions to model such data, particularly in high-dimensional settings, with a large number of ordinal variables.

In this paper, we develop multivariate ordinal graphical model distributions, for which the estimators are computationally tractable. Following the development of classical univariate ordinal distributions, our investigations fall into two categories. In the first category, we investigate multivariate extensions of the log-odds parameterized univariate ordinal distributions. Towards this, we leverage the line of work in Yang et al. (2012), which provides a mechanism to extend a univariate distribution to a multivariate graphical model distribution, by using the univariate distributions to specify node-conditional distributions. Their theoretical treatment requires the univariate distributions to belong to exponential families, while as we show, the log-odds based univariate ordinal distributions, namely *cumulative* and *continuation ratio* models (Agresti, 2010), do not belong to exponential families. Could perhaps the framework of Yang et al. (2012) be nonetheless be extended to this case? We provide a definitive answer in the negative, and show that using these univariate ordinal distributions as node-conditional distributions **cannot** lead to a consistent joint distribution. Whereas, the *consecutive ratio model* (Agresti, 2010) does belong to the exponential family, and can be extended to a multivariate graphical model distribution. However, as we will demonstrate in our experiments, this resulting *novel* class of consecutive ratio MRFs has mixed empirical results.

In the second category, we investigate the multivariate extensions of latent variable quantization based univariate ordinal distributions. We focus on the case where the latent random vector is multivariate Gaussian. Here, we leverage the structure of the ordinal data, and provide a very simple multistage estimator along the lines of Muthén (1984); Jöreskog (1994), that finesses computing the likelihood, and accordingly is computationally tractable, but interestingly, also comes with strong statistical guarantees. We corroborate our findings on both simulated and real data.

2. Multivariate Odds Ratio based Models

In the first part of the paper, we investigate the approach of Yang et al. (2012) i.e., specifying node-conditional distributions via classical univariate ordinal distributions, and then exploring the corresponding joint distribution via Hammersley-Clifford-esque analyses.

2.1. MRFs via Univariate Latent Quantized Ordinal Models

The most popular class of univariate ordinal distributions rely on a generative model that quantizes a latent variable.

Suppose we have a real-valued latent random variable $Z \in \mathbb{R}$ with CDF denoted by $\mathbb{P}[Z \leq z] = g(z - \mu)$, where μ is a location parameter of the distribution. Suppose that the ordinal random variable $Y \in \{0, \ldots, M\}$ can be written as a discretized version of the real-valued variable Z, as $Y = j, \quad \text{iff} \quad Z \in (\theta_{j-1}, \theta_j]$, for some location (or cut-point) parameters $\{\theta_j\}_{j=-1}^{M}$, and with $\theta_{-1} = -\infty, \theta_M = \infty$. It then follows that the probability mass function of the ordinal variable Y can be written as

$$\mathbb{P}[Y = j] = g(\theta_j - \mu) - g(\theta_{j-1} - \mu). \tag{1}$$

A popular distribution for the latent real-valued variable Z is the univariate logistic distribution, where $Z \sim \text{logistic}(\mu, 1)$, so that the function $g(\cdot)$ above is the logistic function, $g(t) = \sigma(t) = 1/(1 + \exp(-t))$. In this case, the distribution above can also be expressed in a more compact form in terms of log-odds ratios as: $\log\left(\frac{\mathbb{P}(Y \leq j)}{\mathbb{P}(Y > j)}\right) = \theta_j - \mu$. Accordingly, this class of ordinal distributions are also called *cumulative ratio* models. We now consider the general framework of Yang et al. (2012), of using univariate ordinal distribution in (1) to specify node-conditional distributions, and deriving a consistent joint distribution.

Let $\mathbf{Y} = (Y_1, \ldots, Y_p)$ be a p-dimensional ordinal random vector. To simplify the notation, in the sequel we assume that the domains of all the random variables $\{Y_s\}_{s=1}^{p}$ are same and equal to $\{0, 1, \ldots, M\}$. Let $G = (V, E)$ be a graph with nodes corresponding to each of the random variables $\{Y_s\}_{s=1}^{p}$. Suppose that for each $s \in V$, we have

$$\mathbb{P}[Y_s = j | Y_{\setminus s}] = g(\theta_{s;j} - \mu_s(Y_{\setminus s})) - g(\theta_{s;j-1} - \mu_s(Y_{\setminus s})), \tag{2}$$

where the location parameter $\mu_s(Y_{\setminus s})$ is an arbitrary function of the rest of the variables and g is the logistic function. We now present the following theorem which shows that these node-conditional distributions do not lead to a consistent joint distribution.

Theorem 1. *Consider a p-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Y_p)$ with domain $\{0, 1, \ldots, M\}^p$. And let $G = (V, E)$ be a graph with nodes corresponding to each of the random variables $\{Y_s\}_{s=1}^{p}$. Suppose that all node-conditional distributions of this random vector follow the univariate cumulative ratio model in (2), where for each $s \in V$, the location parameter $\mu_s(Y_{\setminus s})$ is an arbitrary function of the rest of the variables.*

*Then, for $M \geq 1$, there exist real valued parameters $\{\theta_{s;j}\}_{s \in V, j \in [M]}$ for which the specified node-conditional distributions **are not** consistent with any joint distribution over $\mathbf{Y}$ that is Markov with respect to the graph G with factors of size at most 2.*

The proof of the above theorem can be found in Appendix A and follows the Hammersley Clifford type analysis of Besag (1974).

2.2. MRFs via Continuation Ratio Models

One modification to the cumulative ratio model that has been considered in the literature is that of a closely related log-odds ratio: $\log\left(\frac{\mathbb{P}(Y=j)}{\mathbb{P}(Y>j)}\right) = \theta_j - \mu$. This class of univariate ordinal distributions are also called *continuation ratio* models. From the log-odds ratio above, denoting $\eta_j := \theta_j - \mu$, the Probability Mass Function (PMF) of the random variable Y can be derived as

$$\mathbb{P}[Y = j] := \frac{\exp(\eta_j)}{\prod_{i=0}^{j}\left(1 + \exp(\eta_i)\right)}, \tag{3}$$

for $j = 0, \ldots, M-1$. Then, the probability $\mathbb{P}(Y = M)$ is fixed as: $1 - \sum_{i=0}^{M-1} P(Y = i) = \frac{1}{\prod_{i=0}^{M-1}\left(1+\exp(\eta_i)\right)}$, so that the summation of the PMF equals 1.

As in the previous section, given this univariate ordinal distribution, we ask if we could employ the strategy of Yang et al. (2012), of using these to specify node-conditional distributions, and deriving a consistent joint distribution? Specifically, suppose that for each node $s \in V$ we have

$$\mathbb{P}[Y_s = j | Y_{\setminus s}] = \frac{\exp(\eta_{s;j}(Y_{\setminus s}))}{\prod_{i=0}^{j}\left(1 + \exp(\eta_{s;j}(Y_{\setminus s}))\right)}, \tag{4}$$

where $\eta_{s;j}(Y_{\setminus s}) = \theta_{s;j} - \mu_s(Y_{\setminus s})$ and the location parameter $\mu_s(Y_{\setminus s})$ is an arbitrary function of the rest of the variables. The following theorem shows that these node conditional distributions do not lead to a consistent joint distribution. The proof of this theorem is provided in Appendix B.

Theorem 2. *Consider a p-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Y_p)$ with domain $\{0, 1, \ldots, M\}^p$. And let $G = (V, E)$ be a graph with nodes corresponding to each of the random variables $\{Y_s\}_{s=1}^{p}$. Suppose that all node-conditional distributions of this random vector follow the univariate continuation ratio model in (4), where for each $s \in V$, the location parameter $\mu_s(Y_{\setminus s})$ is an arbitrary function of the rest of the variables.*

*Then, for $M \geq 1$, there exist real valued parameters $\{\theta_{s;j}\}_{s \in [p], j \in [M]}$ for which the specified node-conditional distributions **are not** consistent with any joint distribution over $\mathbf{Y}$ that is Markov with respect to the undirected graph G.*

2.3. MRFs via a Consecutive Ratio model

A key caveat with the univariate cumulative and continuation ratio models is that they do not belong to exponential families, and in particular do not possess the regularities that allow for existence of consistent joint given node-conditionals belonging to these distributions. We now consider the third class of univariate ordinal distribu-

tions called *consecutive ratio* model which is defined as: $\log\left(\frac{\mathbb{P}(Y=j)}{\mathbb{P}(Y=j+1)}\right) = \theta_j - \mu$, for $j = 0, \ldots, M-1$.

As we show below, this ordinal distribution belongs to an exponential family, unlike ordinal distributions in earlier sections.

Proposition 1. *The consecutive ratio model for an ordinal random variable $Y \in \{0, \ldots, M\}$ belongs to an exponential family with sufficient statistics $\{\mathcal{I}[Y \leq j]\}_{j=0}^{M-1}$: $\mathbb{P}[Y] = \exp\Big(\sum_{j=0}^{M-1} \eta_j \mathcal{I}[Y \leq j] - A(\eta)\Big)$, where as before, $\eta_j := \theta_j - \mu$, for $j \in \{0, \ldots, M-1\}$.*

We now consider the counterpart of our key question in the earlier sections. Suppose we use the univariate ordinal distribution above to specify node-conditional distributions for an ordinal random vector $\mathbf{Y} = (Y_1, \ldots Y_p)$. Specifically, suppose that for each node $s \in V$, we have

$$\mathbb{P}[Y_s | Y_{\setminus s}] \propto \exp\Big(\sum_{j=0}^{M-1} \eta_{s;j}(Y_{\setminus s})\, \mathcal{I}[Y_s \leq j]\Big) \qquad (5)$$

where $\eta_{s;j} = \theta_{s;j} - \mu_s(Y_{\setminus s})$, for $j = 0, \ldots, M-1$, where the location parameter $\mu_s(Y_{\setminus s})$ is an arbitrary function of the rest of the variables. Since these node-conditional distributions belong to a univariate exponential family, an application of Proposition 1 of Yang et al. (2012) yields the following theorem:

Theorem 3. *The node-conditional distributions in (5) are consistent with a joint distribution that is Markov with respect to an undirected graph $G = (V, E)$, which in the pairwise case with factors of size atmost two necessarily has the following form: $\mathbb{P}(\mathbf{Y}) \propto \exp\Big(\sum_{s\in V, j\in[M-1]} \theta_{s;j}\, \mathcal{I}[Y_s \leq j] \sum_{(s,t)\in E} \sum_{j,k\in[M-1]} \theta_{st}\, \mathcal{I}[Y_s \leq j]\, \mathcal{I}[Y_t \leq k]\Big)$.*

The distribution in Theorem 3 can be rewritten in the following equivalent form

$$\mathbb{P}(\mathbf{Y}) \propto \exp\Big(\sum_{s\in V, j\in[M-1]} \theta_{s;j}\, \mathcal{I}[Y_s \leq j] + \sum_{(s,t)\in E} \theta_{st}\big(M - Y_s\big)\big(M - Y_t\big)\Big). \qquad (6)$$

Note that the pairwise interaction terms in the above distribution utilize the ordinality of $\mathbf{Y}$ through the term $\big(M - Y_s\big)\big(M - Y_t\big)$.

To estimate the parameters of the consecutive ratio model (6), we solve the following regularized node conditional log likelihood maximization problem at each node $s \in [p]$

$$\arg\min_{\bar{\theta}_s} -\sum_{i=1}^{n} \log \mathbb{P}(\mathbf{y}_{i;s} | \mathbf{y}_{i;\setminus s}) + \lambda_n \sum_{t\in V\setminus s} |\theta_{st}|,$$

where $\{\mathbf{y}_i\}_{i=1}^n$ are training samples and $\bar{\theta}_s = \{\theta_{s;j}\}_{j\in[M-1]} \cup \{\theta_{st}\}_{t\in V\setminus s}$.

We briefly note that existing results on statistical guarantees for estimators of exponential family graphical models (Yang et al., 2015; Tansey et al., 2015) carry over to the consecutive ratio model.

Contrast with Discrete/Nominal Graphical models: We now contrast the consecutive ratio model in (6) with the classical discrete nominal graphical model which treats the random variables at each node as nominal variables. Consider the following discrete graphical model over the random vector $\mathbf{Y}$:

$$\mathbb{P}(\mathbf{Y}) \propto \exp\Big(\sum_{s\in V, j\in[M-1]} \theta_{s;j}\, \mathcal{I}[Y_s = j] + \sum_{(s,t)\in E} \sum_{j,k\in[M-1]} \theta_{st;jk}\, \mathcal{I}[Y_s = j]\, \mathcal{I}[Y_t = k]\Big). \qquad (7)$$

Unlike the consecutive ratio model, the above model doesn't have a common edge parameter θ_{st} for different values of Y_s, Y_t. Each edge in the categorical model is parametrized using M^2 variables. As a result this model does not utilize the ordinality of Y and is also more complex compared to the consecutive ratio model. While this parameterization does encompass the consecutive ratio model parameterization, the key disadvantage is that the nominal graphical model has a larger number of parameters and hence greater sample complexity.

3. Multivariate Latent Quantized Models

In the previous section we considered using the mechanism of Yang et al. (2012) to directly construct multivariate ordinal graphical models from univariate ordinal distributions. In this section we revisit the classical and most popular class of univariate ordinal distributions based on the quantization of a real-valued latent variable. A natural class of multivariate distributions can be obtained by taking a *multivariate* latent random vector, and quantizing this to obtain a multivariate ordinal random vector.

3.1. Probit Graphical Model

The most popular instance of such multivariate quantized ordinal distributions is the case where the multivariate latent random vector is multivariate Gaussian, which is also known as the multivariate probit model (Ashford & Sowden, 1970; Amemiya, 1974). Thus the dependencies are all represented in the latent random vector via the underlying Gaussian distribution.

In the probit model, the ordinal random vector $\mathbf{Y} = (Y_1, \ldots Y_p)$ is assumed to be generated from a latent multivariate Gaussian random vector $Z = (Z_1, \ldots, Z_p)$, where

$Z \sim \mathcal{N}(0, \Sigma)$ and $Z_i \sim \mathcal{N}(0,1)\ \forall i \in [1,p]$. Each Y_i is obtained through discretization of Z_i as follows: $Y_i = k, \quad \text{iff} \quad Z_i \in [\theta_{k-1}^{(i)}, \theta_k^{(i)})$, where $\{\theta_k^{(i)}\}_{k=-1}^{M}$ is the set of thresholds, $\theta_{-1}^{(i)} = -\infty$, $\theta_M^{(i)} = \infty$. Then the density function of $\mathbf{Y}$, $\mathbb{P}(\mathbf{Y}; \Sigma, \Theta)$, is given by

$$\begin{aligned}&\mathbb{P}(\theta_{Y_1-1}^{(1)} \leq Z_1 < \theta_{Y_1}^{(1)}, \ldots, \theta_{Y_p-1}^{(p)} \leq Z_p < \theta_{Y_p}^{(p)}; \Sigma) \\ &= \int_{z \in C(\mathbf{Y}, \Theta)} \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2} z \Sigma^{-1} z^T\right) dz \end{aligned} \tag{8}$$

where $\Theta = \{\theta_k^{(j)} : j \in [1,p], k \in [-1, M]\}$ and $C(\mathbf{Y}, \Theta)$ is the hypercube defined by $[\theta_{Y_1-1}^{(1)}, \theta_{Y_1}^{(1)}) \times \ldots \times [\theta_{Y_p-1}^{(p)}, \theta_{Y_p}^{(p)})$.

Let $\mathbb{Y}_n = \{\mathbf{y}_i\}_{i=1}^n$ be n i.i.d realizations of the random vector $\mathbf{Y}$, drawn from probit model with parameters Θ^*, Σ^*. Then the ℓ_1-regularized Maximum Likelihood (ML) estimator to learn the parameters Σ, Θ from $\mathbb{Y}_n$ takes the form

$$\underset{\Sigma, \Theta}{\text{minimize}} - \sum_{i=1}^{n} \log \mathbb{P}(\mathbf{y}_i; \Sigma, \Theta) + \lambda_n \|\Sigma^{-1}\|_{1,\text{off}} \tag{9}$$

where $\|\cdot\|_{1,\text{off}}$ is the element-wise ℓ_1 norm excluding diagonal entries. It can be seen that the objective is non-convex, and intractable to optimize in general. Accordingly, approximate EM based approaches (Chib & Greenberg, 1998; Guo et al., 2015) have been proposed for learning the model parameters, but these are still relatively computationally demanding, and also do not come with the strong statistical guarantees of the actual regularized MLE solutions.

3.2. A Direct Estimation Method

In the second contribution of the paper, we propose an alternative procedure for the estimation of the unknown parameters Θ, Σ in the probit graphical model distribution in (8). This is a two stage procedure where in the first stage we estimate the thresholds, Θ, from the univariate marginals and in the second stage we estimate the polychoric correlations, Σ, from bivariate marginal distributions.

3.2.1. Estimation of Thresholds

We define $\widehat{\Theta}$, our estimator of Θ as follows

$$\widehat{\theta}_k^{(j)} = \begin{cases} -\infty & \text{if } k = -1 \\ \Phi^{-1}(\frac{1}{n}\sum_{i=1}^n \mathcal{I}(\mathbf{y}_{i,j} \leq k)) & \text{if } k = 0, \ldots, M-1 \\ \infty & \text{if } k = M \end{cases},$$

where $\Phi(.)$ is the CDF of standard normal distribution, $\mathcal{I}(.)$ is the indicator function, $\mathbf{y}_{i,j}$ is the j^{th} coordinate of vector $\mathbf{y}_i$. It can be seen that $\widehat{\Theta}$ consistently estimates Θ^*.

3.2.2. Estimation of Polychoric Correlations and Latent Graph Structure

We present a two step approach for estimation of Σ. In the first step, we compute a raw estimate $\widetilde{\Sigma}$ from the bivariate marginal likelihoods. In the next step we plugin the estimated covariance matrix $\widetilde{\Sigma}$ into the graphical lasso estimator (Friedman et al., 2008) to estimate the sparse latent graph structure and a smoothed estimate $\widehat{\Sigma}$.

Step 1: To estimate each entry of Σ we solve an independent optimization problem. Lets suppose we want to estimate Σ_{jk}, for $j \neq k$. The joint distribution of (Y_j, Y_k) is multinomial with probabilities $\mathbb{P}(Y_j, Y_k; \Theta, \Sigma_{jk}) = \mathbb{P}(\theta_{Y_j-1}^{(j)} \leq Z_j \leq \theta_{Y_j}^{(j)}, \theta_{Y_k-1}^{(k)} \leq Z_k \leq \theta_{Y_k}^{(k)}; \Sigma_{jk})$, where the joint distribution of random variables Z_j, Z_k is bivariate normal with mean $[0,0]$ and covariance $\begin{bmatrix} 1 & \Sigma_{jk} \\ \Sigma_{jk} & 1 \end{bmatrix}$. If Θ^* is known, then one could estimate the unknown parameter Σ_{jk} by maximizing the bivariate marginal log likelihood function, which has the following form

$$\begin{aligned}\ell_{jk}(\sigma; \Theta^*, \mathbb{Y}_n) &= \sum_{a=0}^{M}\sum_{b=0}^{M} \frac{n_{ab}}{n} \log \mathbb{P}(Y_j = a, Y_k = b; \Theta^*, \sigma) \\ &= \sum_{a=0}^{M}\sum_{b=0}^{M} \frac{n_{ab}}{n} \log \phi_{ab;jk}(\sigma; \Theta^*), \end{aligned} \tag{10}$$

where $n_{ab} = \sum_{i=1}^n \mathcal{I}(\mathbf{y}_{i,j} = a, \mathbf{y}_{i,k} = b)$ and $\phi_{ab;jk}(\sigma; \Theta) = \mathbb{P}(Y_j = a, Y_k = b; \Theta, \sigma)$. However, the thresholds Θ^* are unknown. So to estimate Σ_{jk}, we replace Θ^* with its estimator $\widehat{\Theta}$ and maximize the following log likelihood

$$\widetilde{\Sigma}_{jk} = \underset{\sigma \in \mathcal{M}}{\arg\max}\ \ell_{jk}(\sigma; \widehat{\Theta}, \mathbb{Y}_n).$$

where $\mathcal{M}$ is the domain of σ, which is $(-1, 1)$ unless no additional constraint on covariance is placed. Note that this is a one-dimensional optimization problem which, under certain regularity conditions such as smoothness of the objective, allow us to solve to within error ϵ in time $O(1/\epsilon)$ by simply evaluating the objective over a fine grid in $\mathcal{M}$, and selecting the optimal grid point.

Step 2: In this step we plug-in $\widetilde{\Sigma}$ into a parametric Gaussian graphical model estimator to obtain the graph structure and the final covariance matrix. While any consistent parametric Gaussian estimator (e.g., graphical lasso estimator (Friedman et al., 2008), CLIME (Cai et al., 2011), graphical Dantzig selector (Yuan, 2010)) can be used to estimate the latent graph structure, in this work we focus on the graphical lasso estimator (glasso), which involves solving the following optimization problem

$$\widehat{\Sigma} = \underset{\Sigma^{-1} \succ 0}{\arg\min} \langle\langle \Sigma^{-1}, \widetilde{\Sigma} \rangle\rangle - \log\det\left(\Sigma^{-1}\right) + \lambda_n \|\Sigma^{-1}\|_{1,\text{off}}, \tag{11}$$

where $\langle\langle A, B \rangle\rangle$ denotes the trace inner product of A and B.

3.3. Theoretical Properties

In this section, we show that the direct estimation method we proposed in Section 3.2 is not just simple but has strong statistical guarantees. Specifically, we provide an ℓ_∞ bound for the inverse covariance estimator $\widehat{\Sigma}^{-1}$, and show its sparsistency with respect to graphical model structure recovery. For simplicity, we assume that Θ^* is given. However, extension to the case where Θ^* is unknown should be fairly straightforward.

We begin with introducing some notation. Let $\Gamma^* := \Sigma^* \otimes \Sigma^*$, where $\otimes$ denotes the Kronecker matrix product, denote the Hessian of $-\log\det(A)$ evaluated at $(\Sigma^*)^{-1}$. Let S be the set of indices corresponding to all nonzero entries in $(\Sigma^*)^{-1}$, and S^c be the complement of S. We also define $K_{\Sigma^*} := |||\Sigma^*|||_\infty$, and $K_{\Gamma^*} := |||(\Gamma^*_{SS})^{-1}|||_\infty$ for notational simplicity where $|||\cdot|||_\infty$ denotes the maximum absolute row sum of matrix. Let d be the maximum node-degree in the latent graph. Finally, let $\bar{\ell}_{jk}(\sigma;\Theta^*) = \mathbb{E}[\ell_{jk}(\sigma;\Theta^*)]$ be the population version of the sample loss defined in Equation (10). We now state our assumptions.

(C-1) There exists some $\alpha \in (0,1]$ such that $|||\Gamma^*_{S^cS}(\Gamma^*_{SS})^{-1}|||_\infty \leq 1-\alpha$.

(C-2) There is a constant $\delta > 0$ such that $|\Sigma^*_{jk}| \leq 1-\delta, \forall j < k$. Moreover, the likelihood function $\phi_{ab;jk}(\sigma;\Theta^*)$, is strictly positive $\forall|\sigma| \leq 1-\delta$, i.e., $\exists\gamma > 0$ such that $\forall|\sigma| \leq 1-\delta$, $\phi_{ab;jk}(\sigma;\Theta^*) \geq 1/\gamma$.

(C-3) The absolute value of the $1^{st}, 2^{nd}$ and 3^{rd}-order derivatives of $\phi_{ab;jk}(\sigma;\Theta^*)$ w.r.t. σ are upper-bounded respectively by L_1, L_2, L_3, $\forall|\sigma| \leq 1-\delta$. Furthermore, the mild regularity property that $\bar{\ell}_{jk}(\sigma;\Theta^*)$ not have degenerate critical points in $[-1+\delta, 1-\delta]$ holds.

(C-1) is the standard incoherence assumption that is made for the guarantees of glasso estimator (Ravikumar et al., 2011). **(C-2)** is a mild condition which ensures that no two latent variables are perfectly collinear and all the categories of the ordinal variables have non zero probabilities.

Theorem 4. *Consider our estimator (11) for solving a latent Gaussian model with true parameter Σ^*. And suppose conditions (C-1)-(C-3) are satisfied. Then there exist some known quantities c_1, c_2 and c_3 depending on L_1, L_2, L_3, M, α, γ, δ, K_{Σ^*} and K_{Γ^*} such that if $\lambda_n = c_1\sqrt{\log p'/n}$ and n is lower bounded as $n \geq c_2 d^2 \log p'$ where $p' = \max\{n,p\}$, then the inverse of estimate $\widehat{\Sigma}$ satisfies the following bound*

$$\left\|\widehat{\Sigma}^{-1} - (\Sigma^*)^{-1}\right\|_\infty \leq c_3\sqrt{\frac{\log p'}{n}} \tag{12}$$

and, moreover, the graph structure of latent Gaussian encoded in $(\Sigma^*)^{-1}$ is consistently recovered by $\widehat{\Sigma}^{-1}$, as long as $\beta_{\min} := \min_{ij}\left|[(\Sigma^*)^{-1}]_{ij}\right| \geq c_3\sqrt{\frac{\log p'}{n}}$, with probability at least $1-1/p \to 1$.

Proof Sketch: The proof of the theorem involves two main steps. In the first step we show that our estimate $\tilde{\Sigma}$ from step 1 satisfies: $\sup_{j,k}|\tilde{\Sigma}_{jk} - \Sigma^*_{jk}| \leq c_3\sqrt{\frac{\log p'}{n}}$, with high probability. Next, we utilize the consistency properties of glasso (Ravikumar et al., 2008) to show that our estimate $\widehat{\Sigma}$ from step 2 satisfies Equation (12) with high probability. To bound $\sup_{j,k}|\tilde{\Sigma}_{jk} - \Sigma^*_{jk}|$, we utilize recent results of Mei et al. (2017), who study the properties of the stationary points of non-convex empirical risk minimization problems. See Appendix C for the proof of the Theorem.

4. Experiments

In this section we present the performance of Consecutive Ratio model (*Consec model*), and our estimator for probit model (*ProbitDirect*) on various synthetic and real world datasets.

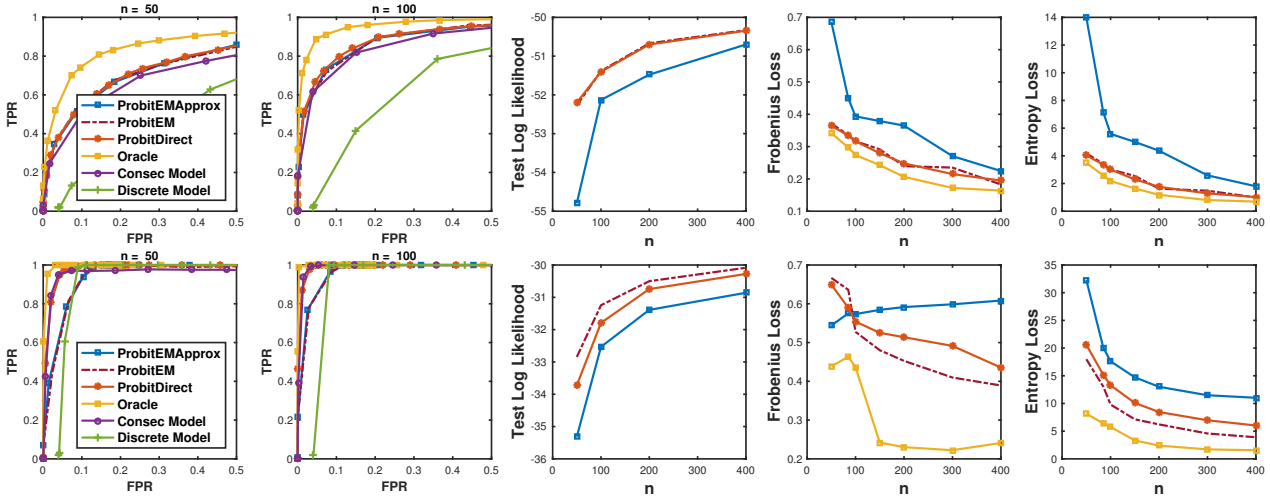
4.1. Synthetic Experiments

Baselines: In the synthetic experiments we compare our estimators with the following estimators:

a) *ProbitEM:* EM+MCMC based estimator for probit model described in Guo et al. (2015). This uses MCMC sampling to approximate the E-step.
b) *ProbitEMApprox:* Approximate EM based estimator for probit model proposed by Guo et al. (2015). This approach uses mean field approximation in the E-step to speed up the EM procedure.
c) *Discrete model:* This model ignores the order in the categories and treats each ordinal variable as a nominal variable. Here we restrict ourselves to a pairwise model. To learn this model we use the pseudo-likelihood based estimator of Jalali et al. (2011).
d) *Oracle:* When the data is generated from a probit model, we also compare with an oracle estimator which has access to the latent variables of the model. Here we run glasso on the latent variables to estimate the graph structure.

Model Selection: The best tuning parameter for all the estimators described above is selected using 5 fold cross validation. For *ProbitEMApprox* we use the cross validation technique proposed in (Guo et al., 2015). For *ProbitEM*, *Discrete model*, *Consec model* and *Oracle* we use the standard cross validation technique where we pick the best tuning parameter based on log likelihood on validation set. For *ProbitDirect* we use the following k-fold cross validation technique. We partition the data set into k subsets. Let $\widehat{\Sigma}_{-i}$ be the covariance matrix output by Step 2 of *ProbitDirect*, when trained using all the subsets except i^{th} subset. And let $\tilde{\Sigma}_i$ be the raw estimate obtained from Step 1 of *ProbitDirect*, using i^{th} subset. We pick a λ which maximizes the following score: $\sum_{i=1}^k \log\det(\widehat{\Sigma}^{-1}_{-i}) - \langle\langle\widehat{\Sigma}^{-1}_{-i}, \tilde{\Sigma}_i\rangle\rangle$.

Figure 1. Comparison of various estimators when the data is generated from a probit model with chain graph structure. Top and bottom rows correspond to $\omega = -0.3$ and $\omega = -0.9$ respectively. The left two columns show ROC curves for $n = 50, 100$. The right three columns show performance on test log likelihood, frobenius and entropy losses.



Evaluation Metrics: We compare the performance of our estimators against baselines, on graph structure recovery, using ROC curves computed by varying the regularization parameter. When the data is generated from a probit model, we compare the parameter estimation performance of *Oracle*, *ProbitEM*, *ProbitEMApprox* and *ProbitDirect* using Frobenius Loss and Entropy Loss which are defined as: Frobenius Loss $= \frac{\|\Sigma^{*-1} - \widehat{\Sigma}^{-1}\|_F}{\|\Sigma^{*-1}\|_F}$, Entropy Loss $= \langle\langle \Sigma^*, \widehat{\Sigma}^{-1} \rangle\rangle - \log\det(\Sigma^* \widehat{\Sigma}^{-1}) - p$, where Σ^* is the true covariance matrix and $\widehat{\Sigma}$ is the estimated covariance matrix. Finally, we also compare *ProbitEM*, *ProbitEMApprox*, *ProbitDirect* on log likelihood computed on 500 test samples (we do not compare with *Discrete model*, *Consec model* because computation of likelihood for these models is infeasible). For comparison on these three metrics, we select the best tuning parameter for each of the methods using cross validation.

Experimental Settings: In all our experiments we fix the number of nodes in the graph to 50 and set number of categories for each ordinal variable to 5. To reduce the variance, we average results over 10 repetitions.

4.1.1. Data from Probit Model

In our first experiment we generate ordinal data from a probit model. We simulate data from a chain graph. The inverse covariance matrix of the latent variables is chosen as follows:

$$\Sigma^{-1}_{j,k} = \begin{cases} \omega^{|j-k|} & \text{if } |j-k| \leq 1 \\ 0 & \text{otherwise} \end{cases}. \tag{13}$$

We pick an $\omega \in (-1, 1)$ in our experiments and set the thresholds (θ) at node j as : $\theta^{(j)} = [-\text{Inf}, -10, -0.7, 0.7, 10, \text{Inf}]$. Finally we scale the covariance matrix so that all the variances are equal to 1. Figure 1 shows the results obtained using $\omega = -0.3, -0.9$. More results for large n and grid and random graphs can be found in Appendix D. We can seen that *ProbitDirect* and *ProbitEM* have similar performance. However, *ProbitDirect* is 1-2 orders of magnitude faster than *ProbitEM*. *ProbitEMApprox* has very poor performance, especially in low sample complexity setting. This could be because of the mean field approximation that is made by Guo et al. (2015), where in the E-step they approximate $\mathbb{E}[Z_j Z_k | Y; \widehat{\Theta}, \widehat{\Sigma}]$ as $\mathbb{E}[Z_j | Y; \widehat{\Theta}, \widehat{\Sigma}] \times \mathbb{E}[Z_k | Y; \widehat{\Theta}, \widehat{\Sigma}]$. This decouples the interactions between any two random variables. Also note that *ProbitDirect* is ≈ 5 times faster than *ProbitEMApprox* (See Table 1 in Appendix).

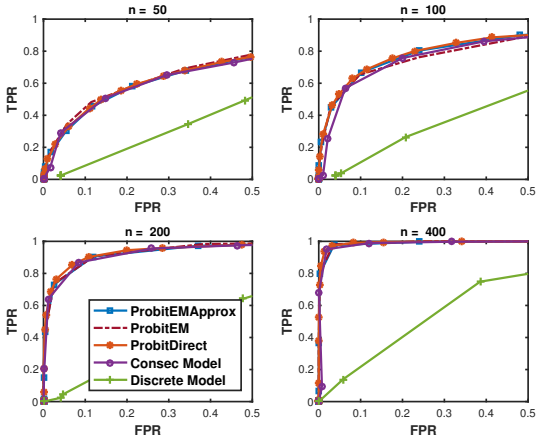


Figure 2. Data sampled from a Consec model with 2D grid structure (10×5 grid). Node specific parameters (θ_s) are uniformly sampled from $[-1, 1]$. Pairwise interaction terms (θ_{st}) are set to 0.1 for all horizontal edges and to -0.1 for all vertical edges.

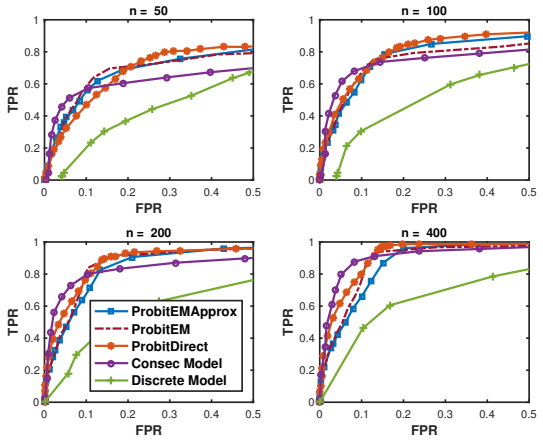


Figure 3. Data sampled from a Consec model with 2D grid structure (10×5 grid). Node specific parameters (θ_s) are uniformly sampled from $[-1, 1]$. Pairwise interaction terms (θ_{st}) are set to 0.3 for all horizontal edges and to -0.3 for all vertical edges.

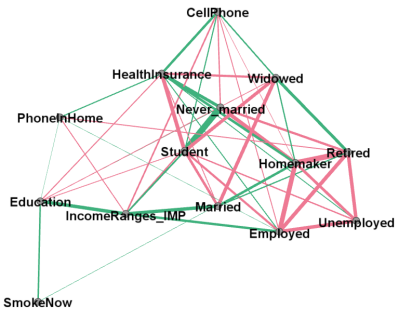


Figure 4. The estimated latent graph structure corresponding to *SmokeNow* and *sociodemographic* indicators. The graph is generated from the marginal distribution of the corresponding variables. Green and red edges represent positive and negative partial correlations respectively. Edge thickness is proportional to the magnitude of the partial correlation.

4.1.2. Data from Consecutive Ratio Model

In this experiment we sample data from *Consec model*. Figures 2, 3 present the results on a grid graph along with the details of exact parameters used. When the interaction between variables is low, *Consec model* has similar performance as other estimators (Figure 2). However, when the interactions are high (Figure 3), its performance degrades. We noticed similar poor performance of *Consec model* on other graphs. This could suggest that either the node conditional likelihood based estimator for Consecutive Ratio model is not efficient or latent graphical models such as Probit model are better models than Consecutive Ratio model.

4.2. Health Information National Trends Survey study

The Health Information National Trends Survey (HINTS) is a nationally representative survey conducted by the National Cancer Institute (NCI) every few years. In this work we analyze HINTS-FDA data which is a special data collected by NCI in partnership with the Food and Drug Administration (FDA) and is made publicly available by NCI. This survey collected detailed information on the following topics: *Tobacco Product Use, Beliefs about Tobacco Products, Beliefs About Cancer, How people access Health Information?, Socio Demographic Indicators*. The survey has 3738 respondents on ≈ 350 questions. Almost all the questions in the survey have either ordinal or categorical responses. Refer to Appendix D for summary statistics of the data and details about preprocessing performed on the data. We treat each question in the survey as a node in the graph and responses of individuals to these questions as samples drawn from the graph. We selected 114 questions from the dataset, that are relevant for our analysis. We fit the probit model using *ProbitDirect* on the selected questions. To choose the best tuning parameter we use 10 fold cross validation. We obtain 95% confidence intervals for the edge strengths of the latent graph through jackknife resampling technique. We place an edge in the graph only if its confidence interval doesn't intersect with $[-0.1, 0.1]$.

Figure 4 shows how various variables related to *sociodemographic* indicators are associated with smoking behavior of a person. In particular, it shows that *SmokeNow* has a very significant association with *Education*. This indicates that if a person is well educated then conditioned on all the other variables, there is lower chance that the person smokes. In Appendix D, we provide an additional Figure 9, which shows how the perceptions of smoking risks vary with smoking behavior. It shows that *SmokeNow* and *FewCigarettesHarmHealth* have a positive partial correlation, indicating that conditioned on the rest of the variables, people who smoke, perceive smoking as less harmful than people who don't smoke. Table 2 in Appendix D describes some highly relevant nodes in the learned graphs, from which several other insights can be obtained. We believe these insights can be helpful in designing efficient strategies for communicating smoking related health information to the public.

5. Acknowledgments

E.Y. acknowledges the support of MSIP/NRF (National Research Foundation of Korea) via NRF-2016R1A5A1012966 and MSIP/IITP (Institute for Information & Communications Technology Promotion of Korea) via ICT R&D program 2016-0-00563, 2017-0-00537. A.S., P.R. acknowledge the support of ARO via W911NF-12-1-0390 and NSF via IIS-1149803, IIS-1447574, DMS-1264033, and NIH via R01 GM117594-01.

References

Agresti, A. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2010.

Allen, G. I. and Liu, Z. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pp. 1–6, 2012.

Amemiya, T. Bivariate probit analysis: Minimum chi-square methods. *Journal of the American Statistical Association*, 69(348):940–944, 1974.

Armstrong, B. and Sloan, M. Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, 129(1):191–204, 1989.

Ashford, J. R. and Sowden, R. R. Multi-variate probit analysis. *Biometrics*, 26(3):535–546, 1970.

Bartolucci, F., Colombi, R., and Forcina, A. An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica*, pp. 691–711, 2007.

Becker, W. E. and Kennedy, P. E. A graphical exposition of the ordered probit. *Econometric theory*, 8(01):127–131, 1992.

Besag, J. Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Stat. Soc. Series B*, 36:192–236, 1974.

Cai, T., Liu, W., and Luo, X. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Chib, S. and Greenberg, E. Analysis of multivariate probit models. *Biometrika*, pp. 347–361, 1998.

De Leon, A. R. Pairwise likelihood approach to grouped continuous model and its extension. *Statistics & probability letters*, 75(1):49–57, 2005.

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

Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. Graphical models for ordinal data. *Journal of Computational and Graphical Statistics*, 24(1):183–204, 2015.

Han, F. and Pan, W. A composite likelihood approach to latent multivariate gaussian modeling of snp data with application to genetic association testing. *Biometrics*, 68 (1):307–315, 2012.

Ising, E. Beitrag zur theorie der ferromagnetismus. *Zeitschrift für Physik*, 31:253–258, 1925.

Jalali, A., Ravikumar, P., Vasuki, V., and Sanghavi, S. On learning discrete graphical models using group-sparse regularization. In *Inter. Conf. on AI and Statistics (AISTATS)*, 14, 2011.

Jöreskog, K. G. On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, 59(3):381–389, 1994.

Lauritzen, S. L. and Wermuth, N. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, pp. 31–57, 1989.

Lindsay, Bruce G. Composite likelihood methods. *Statistical Inference from Stochastic Processes Contemporary Mathematics*, pp. 221239, 1988. doi: 10.1090/conm/080/999014.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. The nonparanormal skeptic. In *International Conference on Machine learning (ICML)*, 2012.

Manning, C. D. and Schütze, H. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

McCullagh, P. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

Mei, S., Bai, Y., and Montanari, A. The landscape of empirical risk for non-convex losses. *Arxiv preprint arXiv:1607.06534*, 2017.

Muthén, B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1):115–132, 1984.

Peterson, B. and Harrell Jr., F. E. Partial proportional odds models for ordinal response variables. *Applied statistics*, pp. 205–217, 1990.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Technical Report 767, Department of Statistics, UC Berkeley, September 2008.

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

Rue, H. and Held, L. *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, 2005.

Speed, T. P. and Kiiveri, H. T. Gaussian Markov distributions over finite graphs. *Annals of Statistics*, 14(1): 138–150, March 1986.

Tansey, W., Padilla, O. H. M., Suggala, A. S., and Ravikumar, P. Vector-space markov random fields via exponential families. In *ICML*, pp. 684–692, 2015.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. Graphical models via generalized linear models. In *Neur. Info. Proc. Sys. (NIPS)*, 25, 2012.

Yang, E., Baker, Y., Ravikumar, P., Allen, G. I., and Liu, Z. Mixed graphical models via exponential families. In *Inter. Conf. on AI and Statistics (AISTATS)*, 17, 2014.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16:3813–3847, 2015.

Yuan, M. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286, 2010.