Learning Exponential Family Graphical Models with Latent Variables using Regularized Conditional Likelihood

Armeen Taeb [a], Parikshit Shah [b], and Venkat Chandrasekaran [c] *

[a] Seminar for Statistics, ETH Zürich
[b] Wisconsin Institutes for Discovery at the University of Wisconsin, Madison
[c] Department of Computing and Mathematical Sciences & of Electrical Engineering, Caltech

October 19, 2020

Abstract

Fitting a graphical model to a collection of random variables given sample observations is a challenging task if the observed variables are influenced by latent variables, which can induce significant confounding statistical dependencies among the observed variables. We present a new convex relaxation framework based on regularized conditional likelihood for latent-variable graphical modeling in which the conditional distribution of the observed variables conditioned on the latent variables is given by an exponential family graphical model. In comparison to previously proposed tractable methods that proceed by characterizing the marginal distribution of the observed variables, our approach is applicable in a broader range of settings as it does not require knowledge about the specific form of distribution of the latent variables and it can be specialized to yield tractable approaches to problems in which the observed data are not well-modeled as Gaussian. We demonstrate the utility and flexibility of our framework via a series of numerical experiments on synthetic as well as real data.

keywords: convex optimization, equivariant estimators, exponential family PCA, pseudolikelihood, semidefinite programming

1 Introduction

Graphical models are multivariate statistical models that provide compact descriptions of joint probability distributions over large collections of variables in terms of products of local compatibility functions, each of which only involves a small number of the variables. We consider an *exponential family* of graphical models in d variables in which the associated distributions factor as a product of functions of one or two variables (see [23] and the references therein) over a domain $\mathcal{X}^d \subseteq \mathbb{R}^d$ with ancillary statistic h:

$$\begin{aligned} f(x;\alpha,\Theta) &\triangleq h(x)\exp\left\{\alpha' x - \tfrac{1}{2}x'\Theta x - \Phi(\alpha,\Theta)\right\} \\ \Phi(\alpha,\Theta) &\triangleq \int_{\mathcal{X}^d} \exp\left\{\alpha' x - \tfrac{1}{2}x'\Theta x\right\} h(x) d\nu(x). \end{aligned} \tag{1.1}$$

In the examples in this paper ν is either the Lebesgue measure or the counting measure (the integral in the definition of Φ is a sum if ν is the counting measure), and the product $h(x)d\nu(x)$ is also called the *base measure*. The parameters $\alpha \in \mathbb{R}^d$ and $\Theta \in \mathbb{S}^d$ are the *natural parameters* of the family, with $\mathbb{S}^d$ denoting the space of $d \times d$ real symmetric matrices. The function $\Phi(\alpha,\Theta)$ is called the *log-partition function* and it serves to normalize f. The set of valid values for the parameters α, Θ are those for which the log-partition is finite:

$$\mathcal{F} \triangleq \left\{(\alpha,\Theta) \in \mathbb{R}^d \times \mathbb{S}^d \mid \Phi(\alpha,\Theta) < \infty\right\}. \tag{1.2}$$

*Correspondence email: armeen.taeb@stat.math.ethz.ch

The set of *valid parameters* $\mathcal{F}$ is a convex subset of $\mathbb{R}^d \times \mathbb{S}^d$, and over the domain $\mathcal{F}$ the log-partition function Φ is convex. The number of parameters required to specify the distribution $f(x; \alpha, \Theta)$ is $\mathcal{O}(d^2)$, which can be prohibitive in problems with a large number of variables d. Consequently, models in which Θ is a *sparse* matrix are of great interest in applications. Such sparse graphical models also have an appealing statistical interpretation as follows. Given a distribution of the form (1.1), one can associate to it a graph consisting of d nodes and edges between those pairs of nodes for which the corresponding $\Theta_{i,j} \neq 0$. The Hammersley-Clifford theorem states that the random variables x_i, x_j at two distinct nodes i, j are independent conditioned on variables at all the other nodes if there is no edge between the nodes i and j, i.e., there is no path between nodes i and j that does not pass through another node. In this manner, the graph corresponding to the sparsity pattern of the matrix Θ encodes the conditional independence or Markov relations underlying the variables.

Several variations of this basic family are possible, such as different types of compatibility functions or compatibility functions consisting of larger subsets of variables. Our discussion can accommodate these extensions, but we stick with the model (1.1) for notational simplicity. Graphical models with random variables that are Gaussian ($\mathcal{X} = \mathbb{R}$) and Bernoulli ($\mathcal{X} = \{-1, +1\}$) are prominent examples of (1.1), and these are respectively called Gaussian graphical models and Ising models. Further, graphical models that are appropriate for data specifying counts ($\mathcal{X} = \mathbb{Z}_+$) or data taking on positive values ($\mathcal{X} = \mathbb{R}_+$) may also be obtained as special cases of (1.1) [3, 24]; see Section 2 for details.

In data analysis problems in which the variables are indexed in an ordered fashion, there are usually reasonable choices for the underlying graph structure; for example, graphs based on nearest-neighbors are often used for specifying time series and spatial models. However, in many applications, a natural choice for the graph structure is not available due to a lack of domain knowledge about the underlying conditional independence relations, and it is of interest to identify a sparse graphical model from sample observations of a collection of variables. A challenge with this task is that there may be latent variables for which it is expensive or impossible to obtain sample observations. Such unobserved variables pose a significant difficulty as graphical model structure is not closed under marginalization; therefore, the edge structure corresponding to the *conditional* distribution of a collection of observed variables conditioned on latent variables is in general different from the *marginal* distribution of the observed variables. The graphical model of the observed variables conditioned on the latent variables signifies those statistical dependencies that are in some sense intrinsic to the observed variables, while the marginal graphical model of the observed variables consists of confounding dependencies that are induced due to marginalization over the latent variables, and this model typically consists of many more edges than the conditional graphical model. In fact, even if the conditional graphical model is compactly described as a product of a small number of pairwise compatibility functions, the marginal model is in general much more complicated as it can consist of higher-order compatibility functions that link together large subsets of the observed variables; see Figure 1.

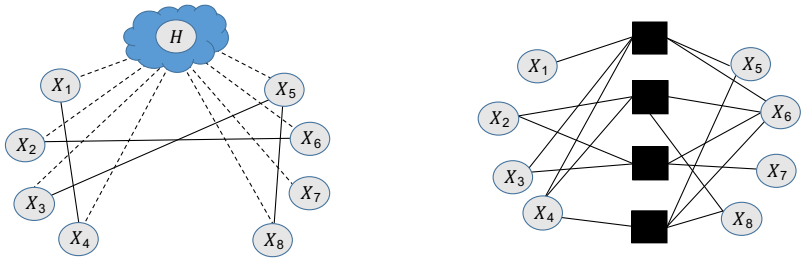


Figure 1: An example of a graphical model over 8 variables $X_1, \ldots, X_8$ — left: the variable H represents an unobserved quantity, solid edges indicate pairwise interactions among observed variables, and dashed edges indicate links between observed and latent variables; right: factor graph where black squares represent higher-order interactions among variables that are linked through the factors.

The problem of learning a graphical model even without latent variables is computationally intractable in general, and accounting for the confounding effects of latent variables is more challenging. There are a number of previous papers that propose computationally efficient approaches based both on combinatorial techniques [5, 6, 11] and on convex relaxation [7, 18] for learning graphical models with latent variables, along with theoretical or empirical demonstrations of the utility of these approaches for particular families of problem instances. These methods proceed by studying the marginal distribution of the observed variables, and their derivation is based on an analysis of the structure of the confounding dependencies among the observed variables induced due to marginalization over the latent variables. Consequently, the development of each of these methods is reliant on assumptions about the

form of the joint distribution of the observed and latent variables – jointly Gaussian observed and latent variables in [7], an Ising model specifying the observed and latent variables in [5, 6, 11], and a conditionally Ising model for the observed variables with Gaussian latent variables in [18].

In this paper we present a new convex relaxation framework for latent-variable graphical model selection in which the conditional graphical model of the observed variables conditioned on the latent variables belongs to an exponential family of the form (1.1). The virtues of convexity are by now self-evident – tractable convex programs are manifestly implementable for moderate-sized problem instances based on off-the-shelf software packages, and in many cases, it is possible to develop special-purpose solvers that can scale to large instances by exploiting problem structure. Perhaps the biggest conceptual distinction in our approach compared to those mentioned in the preceding paragraph is an analysis of the structure of the conditional graphical model of the observed variables conditioned on latent variables rather than the marginal distribution of the observed variables (which as explained previously can be very complicated in general). Our viewpoint leads to tractable convex relaxations for a far broader class of models than those considered in previous work. In particular, the derivation of our method does not require knowledge about the specific form of the distribution of the latent variables. In addition, our framework can also be specialized to settings in which the observed variables are not well-modeled as conditionally Gaussian or Bernoulli (such as with data specifying counts or taking on only positive values). In both these respects, our methodology is more broadly applicable than those described in the papers referenced above.

1.1 Our Contributions

We consider latent-variable graphical models in which a small number of latent variables $z \in \mathbb{R}^r$ influence the observed variables $x \in \mathcal{X}^d$, i.e., $r \ll d$, with the following form for the conditional distribution of $x|z$:

$$f(x|z;\alpha,\Theta,B) \triangleq \exp\left\{(\alpha + Bz)'x - \tfrac{1}{2}x'\Theta x - \Phi(\alpha + Bz,\Theta)\right\} h(x). \tag{1.3}$$

In words, the latent variables influence the natural parameters associated to the node compatibility functions in an affine manner specified by the parameter $B \in \mathbb{R}^{d\times r}$. The form of this conditional distribution is akin to a generalized linear model with x corresponding to the responses and z the covariates, although there are two substantive differences – one is that we do not observe z and the other is that the different components of x are not independent of each other in general after conditioning on z (unless $\Theta_{i,j} = 0,\ \forall i \neq j$) [23]. The form of the conditional distribution (1.3) may also be interpreted as a conditional random field, but again with the distinction that we do not observe z [24]. The model (1.3) encompasses settings in which the joint distribution of the observed variables and the latent variables is given by a pairwise graphical model – such as a Gaussian graphical model or an Ising model jointly over the observed and latent variables – although our setup is more general as we do not assume a specific form for the distribution of the latent variables. It is important to note that even if the conditional graphical model associated to $x|z$ is sparse (i.e., Θ is a sparse matrix), the marginal distribution of x is in general dense depending on the distribution of z; more significantly, the conditional distribution of $x|z$ factorizes as a product of local functions (each depending on one or two variables), but the marginal distribution of x may not be factorizable in such a local manner as the effect of marginalization over z can induce confounding effects that couple all the components of x.

Our objective is to fit a latent-variable graphical model to a sample $\{x^{(k)}\}_{k=1}^n$ of size n of the observed variables. We assume that a selection of an exponential family that is best-suited to the data has been made (i.e., an appropriate choice of $\mathcal{X}$ and h, which in turn induce Φ and $\mathcal{F}$), and we consider the following regularized conditional likelihood optimization problem given user-specified regularization parameters $\lambda, \gamma \geq 0$:

$$\begin{aligned}(\widehat{\alpha},\widehat{\Theta},\widehat{L}) = \underset{\substack{\alpha\in\mathbb{R}^d,\Theta\in\mathbb{S}^d\\ L\in\mathbb{R}^{d\times n}}}{\arg\min}\ & \frac{1}{n}\sum_{k=1}^{n}\left[\Phi(\alpha + L^{(k)},\Theta) - (\alpha + L^{(k)})'x^{(k)} + \tfrac{1}{2}x^{(k)\prime}\Theta x^{(k)}\right] + \lambda\|\Theta\|_{\ell_1} + \gamma\|L\|_\star \\ & \text{s.t.}\,(\alpha + L^{(k)},\Theta) \in \mathcal{F}, \quad k = 1,\ldots,n.\end{aligned} \tag{1.4}$$

The first part of the objective function represents the negative-logarithm of the conditional likelihood, with the vector $L^{(k)} \in \mathbb{R}^d$ denoting the k'th column of $L \in \mathbb{R}^{d\times n}$ and playing the role of $Bz^{(k)}$. Both B and $\{z^{(k)}\}_{k=1}^n$ are unknown, but the matrix with columns given by the elements of the set $\{Bz^{(k)}\}_{k=1}^n$ has small rank if the

dimension r of the latent vector satisfies $r \ll d$; the nuclear norm penalty $\|L\|_\star$ in the second line of the objective function is intended to promote this low-rank structure. The ℓ_1 penalty on Θ is useful for promoting sparsity of the conditional graphical model of the observed variables conditioned on the latent variables. In summary, the optimization problem (1.4) is a convex program, although the log-partition function Φ may be intractable to compute in some cases (e.g., the conditional graphical model is an Ising model). In such situations, one can appeal to further approximations from the literature which continue to preserve the convexity of the problem [3, 23]; see Section 2.2. Finally, if we constrain the off-diagonal entries of the decision variable Θ in (1.3) to be zero, then we obtain a convex relaxation for an exponential-family generalization of principal components analysis (PCA) [9] in which one wishes to fit a model in which the different components of x are independent of each other after conditioning on z (i.e., $\Theta_{i,j} = 0,\ \forall i \neq j$).

In Section 2 we specialize the formulation (1.4) to obtain computationally tractable methods for latent-variable graphical modeling for a range of exponential family graphical models, as well as a symmetry reduction of (1.4) for Gaussian models based on equivariance, which yields a convex program involving $d \times d$ matrices; in contrast, the complexity of solving (1.4) scales with the number of observations n for general models. In Section 3 we present a technique for selecting suitable regularization parameters λ, γ (1.4). In each of these preceding sections, we provide evidence for the effectiveness of our framework via experiments on synthetic data. In Section 4, we demonstrate the performance of our methods on real data. Finally, we conclude with a discussion of future directions in Section 5.

The focus of this paper is on developing a mathematically principled and broadly applicable methodology for latent-variable graphical modeling. We demonstrate the utility and flexibility of our framework empirically on synthetic and real data. We describe a number of questions for future work that concern theoretical analysis of various aspects of our approach in Section 5.

1.2 Notation

We denote the identity matrix by I, with the size being clear from context. The collection of positive-semidefinite matrices in $\mathbb{S}^d$ is denoted $\mathbb{S}^d_+$ and the collection of positive-definite matrices by $\mathbb{S}^d_{++}$.

2 Specializations of Our Framework

We describe specializations of our framework to various exponential family models. Section 2.1 concerns Gaussian models in which the log-partition function can be computed efficiently, and therefore our proposed method is simply the specialization of the convex relaxation (1.4) to the Gaussian case. Furthermore, we show that one can equivalently reformulate the Gaussian specialization of (1.4) as an SDP involving only $d \times d$ matrix decision variables, so that no decision variable has a dimension that scales with n. Section 2.2 concerns several non-Gaussian models for which the likelihood is intractable to compute in general, and therefore computationally efficient approximations of (1.4) are required; we describe one such approach using the pseudo-likelihood approximation of Besag [3]. In both subsections, we discuss how previous approaches for graphical modeling without latent variables and for exponential family PCA may be obtained by restricting our relaxations appropriately. Finally, we present numerical evidence for the effectiveness of our methods in Section 2.3.

2.1 Gaussian Models

Multivariate Gaussians constitute an exponential family of distributions with ancillary statistic $h \equiv 1$ and ν being the Lebesgue measure. The parameter Θ is the precision or inverse covariance matrix and the mean is given by $\Theta^{-1}\alpha$. The corresponding log-partition function and valid parameters are given by:

$$\begin{aligned} \Phi_{\text{gaussian}}(\alpha, \Theta) &= \tfrac{1}{2}\left(\alpha' \Theta^{-1} \alpha - \log\det\Theta + d\log 2\pi\right) \\ \mathcal{F}_{\text{gaussian}} &= \{(\alpha, \Theta) \in \mathbb{R}^d \times \mathbb{S}^d \mid \Theta \succ 0\}. \end{aligned} \tag{2.1}$$

Thus, we obtain the following convex relaxation for latent-variable Gaussian graphical modeling given data $\{x^{(k)}\}_{k=1}^n \subset \mathbb{R}^d$ and user-specified regularization parameters $\lambda, \gamma \geq 0$:

$$\begin{aligned}(\widehat{\alpha}, \widehat{\Theta}, \widehat{L}) = \underset{\substack{\alpha \in \mathbb{R}^d, \Theta \in \mathbb{S}^d \\ L \in \mathbb{R}^{d \times n}}}{\arg\min} \; & \frac{1}{n} \sum_{k=1}^n \left[\tfrac{1}{2} (\alpha + L^{(k)})' \Theta^{-1} (\alpha + L^{(k)}) - (\alpha + L^{(k)})' x^{(k)} + \tfrac{1}{2} x^{(k)\prime} \Theta x^{(k)} \right] \\ & - \frac{1}{2} \log\det \Theta + \lambda \|\Theta\|_{\ell_1} + \gamma \|L\|_\star \\ \text{s.t. } & \Theta \succ 0. \end{aligned} \tag{2.2}$$

Sublevel sets of the function $(\alpha + L^{(k)})' \Theta^{-1} (\alpha + L^{(k)})$ may be expressed via Schur complements, and therefore, this problem is a log-determinant semidefinite program (SDP) that can be solved to a desired precision in polynomial-time.

If we do not account for the confounding effects of latent variables and set $L = 0$, then the convex program (2.2) specializes to the well-known 'Graphical Lasso' method [2, 10, 25], which corresponds to ℓ_1-regularized marginal log-likelihood. On the other hand, if we restrict Θ to be a diagonal matrix we recover a convex relaxation for factor analysis [20].

The dimension of the decision variable L in the convex relaxation (2.2) grows with the number of observations n. Consequently, the computational runtime to solve (2.2) to a desired accuracy scales polynomially with n. We exploit an equivariance property underlying the estimator (2.2) in the Gaussian case to obtain an equivalent relaxation (after preprocessing) in which the dimensions of the decision variables do not depend on n. (The main component of the preprocessing step is a singular value decomposition of a $d \times n$ matrix, but the complexity of this operation scales more modestly with n than that of solving (2.2).) For ease of analysis, we set $\alpha = 0$ (2.2); this restriction may be made with no loss of generality if we center the observations prior to solving (2.2).

Formally, letting $X \in \mathbb{R}^{d \times n}$ denote a data matrix with the observations $\{x^{(k)}\}_{k=1}^n \subset \mathbb{R}^d$ specifying the columns, the objective (2.2) may be written as follows:

$$c(\Theta, L; X) = \tfrac{1}{2n} \mathrm{tr}(L' \Theta^{-1} L) - \tfrac{1}{2} \log\det(\Theta) - \tfrac{1}{n} \mathrm{tr}(L' X) + \tfrac{1}{2n} \mathrm{tr}(\Theta X X') + \lambda \|\Theta\|_{\ell_1} + \gamma \|L\|_\star. \tag{2.3}$$

A key attribute of this expression is that for any matrix $W \in \mathbb{R}^{m \times n}, m \geq n$, satisfying $W'W = I$, one can check that:

$$c(\Theta, LW'; XW') = c(\Theta, L; X). \tag{2.4}$$

The dimensions of the inputs of c here are, in general, different on the left-hand-side versus the right-hand-side, but the expression (2.3) remains valid as long as the dimensions of the inputs/parameters to c are consistent; we allow for this flexibility in our discussion in the sequel. Hence, if the data matrix X is transformed as $X \leftarrow XW'$ then the expression (2.3) remains unchanged with an analogous transformation $L \leftarrow LW'$ applied to the decision variable L (and leaving Θ unchanged). This *equivariance* property enables a reduction in the size of the SDP (2.2), which we formalize via the following result:

Theorem 1. *Given a data matrix $X \in \mathbb{R}^{d \times n}$, let $\Sigma = \frac{1}{n} X X'$ denote the sample covariance matrix and consider the following optimization problem*[1] *with $\lambda, \gamma \geq 0$:*

$$(\widehat{\Theta}, \widehat{H}) = \underset{\substack{\Theta \in \mathbb{S}^d_{++} \\ H \in \mathbb{R}^{d \times d}}}{\arg\min} \; \tfrac{1}{2} \mathrm{tr}(H' \Theta^{-1} H) - \tfrac{1}{2} \log\det \Theta - \mathrm{tr}(H' \sqrt{\Sigma}) + \tfrac{1}{2} \mathrm{tr}(\Theta \Sigma) + \lambda \|\Theta\|_{\ell_1} + \gamma \sqrt{n} \|H\|_\star \tag{2.5}$$

Here $\sqrt{\Sigma}$ denotes the positive-semidefinite square root of Σ. Let $X = UDV'$ be a singular value decomposition of X. Then $(\widehat{\Theta}, \sqrt{n} \widehat{H} U V')$ is an optimal solution of (2.2) with $\alpha = 0$.

Proof. Consider the case $n \leq d$ in which $U \in \mathbb{R}^{d \times n}, V \in \mathbb{R}^{n \times n}$. One can check that:

$$c(\Theta, L; X) = c(\Theta, LVU'; XVU') = c(\Theta, LVU'; \sqrt{n} \sqrt{\Sigma})$$

[1] To ensure that this problem has an optimal solution, it suffices to choose $\lambda > 0$ or to have $\Sigma \succ 0$.

The first equality follows from the equivariance relation (2.4) and the second equality follows from the definition of Σ. Setting $H = \frac{1}{\sqrt{n}} LVU'$, the expression $c(\Theta, LVU'; \sqrt{n}\sqrt{\Sigma})$ is equal to the objective of (2.5). Hence, a feasible (Θ, L) for (2.2) with $\alpha = 0$ leads to a feasible point (Θ, H) for (2.5) with equal cost. In the other direction, consider a feasible (Θ, H) for (2.5), and set $L = \sqrt{n} HUV'$. With this (Θ, L), we consider the three terms of $c(\Theta, L)$ from (2.3) involving L. First, we note using cyclicity of trace that:

$$\tfrac{1}{2n}\mathrm{tr}(L'\Theta^{-1}L) = \tfrac{1}{2}\mathrm{tr}(VU'H'\Theta^{-1}HUV') = \tfrac{1}{2}\mathrm{tr}(H'\Theta^{-1}HUU') \leq \tfrac{1}{2}\mathrm{tr}(H'\Theta^{-1}H),$$

which follows from $V'V = I$ and $I \succeq UU'$. Next, we have that:

$$\tfrac{1}{n}\mathrm{tr}(L'X) = \tfrac{1}{\sqrt{n}}\mathrm{tr}(VU'H'X) = \tfrac{1}{\sqrt{n}}\mathrm{tr}(H'XVU') = \mathrm{tr}(H'\sqrt{\Sigma})$$

using the definition of Σ and the cyclicity of trace. Finally, we observe that:

$$\gamma\|L\|_\star = \gamma\sqrt{n}\|HUV'\|_\star \leq \gamma\sqrt{n}\|H\|_\star\|UV'\|_2 \leq \gamma\sqrt{n}\|H\|_\star$$

by applying the Hölder inequality to the nuclear norm. Therefore, for any feasible (Θ, H) for (2.5), the point $(\Theta, \sqrt{n}HUV')$ is feasible for (2.2) with $\alpha = 0$ and has equal or lower cost.

Consider next the case $n > d$ in which $U \in \mathbb{R}^{d\times d}, V \in \mathbb{R}^{n\times d}$. Set $W = \begin{pmatrix} UV' \\ P \end{pmatrix} \in \mathbb{R}^{n\times n}$ with $P \in \mathbb{R}^{(n-d)\times n}$ so that $WW' = W'W = I$. One can then check that:

$$c(\Theta, L; X) = c(\Theta, LW'; XW') = c(\Theta, (LVU' \;\; LP'); (\sqrt{n}\sqrt{\Sigma} \;\; 0))$$

The first equality follows from the equivariance relation (2.4), and the second equality follows from the properties of W and the definition of Σ. Given a feasible (Θ, H) of (2.5) and setting $L = \sqrt{n}HUV'$, we observe that $LVU' = \sqrt{n}H$ and $LP' = 0$. Thus, with this choice of L the expression $c(\Theta, L; X) = c(\Theta, (\sqrt{n}H \;\; 0); (\sqrt{n}\sqrt{\Sigma} \;\; 0))$ equals the objective of (2.5). Hence, from a feasible point of (2.5), we obtain a feasible point of (2.2) with equal cost. In the other direction, let (Θ, L) be feasible for (2.2) and set $H = \frac{1}{\sqrt{n}} LVU'$. We consider the three terms of the objective from (2.5) involving H. First, we note using the cyclicity of trace that:

$$\tfrac{1}{2}\mathrm{tr}(H'\Theta^{-1}H) = \tfrac{1}{2n}\mathrm{tr}(UV'L'\Theta^{-1}LVU') = \tfrac{1}{2n}\mathrm{tr}(L'\Theta^{-1}LVV') \leq \tfrac{1}{2n}\mathrm{tr}(L'\Theta^{-1}L),$$

which follows from $U'U = I$ and $I \succeq VV'$. Next, we have that:

$$\mathrm{tr}(H'\sqrt{\Sigma}) = \tfrac{1}{\sqrt{n}}\mathrm{tr}(UV'L'\sqrt{\Sigma}) = \tfrac{1}{n}\mathrm{tr}(L'X),$$

using the definition of Σ and the cyclicity of trace. Finally, we observe that:

$$\gamma\sqrt{n}\|H\|_\star = \gamma\|LVU'\|_\star \leq \gamma\|L\|_\star\|VU'\|_2 \leq \gamma\|L\|_\star,$$

by applying the Hölder inequality to the nuclear norm. Thus, for any feasible (Θ, L) for (2.2) with $\alpha = 0$, the point $(\Theta, \frac{1}{\sqrt{n}}LVU')$ is feasible for (2.5) and has lower or equal cost. □

Although this result holds for arbitrary n, it is most relevant when $n \gg d$.

2.2 Non-Gaussian Models

We consider three exponential family graphical models that are relevant in settings in which the observations are not well-modeled as Gaussian. These are derived by considering pairwise graphical models in which the variable at each node conditioned on the variables at all the other nodes is distributed according to a Bernoulli, Poisson, or

exponential random variable [3, 24]; as such the following models represent natural multivariate generalizations of popular univariate exponential families:

Ising models Here $\mathcal{X} = \{-1, +1\}$, the ancillary statistic is $h \equiv 1$, and ν is the counting measure. As the log-partition function is given by a finite sum, the set of valid parameters is not constrained in a significant way:

$$\mathcal{F}_{\text{ising}} = \{(\alpha, \Theta) \in \mathbb{R}^d \times \mathbb{S}^d \mid \Theta_{i,i} = 0,\ i = 1, \ldots, d\}. \tag{2.6}$$

The condition on the diagonal elements of Θ is due to the fact $x_i^2 = 1$ for $x_i \in \{-1, +1\}$, and therefore the diagonal elements of Θ do not offer any degrees of freedom.

Poisson graphical models Here $\mathcal{X} = \mathbb{Z}_+$, the ancillary statistic is $h(x_1, \ldots, x_d) = \prod_{i=1}^d \frac{1}{x_i!}$, and ν is the counting measure. To ensure that each $x_i|x_{\setminus i}$ is distributed as a Poisson random variable, each $\Theta_{i,i} = 0$ for $i = 1, \ldots, d$. The set of valid parameters for which the associated distribution is normalizable is given by:

$$\mathcal{F}_{\text{poisson}} = \{(\alpha, \Theta) \in \mathbb{R}^d \times \mathbb{S}^d \mid \Theta_{i,j} \geq 0, i, j = 1, \ldots, d;\ \Theta_{i,i} = 0, i = 1, \ldots, d\}. \tag{2.7}$$

Exponential graphical models Here $\mathcal{X} = \mathbb{R}_+$, the ancillary statistic is $h \equiv 1$, and ν is the Lebesgue measure. To ensure that each $x_i|x_{\setminus i}$ is distributed as an exponential random variable, each $\Theta_{i,i} = 0$ for $i = 1, \ldots, d$. The set of valid parameters for which the associated distribution is normalizable is given by:

$$\begin{aligned}\mathcal{F}_{\text{exponential}} = \{(\alpha, \Theta) \in \mathbb{R}^d \times \mathbb{S}^d \mid \alpha_i < 0, i = 1, \ldots, d;\ \Theta_{i,j} \geq 0, i, j = 1, \ldots, d;\\ \Theta_{i,i} = 0, i = 1, \ldots, d\}.\end{aligned} \tag{2.8}$$

In each of these three cases, the log-partition is intractable to compute; for such situations, a number of convex approximations of the partition function that are tractable to compute are available in the literature (see [23] and the references therein), and these may be employed as surrogates (1.4) to obtain computationally efficient convex relaxations. In our numerical experiments in Sections 2.3 and 4, we use the following pseudo-likelihood approximation due to Besag [3]:

$$f(x_1, \ldots, x_d|z; \alpha, \Theta, B) \approx \prod_{i=1}^d f(x_i|x_{\setminus i}, z; \alpha, \Theta, B). \tag{2.9}$$

For exponential family distributions of the form (1.1), this approximation replaces partition functions associated to d-dimensional distributions that are potentially expensive to compute by a collection of d one-dimensional partition functions. In the three particular examples above, the diagonal elements of Θ are zero. Further, the ancillary statistic h is a product of the form $h(x_1, \ldots, x_d) = \prod_{i=1}^d \underline{h}(x_i)$, with $\underline{h} = 1$ for the Bernoulli and exponential case and $\underline{h}(y) = \frac{1}{y!}$ for the Poisson case. Consequently, we obtain the following expression:

$$\begin{aligned}\prod_{i=1}^d f(x_i|x_{\setminus i}, z; \alpha, \Theta, B) &= \exp\left\{(\alpha + Bz)'x - x'\Theta x - \sum_{i=1}^d \rho\left(\alpha_i + (Bz)_i - \Theta_{i,\setminus i}x_{\setminus i}\right)\right\} h(x)\\ \rho(u) &\triangleq \int_{\mathcal{X}} \exp\{uy\}\, \underline{h}(y) d\underline{\nu}(y).\end{aligned} \tag{2.10}$$

Here $\underline{\nu}$ represents either the one-dimensional counting measure (Bernoulli, Poisson) or the Lebesgue measure on $\mathbb{R}$ (exponential). Thus, each of the d terms $\rho(\alpha_i + (Bz)_i - \Theta_{i,\setminus i}x_{\setminus i})$ corresponding to the normalization for each $f(x_i|x_{\setminus i})$ entails a one-dimensional integral/sum, and in each of the three examples above, the function ρ is expressible in closed form. With this approximation, we obtain the following regularized conditional pseudo-likelihood optimization problem given data $\{x^{(k)}\}_{k=1}^n \subset \mathcal{X}^d$ and user-specified regularization parameters $\lambda, \gamma \geq 0$:

$$
\begin{aligned}
(\widehat{\alpha}, \widehat{\Theta}, \widehat{L}) = \underset{\substack{\alpha \in \mathbb{R}^d, \Theta \in \mathbb{S}^d \\ L \in \mathbb{R}^{d \times n}}}{\arg\min} \; & \frac{1}{n} \sum_{k=1}^{n} \left[\left(\sum_{i=1}^{d} \rho\left(\alpha_i + L_i^{(k)} - \Theta_{i,\backslash i} x_{\backslash i}^{(k)} \right) \right) - (\alpha + L^{(k)})' x^{(k)} + x^{(k)\prime} \Theta x^{(k)} \right] \\
& + \lambda \|\Theta\|_{\ell_1} + \gamma \|L\|_\star \\
& \text{s.t. } (\alpha + L^{(k)}, \Theta) \in \mathcal{F}, \quad k = 1, \ldots, n.
\end{aligned}
\tag{2.11}
$$

We can specialize this convex relaxation to each of the three examples described above with the corresponding choice of valid parameters in the constraint and the following one-dimensional log-partition functions in the objective:

$$
\rho_{\text{ising}}(u) = \log \cosh(u) \qquad \rho_{\text{poisson}}(u) = \exp(u) \qquad \rho_{\text{exponential}}(u) = -\log(-u). \tag{2.12}
$$

If we do not account for the confounding effects of latent variables and set $L = 0$, then we recover a "coupled" analog of the neighborhood selection approaches of [16, 19, 24], which identify the neighborhood of each node in the graph one at a time by solving d uncoupled ℓ_1-regularized regression problems. One (relatively minor) issue with solving d uncoupled problems is that one subsequently needs to reconcile the solutions to obtain a coherent global model over all the variables, although there are several ways to accomplish this [12, 16, 19, 24]. A more significant issue with solving d uncoupled neighborhood selection problems is that it is not clear how to adapt that method to account for the effects of latent variables. In particular, latent variables can simultaneously influence all the observed variables, which necessitates an approach that jointly estimates the (local) neighborhoods of all the nodes at the same time while also teasing apart the (global) effects of the latent variables, as in (2.11). In another direction, if we set $\Theta = 0$ in (2.11) we obtain a convex relaxation for the exponential family PCA problem [9].

2.3 Empirical demonstrations for Gaussian and non-Gaussian models

We evaluate next the empirical performance of the relaxation for the Gaussian case (2.5) and the relaxation for the non-Gaussian case (2.11) for fitting Ising, Poisson, and exponential graphical models when confounded by latent variables. The following are common elements of the setup for each distributional setting: we consider a collection of $d = 60$ observed variables whose distribution conditioned on some latent variables is given by a Gaussian, Ising, Poisson, or exponential graphical model; the distribution of the latent variables is specified later in each case. We generate two types of graphs with a corresponding $\Theta \in \mathbb{S}^{60}$: a cycle graph and a Erdös Réyni graph with edge probabilities 0.02. We vary the number of latent variables $r = \{1, 2, 3\}$ and generate the matrix $B \in \mathbb{R}^{60 \times r}$ so that the coherence[2] of its column-space is approximately $1.2r/d$. The singular values of B vary for each distribution and are described below. Finally, unless otherwise specified, we set $\alpha \in \mathbb{R}^{60}$ to be the identically zero vector, and the nonzero off-diagonal entries of Θ to be 0.4.

Gaussian setup: The distribution of the observed variables conditioned on zero-mean Ising latent variables is a Gaussian graphical model. The diagonal entries of Θ are set to 1. The singular values of B are chosen to be $\{0.72\}$,$\{0.7, 0.7\}$ and $\{0.68, 0.68, 0.68\}$ for $r = 1, 2, 3$ latent variables, respectively.

Ising setup: The distribution of the observed variables conditioned on independent normally distributed hidden variables is an Ising graphical model. The singular values of B are $\{0.72\}$,$\{0.7, 0.7\}$ and $\{0.68, 0.68, 0.68\}$ for $r = 1, 2, 3$ latent variables, respectively.

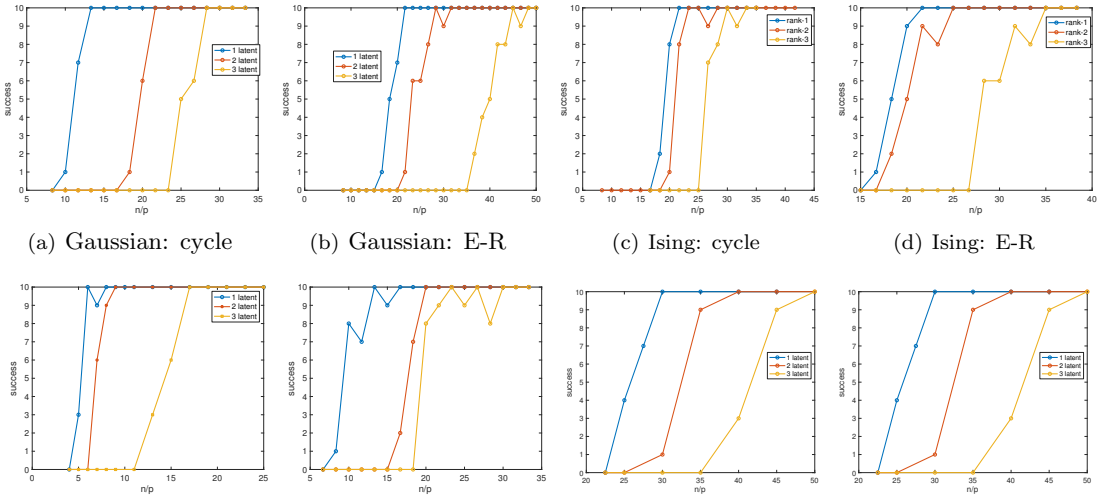
Poisson setup: The distribution of observed conditioned on independent and identically distributed zero-mean Ising hidden variables is a Poisson graphical model. The singular values of B are chosen to be $\{2\}$, $\{1.95, 1.95\}$, $\{1.9, 1.9, 1.9\}$ for $r = 1, 2, 3$ latent variables, respectively.

Exponential setup: The distribution of the observed variables conditioned on independent and identically distributed mean-1 exponential hidden variables is an exponential graphical model. Due to the parameter restriction $\mathcal{F}_{\text{exponential}}$ in (2.8), the entries of Θ must be non-negative, entries of B must be non-positive, and α must consist of negative entries. We set the edge weights (i.e., non-zero entries of Θ) to be 1 and the singular values of B

[2]The coherence of a subspace $\mathcal{S} \subset \mathbb{R}^d$ measures how well $\mathcal{S}$ is aligned with the standard basis vectors in $\mathbb{R}^d$; it is equal to $\max_{i=1,\ldots,d} \|\mathcal{P}_{\mathcal{S}}(e^{(i)})\|_{\ell_2}$, where $e^{(i)}$ is the i'th standard basis vector. The coherence of $\mathcal{S}$ lies in the range $[\sqrt{\dim(\mathcal{S})/d}, 1]$, and this parameter commonly arises in the characterization of statistical identifiability as well as in the analysis of convex relaxations in sparse/low-rank recovery problems.

are chosen to be $\{2\}, \{1.95, 1.95\}, \{1.9, 1.9, 1.9\}$ for $r = 1, 2, 3$ latent variables, respectively. Finally, we set all the entries of α to be equal to -1.

For each problem setting, we generate observations via Gibbs sampling to obtain training data $\{x^{(i)}\}_{i=1}^n \subseteq \mathbb{R}^d$. We supply the data to the estimator (2.5) for the Gaussian model and to the estimator (2.11) for non-Gaussian models (with ρ selected suitably). The regularization parameters λ, γ are chosen with the scaling $\lambda = c_1\sqrt{\frac{d}{n}}$ and $\gamma = c_2\frac{\sqrt{d}}{n}$, for constants c_1, c_2. We evaluate the probability (computed over ten independent trials) that the estimated model correctly identifies the graphical structure as well as the number of latent variables. Figure 2 displays the empirical consistency results for all problem settings. We observe that given sufficient sample size, the estimators (2.5) and (2.11) are successful at correctly identifying the model structure.



(a) Gaussian: cycle (b) Gaussian: E-R (c) Ising: cycle (d) Ising: E-R

(e) Poisson: cycle (f) Poisson: E-R (g) Exponential: cycle (h) Exponential: E-R

Figure 2: Probability of correcting identifying the population graphical model structure and the number of latent variables (computed empirically across 10 trials) with the estimators (2.4) and (2.11) for cycle and Erdös-Rényi graph and $r = 1, 2, 3$ latent variables.

3 Model Selection

The selection of the regularization parameters λ, γ in (1.4) (as well as its specializations and approximations (2.2) and (2.11)) is an important consideration in obtaining a useful model. Standard approaches such as cross-validation tend to yield overly complex models that can overfit to the data [16] (in our context, such models correspond to those in which the graph structure is dense and the number of latent variables is large). To address this issue, several methods have been proposed in the literature based on a notion of *stability* [14, 17], in which a model selection procedure is applied to subsamples of a dataset and the variability of the resulting solutions over the subsamples governs the identification of a suitable regularization parameter or model structure. By combining the ideas in [14, 17], we present a model selection technique that is suited to the context of the present paper in Section 3.1. We demonstrate the utility of these techniques via numerical experiments in Section 3.2.

3.1 Model Selection via Stability

To assess the variability of the structure of the selected models over subsamples, we need an appropriate method to aggregate the models selected over subsamples. Concretely, let $\mathcal{D} = \{x^{(k)}\}_{k=1}^n$ be the given dataset and consider B subsamples $\mathcal{D}^{(l)} \subset \mathcal{D}$, $l = 1, \ldots, B$ with $|\mathcal{D}^{(l)}| = |\mathcal{D}|/2$. Fix regularization parameters λ, γ, and let $(\widehat{\Theta}^{(l)}_{\lambda,\gamma}, \widehat{L}^{(l)}_{\lambda,\gamma})$, $l = 1, \ldots, B$ represent the optimal solutions obtained from (2.2) or (2.11) of the B subsamples (for the purposes of model structure aggregation, α plays no role). To represent the variability in the graphical model structure across

these optimal solutions, form a diagonal matrix $\mathcal{P}^{\text{graph}}_{\lambda,\gamma} \in \mathbb{S}^{\binom{d}{2}}$ as follows:

$$\left(\mathcal{P}^{\text{graph}}_{\lambda,\gamma}\right)_{e,e} = \frac{1}{B}\sum_{l=1}^{B} \mathbb{I}\left(e \in \text{support}(\widehat{\Theta}^{(l)}_{\lambda,\gamma})\right).$$

Here e ranges over the $\binom{d}{2}$ edges and $\mathbb{I}$ denotes the indicator function that equal 1 if its argument is true and 0 otherwise. In words, the diagonal entries of $\mathcal{P}^{\text{graph}}_{\lambda,\gamma}$ lie in $[0,1]$ and they encode the frequencies of the edges appearing in the selected models aggregated over the subsamples. The methods presented in [14, 17] may be described in terms of this matrix, and they were applicable to discrete model selection problems such as graph estimation and variable selection. These ideas were extended recently to low-rank estimation problems based on a geometric reformulation of model selection [21], with a key ingredient being a suitable generalization of the aggregate matrix $\mathcal{P}^{\text{graph}}_{\lambda,\gamma}$. Specifically, for each $\widehat{L}^{(l)}_{\lambda,\gamma}$ let $\mathcal{P}^{(l)}_{\lambda,\gamma} \subset \mathbb{S}^d$ denote the projection operator onto the column-space of $\widehat{L}^{(l)}_{\lambda,\gamma}$. With this notation, the variability in the structure underlying the low-rank estimates across subsamples is specified by the following average projection map:

$$\mathcal{P}^{\text{latent}}_{\lambda,\gamma} = \frac{1}{B}\sum_{l=1}^{B} \mathcal{P}^{(l)}_{\lambda,\gamma}$$

The eigenvalues of $\mathcal{P}^{\text{latent}}_{\lambda,\gamma}$ lie in the range $[0,1]$. We describe next the three main steps of our model selection approach.

Stage 1: Identifying Regularization Parameters The first step is to select appropriate values of λ, γ. Building on the insights of [14], let $\pi^{\text{graph}}_{\lambda,\gamma} = \frac{1}{\binom{d}{2}}[\text{trace}(\mathcal{P}^{\text{graph}}_{\lambda,\gamma}) - \text{trace}(\mathcal{P}^{\text{graph}}_{\lambda,\gamma})^2]$ and $\pi^{\text{latent}}_{\lambda,\gamma} = \frac{1}{d}[\text{trace}(\mathcal{P}^{\text{latent}}_{\lambda,\gamma}) - \text{trace}(\mathcal{P}^{\text{latent}}_{\lambda,\gamma})^2]$ denote the total variabilities in the graphical and latent components, respectively. These parameters lie in the range $[0,1]$, and they are small when the graph structure and the latent subspace are stable across subsamples. For sufficiently large values of λ, γ, the graph structure is completely disconnected and the latent subspace is zero-dimensional; correspondingly, $\pi^{\text{graph}}_{\lambda,\gamma}$ and $\pi^{\text{latent}}_{\lambda,\gamma}$ are both zero. As λ, γ are gradually decreased, more edges and higher-dimensional subspaces are progressively included in the recovered graph structure and latent components, and the values of $\pi^{\text{graph}}_{\lambda,\gamma}$ and $\pi^{\text{latent}}_{\lambda,\gamma}$ begin to increase. When these values reach a desired user-specified threshold, the corresponding λ, γ are set as the regularization parameters. (Typical threshold values for $\pi^{\text{graph}}_{\lambda,\gamma}$ and for $\pi^{\text{latent}}_{\lambda,\gamma}$ are 0.025, thus yielding a total variability of 0.05 as recommended in [14]). This approach provides regularization parameters for which the associated graphical model and latent subspace are sparse/low-dimensional, while exhibiting little overall variability across subsamples.

Stage 2: Identifying Model Structure Solving (2.2) or (2.11) with the regularization parameters obtained from the preceding step tends to lead to models that have small type-II error (formally [14] shows that type-II error in graph structure estimation is small under minimal assumptions). However, to also reduce type-I error it is useful to further restrict the models selected based on a more refined form of stability, as described in [17, 21]. Specifically, while the approach of [14] considers aggregate variability, the methods in [17, 21] suggest selecting a graphical model structure and a latent subspace that are common to a large proportion of the subsamples. Concretely, let $\mathcal{P}^{\text{graph}}$ and $\mathcal{P}^{\text{latent}}$ represent the average projections over subsamples for the values of the regularization parameters chosen from the previous step (we have suppressed the dependence on λ, γ for notational clarity). For the graphical model component, we select those edges corresponding to all those elements on the diagonal of $\mathcal{P}^{\text{graph}}$ that are above a user-specified threshold $\delta^{\text{graph}} \in [0,1]$; for large values of δ^{graph}, these correspond to edges that are chosen in a large proportion of the subsamples. For the latent subspace component, we select the largest-dimensional subspace $\mathcal{C} \in \mathbb{R}^d$ such that $\sigma_{\min}(\mathcal{P}_{\mathcal{C}}\mathcal{P}^{\text{latent}}\mathcal{P}_{\mathcal{C}}) \geq \delta^{\text{latent}}$. Here $\delta^{\text{latent}} \in [0,1]$ is again user-specified, $\mathcal{P}_{\mathcal{C}} \in \mathbb{S}^d$ denotes the projection onto $\mathcal{C}$, and $\sigma_{\min}(\mathcal{P}_{\mathcal{C}}\mathcal{P}^{\text{latent}}\mathcal{P}_{\mathcal{C}})$ is the smallest singular value of the operator $\mathcal{P}_{\mathcal{C}}\mathcal{P}^{\text{latent}}\mathcal{P}_{\mathcal{C}}$ viewed as a self-adjoint map on $\mathcal{C}$. Selecting such a subspace may be accomplished by a singular value decomposition of $\mathcal{P}^{\text{latent}}$, and for large values of δ^{latent}, the selected subspace is one that well-aligned with the subspaces that are chosen in a large proportion of the subsamples. (A typical recommended value for both $\delta^{\text{graph}}, \delta^{\text{latent}}$ is 0.7, as suggested in [17, 21]). As shown in [17] for sparse models and in [21] for low-rank models, such stability-based approaches yield models with small type-I error.

Stage 3: Identifying Model Parameters The output of the preceding step is a stable subset of edges $\mathcal{E}$ for the graphical model and a stable column-space $\mathcal{C}$ for the latent component. With these in hand, we solve either (2.2) or (2.11) with two modifications. First, we add the constraint that Θ must lie in the subspace of matrices in which the entries indexed by $\mathcal{E}^c$ equal zero and the constraint that L must lie in the subspace of matrices in which each column lies in $\mathcal{C}$. Second, we set the regularization parameters $\lambda = \gamma = 0$ as these are no longer required to obtain low-complexity models. Even with these modifications (2.2) and (2.11) continue to be tractable convex optimization problems.

3.2 Experimental Demonstration

We provide empirical demonstration of the utility of the model selection method presented above in terms of the false discovery rate (FDR) and true positive rate (PWR) of the estimated graph structure; the FDR is the expected ratio of the number of estimated edges that are not in the true underlying graph over the total number of estimated edges and the PWR is expected ratio of the number of estimated edges that are in the true underlying graph over the total number of estimated edges.

We consider the setting where the conditional graphical model of 50 observed variables conditioned on two independent normally distributed latent variables is an Ising model, with the population graphical structure being an Erdös-Rényi graph with edge selection probability 0.02 and edge weights 0.4. The coefficient matrix $B \in \mathbb{R}^{50\times 2}$ is a random partial orthogonal matrix sampled uniformly from the Haar measure. We obtain observations (via Gibbs sampling) and compute the FDR and PWR over 10 trials based on the above problem setup using the estimator (2.11) with $\rho = \rho_{\text{ising}}$. Figure 3 demonstrates the graph recovery performance after employing the first stage of our model selection approach as well as combining both the first and second stages. Notice that for moderate n, the first stage of the algorithm yields a graphical structure with PWR ≈ 1 but also high FDR (i.e. many false positives). After the second stage, we substantially reduce FDR without much loss in power. These results provide empirical support for the utility of our two-stage model selection method, and in particular the fact that combing both stages yields graphical models that have small Type-I error as well as small Type-II error.

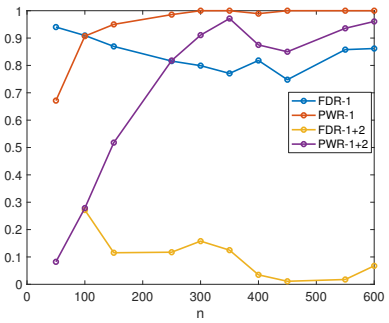


Figure 3: False discovery rate (FDR) and true positive rate (PWR) as a function of n after the first stage of the proposed model selection procedure, denoted by 'FDR-1' and 'PWR-1', and after the second stage of the model selection procedure, denoted by 'FDR-1+2' and 'PWR-1+2'.

4 Experiments with Real Data

In this section, we demonstrate the utility of our latent-variable graphical modeling framework on US Senate voting records data, mi-RNA sequence data, and S&P-500 stock data. We provide comparisons between the graphical structure obtained via our approach and graphical models that do not incorporate latent variables. In addition to examining the difference between the graphical structure of the approach with latent variables to the one without, we also provide quantitative comparison of their prediction performance. Specifically, let $\mathcal{D}_{\text{train}} = \{x^{(k)}\}_{k=1}^{n_{\text{train}}}$ and $\mathcal{D}_{\text{test}} = \{x^{(i)}\}_{k=1}^{n_{\text{train}}}$ denote training and test datasets, respectively. Using (2.2) and (2.11), let $(\widehat{\alpha}_{\text{latent}}, \widehat{\Theta}_{\text{latent}}, \widehat{L}_{\text{latent}})$ denote the estimated parameters based on our approach that incorporates latent variables. Next, let $(\widehat{\alpha}_{\text{no-latent}}, \widehat{\Theta}_{\text{no-latent}})$ denote the estimated parameters based on fixing $L = 0$ so that latent variables are not incorporated. In obtaining these models, we employ the model selection technique described in Section 3 with the various thresholds chosen as stated in the corresponding stages; the one distinction is that when

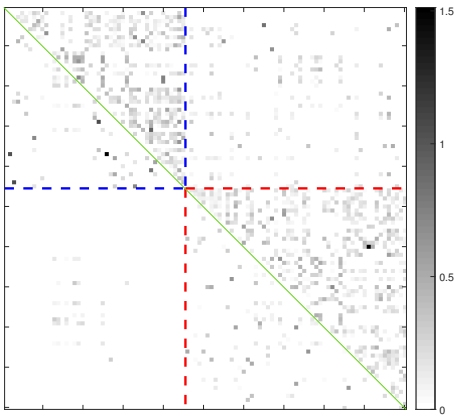


Figure 4: Edges between senator pairs in the graphical model with 8 latent variables (bottom triangle) compared with those of the no latent variable estimate (tom triangle); the two parties zones are separated by red and blue lines: top left correspond to interactions among Democrats, and bottom right among Republicans; white indicates no edge.

we fix $L = 0$ to obtain a graphical model that does not incorporate latent variables, we fix the variability threshold for $\pi_\lambda^{\text{graph}}$ to be 0.05. We evaluate the prediction performance of the two models by comparing negative log (pseudo-)likelihood values on the test data $\mathcal{D}_{\text{test}}$. These values are obtained by solving unregularized (i.e., $\lambda = \gamma = 0$) and suitably constrained versions of (2.2) and (2.11). In particular, for the setting with latent variables, we consider the optimal values of these problems with the additional constraints $\alpha = \widehat{\alpha}_{\text{latent}}, \Theta = \widehat{\Theta}_{\text{latent}}, \text{col-space}(L) \subseteq \text{col-space}(\widehat{L}_{\text{latent}})$, and for the setting with no latent variables we consider the optimal values with the constraints $\alpha = \widehat{\alpha}_{\text{no-latent}}, \Theta = \widehat{\Theta}_{\text{no-latent}}, L = 0$.

4.1 Senate voting records data

We apply our latent-variable modeling framework to the 109th Senate voting record dataset. The dataset was obtained from the website of the US Congress (`http://www.senate.gov`). It contains the voting records of the 100 senators – 55 Republicans, 44 Democrats, and one Independent – of the 109th congress (January 3, 2005 January 3, 2007) on 645 bills on which the Senate voted. The votes are recorded as $+1$ for "yes and -1 for "no. The data contains missing votes as some senators abstained on a small number of bills. The missing values (missed votes) for each senator were imputed with the majority vote of that senators party on that particular bill and the missing votes of the Independent Senator Jeffords were imputed with the Democratic majority vote (because he caucused with the Democrats). Finally, we exclude bills where the "yes/no" proportion fell outside the interval $[0.3, 0.7]$. This results in $n = 479$ votes across $d = 100$ senators to yield a dataset of $\mathcal{D} = \{x^{(k)}\}_{k=1}^{479} \subset \{-1, +1\}^{100}$. We take the first 383 samples as training set $\mathcal{D}_{\text{train}}$ and the remaining 96 samples as test data $\mathcal{D}_{\text{test}}$.

We fit Ising models with and without latent variables to this dataset. We obtain a latent-variable graphical model with $r = 8$ latent variables and a conditional graphical model with the number of edges equal to 6% of the total number of pairs of variables. In contrast, the model without latent variables is given by a graphical model with edge density $\approx 24\%$. The edge weights of the graph structure in the latent-variable graphical model are shown in the bottom half of Figure 5 and the edge weights of the graphical model without latent variables are shown in the top half of Figure 5. The majority of the interactions in the estimated edges in the graphical models occur between individuals in the same party. The incorporation of latent variables substantially reduces the number of edges as many confounding dependencies are removed. Examining the positive interactions in the model with latent variables, the strongest edge among the Democrats is between senators Pryor-Lautenburg and among the Republicans is between senators Robert-Inhofe. We observe that conditioning on the latent variables induces some negative dependencies between senators in the same party, notably Pryor-Baucus & Reed-Levin among the Democrats and Enzi-Coburn & Sessions-Cornyn among the Republicans. Finally, the negative log (pseudo-)likelihoods evaluated on the test data (in the manner described above) yield values of 38.3 without latent variables and 33.8 with latent variables, suggesting that our approach which incorporates latent variables more accurately models Senate voting records.

4.2 mi-RNA sequence data

Next, we demonstrate the utility of our approach in estimating an miRNA inhibitory network for Level III breast cancer miRNA expressions (downloaded from `http://tcga-data.nci.nih.gov/tcga/`). The data consist of 262 miRNAs and 544 subjects. Of the 262 miRNAs, we extract 27 that were considered by [24] (after a hierarchical clustering step) as their interactions are well modeled by negative dependencies (since (2.7) only allows for negative dependencies or equivalently Θ non-negative). The data consisting of the selected 27 miRNAs were adjusted for possible over-dispersion using a power transform [1]. After performing these pre-processing steps, we obtain training data $\mathcal{D}_{\text{train}} = \{x^{(k)}\}_{k=1}^{544} \subseteq \mathbb{R}^{27}$ that is well-modeled by a Poisson distribution.

We fit Poisson graphical models with and without latent variables. We obtain a latent-variable graphical model consisting of $r = 2$ latent variables and a conditional graphical model in which the number of edges is 3% of the total number of pairs of variables. The graphical model that does not incorporate latent variables has an edge density of $\approx 18\%$. The corresponding graphs are displayed in the bottom triangle and the top triangle of Figure 5, respectively. We observe that incorporating latent variables in the graphical model removes dependencies between pairs of miRNAs that have similar primary function. Specifically, the strongest edges in the graphical model without latent variables that are not part of the graphical model that incorporates latent variables are among the miRNAs '632' (promotes cell proliferation in carcinoma cancer) and '215' (early indicator of carcinoma cancer); '186' and '132' (both are colorectal cancer suppressants); and '374' and '9-1' (both are prostate cancer suppressants). Further, the majority of the edges in the graphical model with latent variables are among miRNAs that have different functionalities. Specifically, the five strongest edges in this graph are between the pairs: '449b' (breast cancer suppressant) and '577' (lung cancer suppressant); '192' (oncogene for prostate cancer) and '518c' (inhibits gastric cell growth); '449b' and '518c' (both are breast cancer suppressants), '449b' (breast cancer suppresant) and '143' (down-regulated in lung cancer); and '518c' (inhibits gastric cell growth) and '141' (biomarker in prostate cancer). Of these five strongest edges, only the one linking '449b' and '518c' is between similar functioning miRNAs, and this edge is also present in the graphical model without latent variables. In summary, these observations suggest that in the latent variable graphical model, the latent variables may correspond to commonalities in the biological functions of various miRNAs, and the associated confounding edges are not present in the conditional graphical model. We observe a similar feature in the experimental results in the next subsection with stock return data.

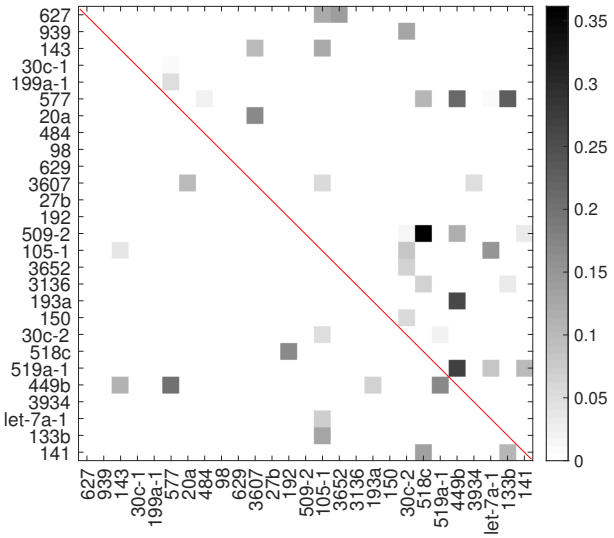


Figure 5: Edges between miRNAs in the graphical model conditioned on two latent variables (bottom triangle) and in the graphical model without latent variables (top triangle).

4.3 Stock data

We analyze monthly stock returns of $d = 40$ companies from the Standard and Poor index over the period March 1982 to March 2016, which leads to a total of $n = 408$ observations. We set aside $n_{\texttt{test}} = 58$ observations as the test set, and the remaining $n_{\texttt{train}} = 350$ observations as the training set. In this experiment, we apply the convex relaxation (2.2) for fitting Gaussian graphical models conditioned on latent variables.

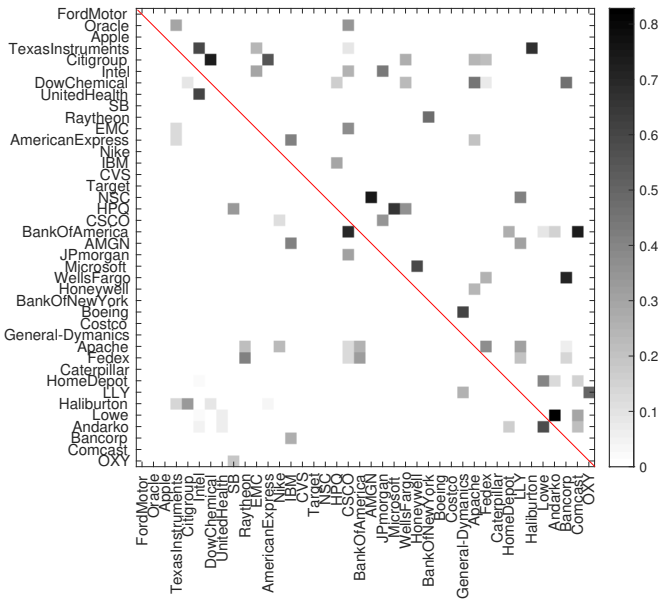


Figure 6: Bottom triangle: edges among pairs of companies obtained with the conditional likelihood procedure; top triangle: graphical model without incorporating latent variables; white indicates no edge.

We obtain a latent-variable graphical model with $r = 3$ latent variables and a conditional graphical model with edge density $\approx 3.8\%$. The magnitudes of the partial correlations corresponding to this conditional graphical model are displayed in the bottom triangle of Figure 6. The strongest five edges in this graph are between companies Andarko - Lowe, United Health - Intel, Bank of America - Cisco, IBM - Amgen, Fedex - Raytheon. Note that in the Standard Industrial Classification system[3] for grouping these companies, all of these pairs are in different classes. We also obtain a graphical model that does not incorporate latent variables, and the graph structure of this model has edge density $\approx 6.5\%$; the corresponding magnitudes of the partial correlations are shown in the top triangle of Figure 6. In contrast to the previous model that incorporated latent variables, four of the five strongest edges in this graphical model without latent variables graph are between companies in the same category: Texas Instruments - Intel, HPQ - Microsoft, Wellsfargo - Bancorp, and Boeing - General Dynamics. The negative log likelihoods (based on the procedure described previously in this section) evaluated on the test data yield values of 18.1 for the model that incorporates latent variables as compared to 22.2 for the model without latent variables, which suggests that accounting for the confounding effects of latent variables yields a better fit to stock data.

5 Discussion

In this paper, we describe a new convex relaxation framework for learning a latent variable graphical model given sample observations of a collection of variables. Specifically, we fit the observations to a model in which the conditional distribution of the observed variables conditioned on latent variables is given by an exponential family graphical model (1.1). Our approach is based on regularized conditional likelihood and we demonstrate the utility and flexibility of our method with both synthetic data as well as real data from a variety of problem domains.

There are several interesting directions for further investigation arising from our work that we outline below:

Consistency of the estimators (2.2) **and** (2.11): In Section 2.3, we presented empirical evidence that the estimators (2.2) and (2.11) consistently identify the structure of a latent variable graphical model in various settings and with several types of distributions. Given this empirical demonstration, it would be of interest to provide theoretical support for the consistency of our method by leveraging prior work such as [7].

Comparison of [7] and estimator (2.2) **for Gaussian graphical modeling:** The authors in [7] develop a convex relaxation for the problem of latent variable graphical modeling in settings with jointly Gaussian observed and latent variables. Their approach proceeds by considering the marginal distribution of the observed variables and

[3]See the U.S. SEC website at `http://www.sec.gov/info/edgar/siccodes.html`.

explicitly characterizing the influence of the latent variables on the observed variables upon marginalization. This characterization and the underlying assumption of joint Gaussianity are central to the derivation of the relaxation in [7]. In contrast, the derivation of our estimator (2.2) requires no knowledge of the distribution of the latent variables, and only assumes that the conditional distribution of the observed variables conditioned on the latent variables is given by a Gaussian graphical model. This distinction suggests that the framework in this paper is more flexible and may be more robust to different distributions for the latent variables. A natural question is to develop further theoretical and empirical understanding of comparative advantages of each approach for latent variable Gaussian graphical modeling.

Theoretical support for the proposed model selection procedure: In Section 3, we described a model selection procedure – based on a notion of *stability* – for selecting regularization parameters λ, γ in 2.2 and in 2.11 as well as for identifying suitable model structures. Our approach is based on combining the ideas of [14, 17] for the graphical model component and [21] for the latent subspace. The empirical demonstrations in Section 3.2 suggest that our procedure provides good control over both Type-I and Type-II errors in estimating a population graphical model. These are perhaps to be expected based on the theoretical analyses in [14, 17, 21], and it would be useful to combine and formalize these results in our context by showing that the Type-I and Type-II errors can be provably controlled under appropriate assumptions.

Better regularizers: The nuclear norm regularizer in (1.4) (and its specializations) is agnostic to the type or form of the latent variables and only encourages low-rankness (i.e., few latent variables). If a data analyst has access to additional information about potential latent variables (e.g., the latent variables take on non-negative or categorical values) or wishes to fit to models in which the latent variables have additional structure, one can design tighter convex regularizers than the nuclear norm [8]. As an example, if the latent variables take on binary values, a tighter regularizer than the nuclear norm is the max-2 norm. Thus, an exciting direction is to investigate the computational and statistical tradeoffs underlying these tighter regularizers for latent variable graphical modeling.

Tailored computational methods: We solve the convex program (2.11) via an ADMM procedure that we implemented ourselves [4]. The most costly component of this algorithm is computing a singular-value decomposition of $d \times n$ matrices, which can be prohibitive when the sample size or the number of variables are large. For the convex program (2.4), we use the off-the-shelf logDetPPA solver [22] and it tends to be prohibitively expensive beyond $d \approx 500$ on standard contemporary workstations. Fast solvers for the graphical Lasso (and its variants) that can handle up to tens of thousands and sometimes millions of variables by exploiting problem-specific structure have been proposed previously [13, 15]. Designing similar custom solvers for the relaxations (2.11) and (2.4) proposed in this paper would enable a broader application of our methods in problems with a large number of variables.

References

[1] G. Allen and Z. Liu, *A local poisson graphical model for inferring networks from sequencing data*, IEEE Transactions on NanoBioscience, 12 (2013), pp. 1–10.

[2] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, *Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data*, Journal of Machine Learning Research, 9 (2008), pp. 485–516.

[3] J. Besag, *Spatial interaction and spatial analysis of lattice systems*, Journal of Royal Statistical Society (Series B), 36 (1974), pp. 192–236.

[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 (2010), pp. 1–122.

[5] G. Bresler and R. Buhai, *learning restricted boltzmann machines with few latent variables*, arXiv:2006.04166, (2020).

[6] G. Bresler, F. Koehler, A. Moitra, and E. Mossel, *Learning restricted boltzmann machines via influence maximization*, arXiv:1805.10262, (2018).

[7] V. Chandrasekaran, P. A. Parillo, and A. S. Willsky, *Latent variable graphical model selection via convex optimization*, Annals of Statistics, 40 (2012), pp. 1935–1967.

[8] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky, *Convex geometry of linear inverse problems*, Foundations of Computational Mathematics, 12 (2012), pp. 805–849.

[9] M. Collins, S. Dasgupta, and R. Schapire, *A generalization of principal component analysis to the exponential family*, Neural Information Processing Systems, (2002).

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

[11] S. Goel, *Learning ising and potts models with latent variables*, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, 18 (2020), pp. 3557–3566.

[12] H. Höfling and R. Tibshirani, *Estimation of sparse binary pairwise markov networks using pseudo-likelihoods*, Journal of machine learning research, 10 (2009), pp. 883–906.

[13] C. Hsieh, M. Sustik, I. Dhillon, P. Ravikumar, and R. Poldrack, *Big & quic: sparse inverse covariance estimation for a million variables*, Advanced Neural Information Processing System, (2013).

[14] H. Liu, K. Roeder, and L. Wasserman, *Stability approach to regularization selection (StARS) for high dimensional graphical models*, International Conference on Neural Information Processing Systems, 2 (2010), pp. 1432–1440.

[15] s. Ma, L. Xue, and H. Zou, *Alternating direction methods for latent variable graphical model selection*, Neural Computations, 25 (2013), pp. 2172–2198.

[16] N. Meinshausen and P. Bühlmann, *High dimensional graphs and variable selection with the lasso*, Annals of Statistics, 34 (2006), pp. 1436–1462.

[17] N. Meinshausen and P. Bühlmann, *Stability selection*, Journal of Royal Statistical Methodology (Series B), 72 (2010), pp. 417–473.

[18] F. Nussbaum and J. Giesen, *Ising models with latent conditional gaussian variables*, Proceedings of Machine Learning Research, 13 (2019), pp. 1–9.

[19] P. Ravikumar, M. Wainwright, and J. Lafferty, *High-dimensional ising model selection using ℓ_1-regularized logistic regression*, Annals of Statistics, 3 (2010), pp. 1287–1319.

[20] A. Shapiro, *Weighted minimum trace factor analysis*, Psychometrika, 77 (1982), pp. 243–264.

[21] A. Taeb, P. Shah, and V. Chandrasekaran, *False discovery and its control in low-rank estimation*, Journal of Royal Statistical Society (Series B), 82 (2020), pp. 997–1027.

[22] K. C. Toh, M. J. Todd, and R. H. Tutuncu, *SDPT3 - a matlab software package for semidefinite-quadratic-linear programming.* 2016.

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

[24] E. Yang, P. Ravikumar, G. Allen, and Z. Liu, *On graphical models via univariate exponential family distributions*, Journal of Machine Learning Research, 16 (2015), pp. 3813–3847.

[25] M. Yuan and Y. Lin, *Model selection and estimation in the gaussian graphical model*, Biometrika, 94 (2007), pp. 19–35.