Structure Learning of Mixed Graphical Models

Jason D. Lee and Trevor J. Hastie

Stanford University

Discussion Led by Shaobo Han, Duke University
Nov 8, 2013
Outline

1. Mixed Graphical Model
2. Parameter Estimation
3. Penalty Terms
4. Experimental Results
Introduction

- Typical data sources contain both continuous and discrete variables: population survey data, genomics data, url-click pairs etc.
- The census survey dataset\(^1\) consists of 11 variables, 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)

- Problem of interest: learning the edge structure and parameters of sparse undirected graphical models (Markov random fields), over continuous and discrete variables

---

\(^1\)The dataset was assembled by Steve Miller of OpenBI.com from the March 2011 Supplement to Current Population Survey data.
Pairwise Markov random field

Denote \(\{x_s\}_{s=1}^{p}\) the continuous variables, and \(\{y_j\}_{j=1}^{q}\) the discrete variables (discrete \(y_j\) takes \(L_j\) states), the joint density

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

is parameterized by \(\Theta = [\{\beta_{st}\}, \{\alpha_s\}, \{\rho_{sj}\}, \{\phi_{rj}\}]^2\). The model parameters are \(\beta_{st}\) continuous-continuous edge potential, \(\alpha_s\) continuous node potential, \(\rho_{sj}(y_j)\) continuous-discrete edge potential, and \(\phi_{rj}(y_r, y_j)\) discrete-discrete edge potential.

\(^2\rho_{sj}(y_j)\) 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.
Conditional Independence

In the pairwise mixed graphical model, there are 3 types of edges:

1. $\beta_{st}$ is a scalar, $\beta_{st} = 0$ implies $x_s \perp\!\!\!\perp x_t | (x_s, t, y)$
2. $\rho_{sj}$ is a vector of length $L_j$, if $\rho_{sj}(y_j) = 0$ for all values of $y_j$, then $y_j \perp\!\!\!\perp x_s | (x_s, y_j)$
3. $\phi_{rj}$ 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 $y_r \perp\!\!\!\perp y_j | (x, y_r, y_j)$

Figure 1: Symmetric matrix represents the parameter $\Theta$ of the model. This example has $p = 3$, $q = 2$, $L_1 = 2$ and $L_2 = 3$. 
Continuous variables only: multivariate Gaussian distribution

\[ p(x) \propto \exp \left( -\frac{1}{2}(x - \mu)^T B(x - \mu) \right), \quad B = \{\beta_{st}\}, \quad \mu = B^{-1} \alpha \quad (2) \]

Discrete variables only: pairwise Markov random field

\[ p(y) \propto \exp \left( \sum_{j=1}^{q} \sum_{r=1}^{q} \phi_{rj}(y_r, y_j) \right) \quad (3) \]

Sampling from the joint distribution

\[ p(x|y) = \mathcal{N}(B^{-1} \gamma(y), B^{-1}), \quad \{\gamma(y)\}_s = \alpha_s + \sum_j \rho_{sj}(y_j) \]

\[ p(y) \propto \exp \left( \sum_{j=1}^{q} \sum_{r=1}^{j} \phi_{rj}(y_r, y_j) + \frac{1}{2} \gamma(y)^T B^{-1} \gamma(y) \right) \quad (4) \]
Model Properties (2/2)

- Conditional distribution of $x_s$: univariate Gaussian distribution

$$
p(x_s|x_{\setminus s}, y; \Theta) = \mathcal{N}\left( \frac{\alpha_s + \sum_j \rho_{sj}(y_j) - \sum_{t \neq s} \beta_{st} x_t}{\beta_{ss}}, \beta_{ss}^{-1} \right) \tag{5}
$$

- Conditional distribution of $y_r$ (with $L_r$ states): multinomial distribution

$$
p(y_r = k | y_{\setminus r}, x; \Theta) = \frac{\exp \left( \sum_s \rho_{sr}(y_r)x_s + \phi_{rr}(y_r, y_r) + \sum_{j \neq r} \phi_{rj}(y_r, y_j) \right)}{\sum_{\ell=1}^{L_r} \exp \left( \sum_s \rho_{sr}(\ell)x_s + \phi_{rr}(\ell, \ell) + \sum_{j \neq r} \phi_{rj}(\ell, y_j) \right)} \tag{6}
$$
• **Maximum likelihood** Given samples $(x_i, y_i)_{i=1}^n$, find the MLE of $\Theta$ which minimizes $\ell(\Theta) = - \sum_{i=1}^n \log p(x_i, y_i; \Theta)$, where

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

(7)

• **Pseudolikelihood** [Besag (1975)]\(^3\) computationally efficient and consistent estimator formed by products of all the conditionals

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

(8) is **jointly convex** in all the parameters $\{\beta_{ss}, \beta_{st}, \alpha_s, \phi_{rj}, \rho_{sj}\}$ over the region $\beta_{ss} > 0$

Figure 2:  (a) A small 2d lattice. (b) The representation used by pseudo likelihood. Solid nodes are observed neighbors. Based on Figure 19.13 of [Murphy (2012)]

- **Separate node-wise regression** A simple baseline model, fit the univariate conditional distribution of each node $p(x_s | x_{\bar{s}}, y, \Theta)$ or $p(y_r | x, y_{\bar{r}}, \Theta)$ separately by a generalized linear model (e.g. Poisson, categorical, Gaussian)

---

Penalty Terms

For edges that involves discrete terms, the absence of of that edge requires the entire matrix $\phi_{rj}$ or vector $\rho_{sj}$ to be zero. This motivates the following regularized optimization problem

$$\min_\Theta \ell(\Theta) + \lambda \left( \sum_{s<t} 1[\beta_{st} \neq 0] + \sum_{sj} 1[\rho_{sj} \neq 0] + \sum_{r<j} 1[\phi_{rj} \neq 0] \right)$$

This problem is non-convex, so we replace the $\ell_0$ sparsity and group $\ell_0$ sparsity penalties with the group $\ell_1/\ell_2$ (group lasso) norm. [Yuan and Lin (2006)]

$$\min_\Theta \ell(\Theta) + \lambda \left( \sum_{s=1}^{p} \sum_{t=1}^{s-1} |\beta_{st}| + \sum_{s=1}^{p} \sum_{j=1}^{q} ||\rho_{sj}||_2 + \sum_{j=1}^{q} \sum_{r=1}^{j-1} ||\phi_{rj}||_F \right)$$

---

Calibrated Regularizers

Introduce weights $w_g$ such that each parameter set $\theta_g$ is treated equally (under the independence model, $p(x, y) = \prod_{s=1}^p p(x_s) \prod_{r=1}^q p(y_r) := p_F$)

$$
\min_{\Theta} \ell(\Theta) + \lambda \left( \sum_{s=1}^p \sum_{t=1}^{s-1} w_{st} |\beta_{st}| + \sum_{s=1}^p \sum_{j=1}^q w_{sj} \|\rho_{sj}\|_2 + \sum_{j=1}^q \sum_{r=1}^{j-1} w_{rj} \|\phi_{rj}\|_F \right)
$$

KKT conditions [Friedman et al. (2007)]\textsuperscript{7}: the parameter group $\theta_g$ is non-zero if $\|\partial\ell/\partial \theta_g\| > \lambda w_g$. We wish $E_{p_F} \|\partial\ell/\partial \theta_g\| = \text{constant} \times w_g$. However, it is simpler to compute in closed form $E_{p_F} \|\partial\ell/\partial \theta_g\|^2$, so we choose $w_g \propto \sqrt{E_{p_F} \|\partial\ell/\partial \theta_g\|^2}$

$$
w_{st} = \sigma_s \sigma_t, \quad w_{sj} = \sigma_s \sqrt{\sum_a p_a (1 - p_a)}, \quad p_a = \text{Pr}(y_r = a)
$$

$$
w_{rj} = \sqrt{\sum_a p_a (1 - p_a) \sum_b q_b (1 - q_b)}, \quad q_b = \text{Pr}(y_j = b) \tag{9}
$$

Synthetic Experiments

Data are sampled from a true model with 10 continuous variables and 10 binary variables, $\lambda = 5\sqrt{\log p + q}/n$.

Figure 3: (a) Graphical structure with 4 continuous and 4 discrete variables. (b) Probability of recovery. Results are averaged over 100 trials.
Figure 4: Pseudolikelihood outperforms maximum likelihood across all sample sizes on both evaluation criteria.

- Although asymptotic theory suggests that maximum likelihood is more efficient than the pseudolikelihood, this analysis is applicable because of the finite sample regime and misspecified model.
The regularization parameter $\lambda$ is varied over the interval $[5 \times 10^{-5}, 0.7]$ at 50 points equispaced on log-scale all experiments.

**Figure 5**: Model selection under different training set sizes. All the evaluations are done using a holdout test set of size 100,000 for the survey experiments. The saturated model has 55 edges.
Survey Experiments (3/4): Separate Regression v.s. Pseudolikelihood

Figure 6: For low levels of regularization and at small training sizes, the pseudolikelihood seems to overfit less. For large sample sizes, they perform similar.

\[ p(x, y|f; \Theta) \propto \exp \left( \sum_{s,t} -\frac{1}{2} \beta_{st} x_s x_t + \sum_{s} \alpha_s x_s + \sum_{s,j} \rho_{sj}(y_j x_s) \right) \]

\[ + \sum_{r<j} \phi_{rj}(y_r, y_j) + \sum_{s,\ell} \gamma_{s\ell} x_s f_{\ell} + \sum_{\ell, r} \eta_{\ell r}(y_r) f_{\ell} \]  

(10)

Figure 7: Using the conditional model (10), we model only 3 variables log wage, education (5) and jobclass (2); The other 8 variables are only used as features.