Structure Learning of Mixed Graphical Models

Jason D. Lee
Institute of Computational and Mathematical Engineering
Stanford University

Trevor J. Hastie
Department of Statistics
Stanford University

Abstract

We consider the problem of learning the structure of a pairwise graphical model over continuous and discrete variables. We present a new pairwise model for graphical models with both continuous and discrete variables that is amenable to structure learning. In previous work, authors have considered structure learning of Gaussian graphical models and structure learning of discrete models. Our approach is a natural generalization of these two lines of work to the mixed case. The penalization scheme is new and follows naturally from a particular parametrization of the model.

1 Introduction

Many authors have considered the problem of learning the edge structure and parameters of sparse undirected graphical models. We will focus on using the l_1 regularizer to promote sparsity. This line of work has taken two separate paths: one for learning continuous valued data and one for learning discrete valued data. However, typical data sources contain both continuous and discrete variables: population survey data, genomics data, url-click pairs etc. In this work, we consider learning mixed models with both continuous variables and discrete variables.

For only continuous variables, previous work assumes a multivariate Gaussian (Gaussian graphical) model with mean 0 and inverse covariance Θ. Θ is then estimated via the graphical lasso by minimizing the regularized negative log-likelihood $\ell(\Theta) + \lambda \|\Theta\|_1$. Several efficient methods for solving this can be found in [9, 2].

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

Because the graphical lasso problem is computationally challenging, several authors considered methods related to the pseudolikelihood (PL) and node-wise l_1 least squares [20, 10, 21]. For discrete models, previous work focuses on estimating a pairwise Markov random field of the form $p(y) \propto \exp \sum_{r \leq j} \phi_{rj}(y_r, y_j)$. The maximum likelihood problem is intractable for models with a moderate to large number of variables (high-dimensional) because it requires evaluating the partition function and its derivatives. Again previous work has focused on the pseudolikelihood approach [11, 23, 25, 12, 13, 17, 22].

Our main contribution here is to propose a model that connects the discrete and continuous models. The conditional distributions of this model are two widely adopted and well understood models: multiclass logistic regression and Gaussian linear regression. In addition, in the case of only discrete variables, our model is a pairwise Markov random field; in the case of only continuous variables, it is a Gaussian graphical model. Our proposed model leads to a natural scheme for structure learning that generalizes the graphical Lasso. Since each parameter block is of different size, we also derive a calibrated weighting scheme to penalize each edge fairly.

In Section 2, we introduce our new mixed graphical model and discuss previous approaches to modeling mixed data. Section 3 discusses the pseudolikelihood approach to parameter estimation and connections to generalized linear models. Section 4 discusses a natural method to perform structure learning in the mixed model. Section 5 presents the calibrated regularization scheme and Section 6 discusses two methods for solving the optimization problem. Finally, Section 7 discusses a conditional random field extension and Section 8 presents empirical results on a census population survey dataset and synthetic experiments.

2 Mixed Graphical Model

We propose a pairwise graphical model on continuous and discrete variables. The model is a pairwise Markov

random field with density

$$p(x,y;\Theta) \propto \exp\left(\sum_{s=1}^{p}\sum_{t=1}^{p} -\frac{1}{2}\beta_{st}x_s x_t + \sum_{s=1}^{p}\alpha_s x_s + \sum_{s=1}^{p}\sum_{j=1}^{q}\rho_{sj}(y_j)x_s + \sum_{j=1}^{q}\sum_{r=1}^{q}\phi_{rj}(y_r,y_j)\right). \quad (1)$$

Here x_s denotes the sth of p continuous variables, and y_j the jth of q discrete variables. The joint model is parametrized by $\Theta = [\{\beta_{st}\},\{\alpha_s\},\{\rho_{sj}\},\{\phi_{rj}\}]$[1]. The discrete y_r takes on L_r states. The model parameters are β_{st} continuous-continuous edge potential, α_s continuous node potential, $\rho_{sj}(y_j)$ continuous-discrete edge potential, and $\phi_{rj}(y_r,y_j)$ discrete-discrete edge potential.

The two most important features of this model are:

1. the conditional distributions are given by Gaussian linear regression and multiclass logistic regressions;

2. the model simplifies to a multivariate Gaussian in the case of only continuous variables and simplifies to the usual discrete pairwise Markov random field in the case of only discrete variables.

The conditional distributions of a graphical model are of critical importance. The absence of an edge corresponds to two variables being conditionally independent. The conditional independence can be read off from the conditional distribution of a variable on all others. For example in the multivariate Gaussian model, x_s is conditionally independent of x_t iff the partial correlation coefficient is 0. Our mixed model has the desirable property that the two type of conditional distributions are simple Gaussian linear regressions and multiclass logistic regressions. This follows from the pairwise property in the joint distribution. In more detail:

1. The conditional distribution of y_r given the rest is multinomial, with probabilities defined by a multiclass logistic regression where the covariates are the other variables x_s and $y_{\setminus r}$ (denoted collectively by z in the right-hand side):

$$p(y_r = k|y_{\setminus r},x;\Theta) = \frac{\exp\left(\omega_k^T z\right)}{\sum_{l=1}^{L_r}\exp\left(\omega_l^T z\right)} = \frac{\exp\left(\omega_{0k}+\sum_j \omega_{kj}z_j\right)}{\sum_{l=1}^{L_r}\exp\left(\omega_{0l}+\sum_j \omega_{lj}z_j\right)} \quad (2)$$

   Here we use a simplified notation, which we make explicit in Section 3.1. The discrete variables are represented as dummy variables for each state, e.g. $z_j = 1[y_u = k]$, and for continuous variables $z_s = x_s$.

2. The conditional distribution of x_s given the rest is Gaussian, with a mean function defined by a linear regression with predictors $x_{\setminus s}$ and y_r.

$$E(x_s|x_{\setminus s},y_r;\Theta) = \omega^T z = \omega_0 + \sum_j z_j\omega_j \quad (3)$$

$$p(x_s|x_{\setminus s},y_r;\Theta) = \frac{1}{\sqrt{2\pi}\sigma_s}\exp\left(-\frac{1}{2\sigma_s^2}(x_s-\omega^T z)^2\right).$$

   As before, the discrete variables are represented as dummy variables for each state $z_j = 1[y_u = k]$ and for continuous variables $z_s = x_s$.

The exact form of the conditional distributions (2) and (3) are given in (9) and (8) in Section 3.1, where the regression parameters ω_j are defined in terms of the parameters Θ.

The second important aspect of the mixed model is the two special cases of only continuous and only discrete variables.

1. Continuous variables only. The pairwise mixed model reduces to the familiar multivariate Gaussian parametrized by the symmetric positive-definite inverse covariance matrix $B = \{\beta_{st}\}$ and mean $\mu = B^{-1}\alpha$,

$$p(x) \propto \exp\left(-\frac{1}{2}(x-B^{-1}\alpha)^T B(x-B^{-1}\alpha)\right).$$

2. Discrete variables only. The pairwise mixed model reduces to a pairwise discrete (second-order interaction) Markov random field,

$$p(y) \propto \exp\left(\sum_{j=1}^{q}\sum_{r=1}^{q}\phi_{rj}(y_r,y_j)\right).$$

Although these are the most important aspects, we can characterize the joint distribution further. The conditional distribution of the continuous variables given

[1] $\rho_{sj}(y_j)$ is a function taking L_j values $\rho_{sj}(1),\ldots,\rho_{sj}(L_j)$. Similarly, $\phi_{rj}(y_r,y_j)$ is a bivariate function taking on $L_r \times L_j$ values. Later, we will think of $\rho_{sj}(y_j)$ as a vector of length L_j and $\phi_{rj}(y_r,y_j)$ as a matrix of size $L_r \times L_j$.

the discrete follow a multivariate Gaussian distribution, $p(x|y) = \mathcal{N}(\mu(y), B^{-1})$. Each of these Gaussian distributions share the same inverse covariance matrix B, since all the parameters are pairwise. The mean parameter depends additively on the value of the discrete variables. By standard multivariate Gaussian calculations,

$$p(x|y) = \mathcal{N}(B^{-1}\gamma(y), B^{-1}) \tag{4}$$

$$\{\gamma(y)\}_s = \alpha_s + \sum_j \rho_{sj}(y_j) \tag{5}$$

$$p(y) \propto \exp\left(\sum_{j=1}^{q}\sum_{r=1}^{j}\phi_{rj}(y_r, y_j) + \frac{1}{2}\gamma(y)^T B^{-1}\gamma(y)\right) \tag{6}$$

2.1 Related work on mixed graphical models

Lauritzen [15] proposed a type of mixed graphical model, with the property that conditioned on discrete variables, $p(x|y) = \mathcal{N}(\mu(y), \Sigma(y))$. The homogeneous mixed graphical model enforces common covariance, $\Sigma(y) \equiv \Sigma$. Thus our proposed model is a special case of Lauritzen's mixed model with the following assumptions: common covariance, additive mean assumptions and the marginal $p(y)$ factorizes as a pairwise discrete Markov random field. With these three assumptions, the full model simplifies to the mixed pairwise model presented. Although the full model is more general, the number of parameters scales exponentially with the number of discrete variables, and the conditional distributions are not as convenient. For each state of the discrete variables there is a mean and covariance. Consider an example with q binary variables and p continuous variables; the full model requires estimates of 2^q mean vectors and covariance matrices in p dimensions. Even if the homogeneous constraint is imposed on Lauritzen's model, there are still 2^q mean vectors for the case of binary discrete variables. The full mixed model is very complex and cannot be easily estimated from data without some additional assumptions. In comparison, the mixed pairwise model has number of parameters $O((p+q)^2)$ and allows for a natural regularization scheme which makes it appropriate for high dimensional data.

There is a line of work regarding parameter estimation in undirected mixed models that are decomposable: any path between two discrete variables cannot contain only continuous variables. These models allow for fast exact maximum likelihood estimation through node-wise regressions, but are only applicable when the structure is known and $n > p$ [7]. There is also related work on parameter learning in directed mixed graphical models. Since our primary goal is to learn the graph structure, we forgo exact parameter estimation and use the pseudolikelihood. Similar to the exact maximum likelihood in decomposable models, the pseudolikelihood can be interpreted as node-wise regressions that enforce symmetry.

To our knowledge, this work is the first to consider convex optimization procedures for learning the edge structure in mixed graphical models.

3 Parameter Estimation: Maximum Likelihood and Pseudolikelihood

Given samples $(x_i, y_i)_{i=1}^n$, we want to find the maximum likelihood estimate of Θ. This can be done by minimizing the negative log-likelihood of the samples:

$$\ell(\Theta) = -\sum_{i=1}^{n} \log p(x_i, y_i; \Theta) \text{ where}$$

$$\log p(x, y; \Theta) = \sum_{s=1}^{p}\sum_{t=1}^{p} -\frac{1}{2}\beta_{st}x_s x_t + \sum_{s=1}^{p}\alpha_s x_s$$

$$+\sum_{s=1}^{p}\sum_{j=1}^{q}\rho_{sj}(y_j)x_s + \sum_{j=1}^{q}\sum_{r=1}^{j}\phi_{rj}(y_r, y_j) - \log Z(\Theta)$$

The negative log-likelihood is convex, so standard gradient-descent algorithms can be used for computing the maximum likelihood estimates. The major obstacle here is $Z(\Theta)$, which involves a high-dimensional integral. Since the pairwise mixed model includes both the discrete and continuous models as special cases, maximum likelihood estimation is at least as difficult as the two special cases, the first of which is a well-known computationally intractable problem. We defer the discussion of maximum likelihood estimation to Supplementary Material.

3.1 Pseudolikelihood

The pseudolikelihood method [5] is a computationally efficient and consistent estimator formed by products of all the conditional distributions:

$$\tilde{\ell}(\Theta|x, y) = -\sum_{s=1}^{p}\log p(x_s|x_{\setminus s}, y; \Theta) - \sum_{r=1}^{q}\log p(y_r|x, y_{\setminus r}; \Theta) \tag{7}$$

The conditional distributions $p(x_s|x_{\setminus s}, y; \theta)$ and $p(y_r = k|y_{\setminus r,}, x; \theta)$ take on the familiar form of linear Gaussian and (multiclass) logistic regression, as we pointed out in (2) and (3). Here are the details:

- The conditional distribution of a continuous variable x_s is Gaussian with a linear regression model

for the mean, and unknown variance.

$$p(x_s|x_{\setminus s}, y; \Theta) = \frac{\sqrt{\beta_{ss}}}{\sqrt{2\pi}} \exp(a) \tag{8}$$
$$a = \frac{-\beta_{ss}}{2}\left(\frac{\alpha_s + \sum_j \rho_{sj}(y_j) - \sum_{t\neq s}\beta_{st}x_t}{\beta_{ss}} - x_s\right)^2$$

- The conditional distribution of a discrete variable y_r with L_r states is a multinomial distribution, as used in (multiclass) logistic regression. Whenever a discrete variable is a predictor, each of its levels contribute an additive effect; continuous variables contribute linear effects.

$$p(y_r|y_{\setminus r,}, x; \Theta) = \frac{\exp(b_{y_r})}{\sum_{l=1}^{L_r}\exp(b_l)} \tag{9}$$
$$b_l = \left(\sum_s \rho_{sr}(l)x_s + \phi_{rr}(l,l) + \sum_{j\neq r}\phi_{rj}(l, y_j)\right)$$

A generic parameter block, θ_{uv}, corresponding to an edge (u, v) appears twice in the pseudolikelihood, once for each of the conditional distributions $p(z_u|z_v)$ and $p(z_v|z_u)$.

Proposition 1. *The negative log pseudolikelihood in* (7) *is jointly convex in all the parameters* $\{\beta_{ss}, \beta_{st}, \alpha_s, \phi_{rj}, \rho_{sj}\}$ *over the region* $\beta_{ss} > 0$.

We prove Proposition 1 in the Supplementary Material.

3.2 Separate node-wise regression

A simple approach to parameter estimation is via separate node-wise regressions; a generalized linear model is used to estimate $p(z_s|z_{\setminus s})$ for each s. Separate regressions were used in [20] for the Gaussian graphical model and [22] for the Ising model. The method can be thought of as an asymmetric form of the pseudolikelihood since the pseudolikelihood enforces that the parameters are shared across the conditionals. Thus the number of parameters estimated in the separate regression is approximately double that of the pseudolikelihood, so we expect that the pseudolikelihood outperforms at low sample sizes and low regularization regimes. The node-wise regression was used as our baseline method since it is straightforward to extend it to the mixed model. As we predicted, the pseudolikelihood or joint procedure outperforms separate regressions; see top left box of Figures 4 and 5. [19, 18] confirm that the separate regressions are outperformed by pseudolikelihood in numerous synthetic settings.

Recent work[2] [26, 27] extend the separate node-wise regression model from the special cases of Gaussian and categorical regressions to generalized linear models, where the univariate conditional distribution of each node $p(x_s|x_{\setminus s})$ is specified by a generalized linear model (e.g. Poisson, categorical, Gaussian). By specifying the conditional distributions, [4] show that the joint distribution is also specified. Thus another way to justify our mixed model is to define the conditionals of a continuous variable as Gaussian linear regression and the conditionals of a categorical variable as multiple logistic regression and use the results in [4] to arrive at the joint distribution in (1). However, the neighborhood selection algorithm in [26, 27] is restricted to models of the form $p(x) \propto \exp\left(\sum_s \theta_s x_s + \sum_{s,t}\theta_{st}x_s x_t + \sum_s C(x_s)\right)$. In particular, this procedure cannot be applied to edge selection in our pairwise mixed model in (1) or the categorical model in (2) with greater than 2 states. Our baseline method of separate regressions is closely related to the neighborhood selection algorithm they proposed; the baseline can be considered as a generalization of [26, 27] to allow for more general pairwise interactions with the appropriate regularization to select edges. Unfortunately, the theoretical results in [26, 27] do not apply to the baseline method, nor the joint pseudolikelihood.

4 Conditional Independence and Penalty Terms

In this section, we show how to incorporate edge selection into the maximum likelihood or pseudolikelihood procedures. In the graphical representation of probability distributions, the absence of an edge $e = (u, v)$ corresponds to a conditional independency statement that variables x_u and x_v are conditionally independent given all other variables [14]. We would like to maximize the likelihood subject to a penalization on the number of edges since this results in a sparse graphical model. In the pairwise mixed model, there are 3 type of edges

1. β_{st} is a scalar that corresponds to an edge from x_s to x_t. $\beta_{st} = 0$ implies x_s and x_t are conditionally independent given all other variables. This parameter is in two conditional distributions, corresponding to either x_s or x_t is the response variable, $p(x_s|x_{\setminus s}, y; \Theta)$ and $p(x_t|x_{\setminus t}, y; \Theta)$.
2. ρ_{sj} is a vector of length L_j. If $\rho_{sj}(y_j) = 0$ for

[2]The current paper was submitted to arXiv.org on May 22nd, 2012 (`http://arxiv.org/abs/1205.5012`). [26] appeared on the homepage of the authors in November 2012 (`http://www.stat.rice.edu/~gallen/eyang_glmgm_nips2012.pdf`) and was published on December 2012 at NIPS 2012. The long version [27] was submitted to arXiv.org on January 17th, 2013 (`http://arxiv.org/abs/1301.4183`).

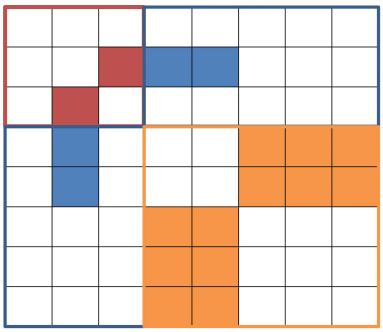


Figure 1: Symmetric matrix represents the parameters Θ of the model. This example has $p = 3$, $q = 2$, $L_1 = 2$ and $L_2 = 3$. The red square corresponds to the continuous graphical model coefficients B and the solid red square is the scalar β_{st}. The blue square corresponds to the coefficients ρ_{sj} and the solid blue square is a vector of parameters $\rho_{sj}(\cdot)$. The orange square corresponds to the coefficients ϕ_{rj} and the solid orange square is a matrix of parameters $\phi_{rj}(\cdot,\cdot)$. The matrix is symmetric, so each parameter block appears in two of the conditional probability regressions.

all values of y_j, then y_j and x_s are conditionally independent given all other variables. This parameter is in two conditional distributions, corresponding to either x_s or y_j being the response variable: $p(x_s|x_{\setminus s}, y; \Theta)$ and $p(y_j|x, y_{\setminus j}; \Theta)$.

3. ϕ_{rj} is a matrix of size $L_r \times L_j$. If $\phi_{rj}(y_r, y_j) = 0$ for all values of y_r and y_j, then y_r and y_j are conditionally independent given all other variables. This parameter is in two conditional distributions, corresponding to either y_r or y_j being the response variable, $p(y_r|x, y_{\setminus r}; \Theta)$ and $p(y_j|x, y_{\setminus j}; \Theta)$.

For edges that involve discrete variables, the absence of that edge requires that the entire matrix ϕ_{rj} or vector ρ_{sj} is 0. The form of the pairwise mixed model motivates the following regularized optimization problem

$$\text{minimize}_\Theta \; \ell_\lambda(\Theta) = \ell(\Theta) \qquad (10)$$
$$+ \lambda \left(\sum_{s=1}^{p} \sum_{t=1}^{s-1} |\beta_{st}| + \sum_{s=1}^{p} \sum_{j=1}^{q} \|\rho_{sj}\|_2 + \sum_{j=1}^{q} \sum_{r=1}^{j-1} \|\phi_{rj}\|_F \right).$$

For scalars, we use the absolute value (l_1 norm), for vectors we use the l_2 norm, and for matrices we use the Frobenius norm. This choice corresponds to the standard relaxation from group l_0 to group l_1/l_2 (group lasso) norm [1, 28].

5 Calibrated regularizers

In (10) each of the group penalties are treated as equals, irrespective of the size of the group. We suggest a calibration or weighting scheme to balance the load in a more equitable way. We introduce weights for each group of parameters and show how to choose the weights such that each parameter set is treated equally under p_F, the fully-factorized independence model [3]

$$\text{minimize}_\Theta \ell_\lambda(\Theta) = \ell(\Theta) + \qquad (11)$$
$$\lambda \left(\sum_{t<s} w_{st}|\beta_{st}| + \sum_{s,j} w_{sj} \|\rho_{sj}\|_2 + \sum_{r<j} w_{rj} \|\phi_{rj}\|_F \right)$$

Based on the KKT conditions [8], the parameter group θ_g is non-zero if

$$\left\| \frac{\partial \ell}{\partial \theta_g} \right\| > \lambda w_g$$

where θ_g and w_g represents one of the parameter groups and its corresponding weight. Now $\frac{\partial \ell}{\partial \theta_g}$ can be viewed as a generalized residual, and for different groups these are different dimensions—e.g. scalar/vector/matrix. So even under the independence model (when all terms should be zero), one might expect some terms $\left\| \frac{\partial \ell}{\partial \theta_g} \right\|$ to have a better random chance of being non-zero (for example, those of bigger dimensions). Thus for all parameters to be on equal footing, we would like to choose the weights w such that

$$E_{p_F} \left\| \frac{\partial \ell}{\partial \theta_g} \right\| = \text{constant} \times w_g$$

However, it is simpler to compute in closed form $E_{p_F} \left\| \frac{\partial \ell}{\partial \theta_g} \right\|^2$, so we choose

$$w_g \propto \sqrt{E_{p_F} \left\| \frac{\partial \ell}{\partial \theta_g} \right\|^2}$$

where p_F is the fully factorized (independence) model. In the Supplementary Material, we show that the weights can be chosen as

$$w_{st} = \sigma_s \sigma_t$$
$$w_{sj} = \sigma_s \sqrt{\sum_a p_a(1-p_a)}$$
$$w_{rj} = \sqrt{\sum_a p_a(1-p_a) \sum_b q_b(1-q_b)}$$

[3] Under the independence model p_F is fully-factorized $p(x, y) = \prod_{s=1}^{p} p(x_s) \prod_{r=1}^{q} p(y_r)$

σ_s is the standard deviation of the continuous variable x_s. $p_a = Pr(y_r = a)$ and $q_b = Pr(y_j = b)$. For all 3 types of parameters, the weight has the form of $w_{uv} = \mathbf{tr}(\mathbf{cov}(z_u))\mathbf{tr}(\mathbf{cov}(z_v))$, where z represents a generic variable and $\mathbf{cov}(z)$ is the variance-covariance matrix of z.

6 Optimization Algorithms

In this section, we discuss two algorithms for solving (10): the proximal gradient and the proximal newton methods. This is a convex optimization problem that decomposes into the form $f(x) + g(x)$, where f is smooth and convex and g is convex but possibly non-smooth. In our case f is the negative log-pseudolikelihood and g are the group sparsity penalties.

6.1 Proximal Gradient

Problems of this form are well-suited for the proximal gradient and accelerated proximal gradient algorithms [6, 3] as long as the proximal operator of g can be computed. The proximal gradient iteration is given by

$$x_{k+1} = prox_t\left(x_k - t\nabla f(x_k)\right)$$

where t is determined by line search and $prox_t(x) = \text{argmin}_u \frac{1}{2t}\left\|x-u\right\|^2 + g(u)$. The theoretical convergence rates and properties of the proximal gradient algorithm and its accelerated variants are well-established [3]. The proximal gradient method achieves linear convergence rate of $O(c^k)$ when the objective is strongly convex and the sublinear rate $O(1/k)$ for non-strongly convex problems.

6.2 Proximal Newton Algorithms

This section borrows heavily from [23], [24] and [16]. The class of proximal Newton algorithms is a 2nd order analog of the proximal gradient algorithms with a quadratic convergence rate [16]. It attempts to incorporate 2nd order information about the smooth function f into the model function. At each iteration, it minimizes a quadratic model centered at x_k

$$\begin{aligned}&\nabla f(x_k)^T(u-x_k) + \frac{1}{2t}(u-x_k)^T H(u-x_k) + g(u)\\ &:= Hprox_t\left(x_k - tH^{-1}\nabla f(x_k)\right)\end{aligned}$$

where $H = \nabla^2 f(x_k)$. The $Hprox$ operator is analogous to the proximal operator, but in the $\|\cdot\|_H$-norm. It simplifies to the proximal operator if $H = I$, but in the general case of positive definite H there is no closed-form solution for many common non-smooth $g(x)$ (including l_1 and group l_1). However if the proximal operator of g is available, each of these subproblems can be solved efficiently with proximal gradient.

Algorithm 1 Proximal Newton

repeat

Solve subproblem

$$p_k = Hprox_t\left(x_k - tH_k^{-1}\nabla f(x_k)\right) - x_k.$$

Find t to satisfy Armijo line search condition

$$f(x_k + tp_k) + g(x_k + tp_k) \leq f(x_k) + g(x_k) - \frac{t\alpha}{2}\left\|p_k\right\|^2$$

Set $x_{k+1} = x_k + tp_k$

$k = k + 1$

until $\frac{\|x_k - x_{k+1}\|}{\|x_k\|} < tol$

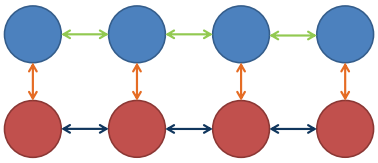
Theoretical analysis in [16] suggests that proximal Newton methods generally require fewer outer iterations (evaluations of $Hprox$) than first-order methods while providing higher accuracy because they incorporate 2nd order information. We have confirmed empirically that the proximal Newton methods are faster when n is very large or the gradient is expensive to compute. The hessian matrix H can be replaced by a quasi-newton approximation such as BFGS/L-BFGS/SR1. In our implementation, we use the `PNOPT` implementation [16].

7 Conditional Model

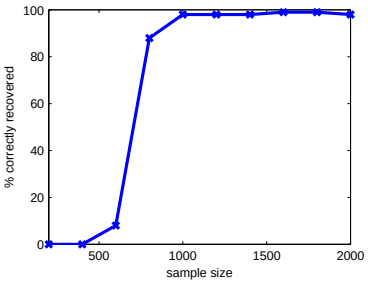
We can generalize our mixed model to include a conditional model by incorporating features; this is a type of conditional random field. Conditional models only model the conditional distribution $p(z|f)$, as opposed to the joint distribution $p(z, f)$, where z are the variables of interest to the prediction task and f are features.

In addition to observing x and y, we observe features f and we build a graphical model for the conditional distribution $p(x, y|f)$. Consider a full pairwise model $p(x, y, f)$ of the form (1). We then choose to only model the joint distribution over only the variables x and y to give us $p(x, y|f)$ which is of the form

$$\exp\left(\sum_{s,t} -\frac{1}{2}\beta_{st}x_s x_t + \sum_s \alpha_s x_s + \sum_{s,j}\rho_{sj}(y_j)x_s + \sum_{r<j}\phi_{rj}(y_r, y_j) + \sum_{s,l}\gamma_{ls}x_s f_l + \sum_{l,r}\eta_{lr}(y_r)f_l\right)$$



(a) True Graph Structure



(b) Probability of recovery

Figure 2: Figure 2a shows a smaller version of the graph used in the synthetic experiments. The graph used in the experiment has 10 continuous and 10 discrete variables. Blue nodes are continuous variables, red nodes are binary variables and the orange, green and dark blue lines represent the 3 types of edges. Figure 2b is a plot of the probability of correct edge recovery at a given sample size. Results are averaged over 100 trials.

8 Experimental Results

We present experimental results on synthetic data, survey data and on a conditional model.

8.1 Synthetic Experiments

In the synthetic experiment, the training points are sampled from a true model with 10 continuous variables and 10 binary variables. The edge structure is shown in Figure 2a. λ is chosen as $5\sqrt{\frac{\log p+q}{n}}$ as suggested in several theoretical studies [22, 13]. We see from the experimental results that recovery of the correct edge set undergoes a sharp phase transition, as expected. With $n = 1000$ samples, we are recovering the correct edge set with probability nearly 1.

8.2 Survey Experiments

The survey dataset we consider consists of 11 variables, of which 2 are continuous and 9 are discrete: age (continuous), log-wage (continuous), year(7 states), sex(2 states),marital status (5 states), race(4 states), education level (5 states), geographic region(9 states), job class (2 states), health (2 states), and health insurance (2 states). The dataset was assembled by Steve Miller of OpenBI.com from the March 2011 Supplement to Current Population Survey data. All the evaluations are done using a holdout test set of size $100,000$ for the survey experiments. The regularization parameter λ is varied over the interval $[5 \times 10^{-5}, .7]$ at 50 points equispaced on log-scale for all experiments.

8.2.1 Model Selection

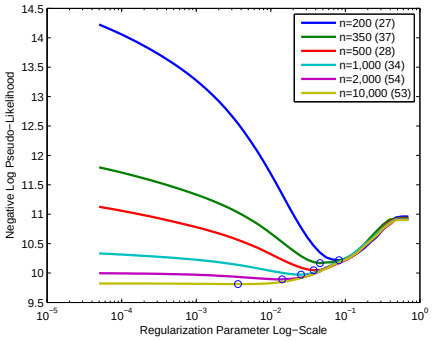


Figure 3: Model selection under different training set sizes. Circle denotes the lowest test set negative log pseudolikelihood and the number in parentheses is the number of edges in that model at the lowest test negative log pseudolikelihood. The saturated model has 55 edges.

In Figure 3, we study the model selection performance of learning a graphical model over the 11 variables under different training samples sizes. We see that as the sample size increases, the optimal model is increasingly dense, and less regularization is needed.

8.2.2 Comparing against Separate Regressions

A sensible baseline method to compare against is a separate regression algorithm. This algorithm fits a linear Gaussian or (multiclass) logistic regression of each variable conditioned on the rest. We can evaluate the performance of the pseudolikelihood by evaluating $-\log p(x_s|x_{\setminus s}, y)$ for linear regression and $-\log p(y_r|y_{\setminus r}, x)$ for (multiclass) logistic regression. Since regression is directly optimizing this loss function, it is expected to do better. The pseudolikelihood objective is similar, but has half the number of parameters as the separate regressions since the coefficients are shared between two of the conditional likelihoods. From Figures 4 and 5, we can see that the pseudolikelihood performs very similarly to the separate regressions and sometimes even outperforms regression. The benefit of the pseudolikelihood is that we have learned parameters of the joint distribution $p(x, y)$ and not just of the conditionals $p(x_s|y, x_{\setminus s})$. On the test

dataset, we can compute quantities such as conditionals over arbitrary sets of variables $p(y_A, x_B | y_{A^C}, x_{B^C})$ and marginals $p(x_A, y_B)$ [14]. This would not be possible using the separate regressions.

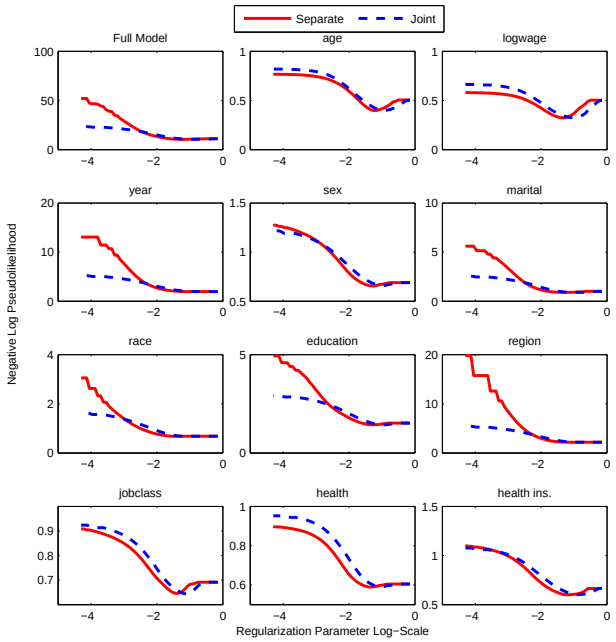


Figure 4: Separate Regression vs Pseudolikelihood $n = 100$. y-axis is the appropriate regression loss for the response variable. For low levels of regularization and at small training sizes, the pseudolikelihood seems to overfit less; this may be due to a global regularization effect from fitting the joint distribution as opposed to separate regressions.

8.2.3 Conditional Model

Using the conditional model, we model only the 3 variables logwage, education(5) and jobclass(2). The other 8 variables are only used as features. The conditional model is then trained using the pseudolikelihood. We compare against the generative model that learns a joint distribution on all 11 variables. From Figure 6, we see that the conditional model outperforms the generative model, except at small sample sizes.

References

[1] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4:1–106, 2011.

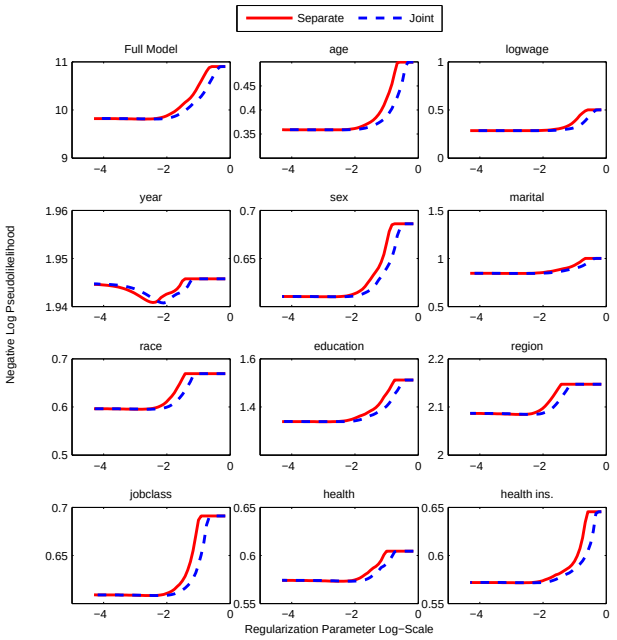


Figure 5: Separate Regression vs Pseudolikelihood $n = 10,000$. y-axis is the appropriate regression loss for the response variable. At large sample sizes, separate regressions and pseudolikelihood perform very similarly. This is expected since this is nearing the asymptotic regime.

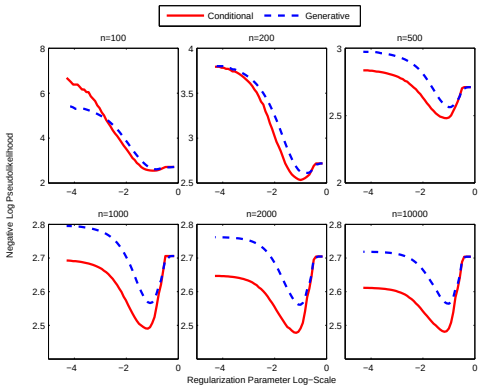


Figure 6: Conditional Model vs Generative Model at various sample sizes. y-axis is test set performance is evaluated on negative log pseudolikelihood of the conditional model.

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

[3] A. Beck and M. Teboulle. Gradient-based algorithms with applications to signal recovery problems. *Convex Optimization in Signal Processing*

and Communications, pages 42–88, 2010.

[4] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

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

[6] P.L. Combettes and J.C. Pesquet. Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212, 2011.

[7] D. Edwards. *Introduction to graphical modelling*. Springer, 2000.

[8] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

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

[10] J. Friedman, T. Hastie, and R. Tibshirani. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Technical report, Technical Report, Stanford University, 2010.

[11] J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint structure estimation for categorical markov networks. *Submitted. Available at http://www. stat. lsa. umich. edu/˜ elevina*, 2010.

[12] H. Höfling and R. Tibshirani. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *The Journal of Machine Learning Research*, 10:883–906, 2009.

[13] A. Jalali, P. Ravikumar, V. Vasuki, S. Sanghavi, UT ECE, and UT CS. On learning discrete graphical models using group-sparse regularization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[14] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.

[15] S.L. Lauritzen. *Graphical models*, volume 17. Oxford University Press, USA, 1996.

[16] J.D. Lee, Y. Sun, and M.A. Saunders. Proximal newton-type methods for minimizing convex objective functions in composite form. *arXiv preprint arXiv:1206.1623*, 2012.

[17] S.I. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of markov networks using l1regularization. In *NIPS*, 2006.

[18] Q. Liu and A. Ihler. Learning scale free networks by reweighted l1 regularization. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[19] Q. Liu and A. Ihler. Distributed parameter estimation via pseudo-likelihood. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

[20] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[21] J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

[22] P. Ravikumar, M.J. Wainwright, and J.D. Lafferty. High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[23] M. Schmidt. *Graphical Model Structure Learning with l1-Regularization*. PhD thesis, University of British Columbia, 2010.

[24] M. Schmidt, D. Kim, and S. Sra. Projected newton-type methods in machine learning. 2011.

[25] M. Schmidt, K. Murphy, G. Fung, and R. Rosales. Structure learning in random fields for heart motion abnormality detection. *CVPR. IEEE Computer Society*, 2008.

[26] E. Yang, P. Ravikumar, G. Allen, and Z. Liu. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems 25*, pages 1367–1375, 2012.

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

[28] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.