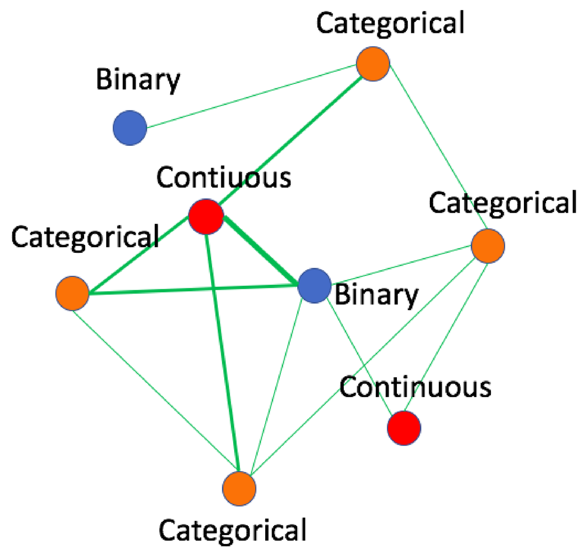
CS229
Machine Learning
Autumn 2016

Learning the Network Structure of Heterogeneous Data via Pairwise Exponential MRF

Jongho Kim, Youngsuk Park
Email:{jkim22, youngsuk}@stanford.edu

Motivation

- Typical Multivariate distributions (e.g. GMRF and Ising model) can model data on the homogeneous domain.
- But, the entities of real data can have heterogeneous domain (e.g. binary, category, and continuous domain).
- Need to define a proper multivariate distribution for heterogeneous data.
- Propose a subset of multivariate exponential family distribution, pairwise exponential Markov Random Fields, able to reveal the Markov (Network) structure across the entities.



Model: Pairwise Exponential Markov Random Field

- For a Markov Random Fields described by $G = (V, E)$ with $|V| = p$, a random vector $X = \{X_1, \ldots, X_p\}$ is defined over (heterogeneous) domains $\mathcal{X} = \otimes\{\mathcal{X}_r\}_{r=1}^p$.
- Let each node X_r have $B_r(X_r)$ be (m_r-dimensional) node-potential and scalar base measure $C_r(X_r)$, by which we can design the conditional distribution on X_r given others $X_{\setminus r}$.

Definition A random vector X is defined as a PE-MRF if, for $x = \{x_1, \ldots, x_p\} \in \mathcal{X}$, it follows the joint distribution

$$p(x;\boldsymbol{\theta}) = \exp\{\sum_{r=1}^{p} \theta_r^T B_r(x_r) + \sum_{s,t=1}^{p} \left\langle \Theta_{st}, B_t(x_t)B_s(x_s)^T \right\rangle_F + \sum_{r=1}^{p} C_r(x_r) - A(\boldsymbol{\theta})\}.$$

- $\boldsymbol{\theta}$ is the natural parameter with node-parameter $\theta_r \in \mathbb{R}^{m_r}$ and edge-parameter $\Theta_{st} \in \mathbb{R}^{m_s \times m_t}$.
- The log-partition function $A(\theta)$ should be finite.

Property.

- PE-MRF is exponential family where sufficient statistics and natural parameter interact linearly and nodes have a pairwise relationship.
- Include well-known distributions such as GMRF, Ising model, discrete model.
- The $\{\Theta_{st}\}_{s,t=1}^p$ explicitly reveal underlying Markov structure.

Estimation: Approximated Maximum Likelihood

- Maximum likelihood would be a typical approach but it entails intractable $A(\boldsymbol{\theta})$.
- Instead, replace $A(\theta)$ with a tractable upperbound $U(\boldsymbol{\theta})$

$$\underset{\theta}{\text{minimize}} \quad \underbrace{-\langle \boldsymbol{\theta}, \overline{\boldsymbol{B}(\mathbf{x})} \rangle + U(\boldsymbol{\theta})}_{\text{approximated likelihood function}} + \underbrace{R_\lambda(\boldsymbol{\theta})}_{\text{group lasso}} .$$

- Here, $\overline{\boldsymbol{B}(\mathbf{x})}$ is the averaged sufficient statistic $B(x^{(i)})$ over the n samples. And the group lasso $R_\lambda(\boldsymbol{\theta}) = \lambda \sum_{s \neq t} w_{st} \|\Theta_{st}\|_F$ [2] encourages the st-th block, for every $s \neq t$, to be a zero matrix.
- Derive a convex upperbound $U(\boldsymbol{\theta})$. (Sketch: Express $A(\theta)$ as an entropy $H(X)$, simplify it into $\equiv H(B_1(X_1), \ldots, B_p(X_p))$, and use the fact that Gaussian distribution gives maximum entropy among any pdf under the same covariance matrix.)

Theorem. For a PE-MRF, the approximated negative maximum log-likelihood problem becomes

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

where $d = \sum_{r=1}^p m_r$, $\boldsymbol{\Theta}$ is the augmented natural parameter of $\boldsymbol{\theta}$, $\overline{\mathbf{B}_{aug}(\mathbf{x})}$ is augmented and shifted version of $\overline{\boldsymbol{B}(\mathbf{x})}$, and $R_\lambda(\boldsymbol{\Theta}) \equiv R_\lambda(\boldsymbol{\theta})$.

- Call it *group graphical lasso* that extends the classic *graphical lasso* problem to general setting.

Algorithm

- Formulate the problem below and solve via alternating direction method of multipliers (ADMM) [1]

$$\min_{\boldsymbol{Z}=\boldsymbol{\Theta}, \boldsymbol{\Theta} \in \mathbf{S}_{++}^{d+1}} \langle \boldsymbol{\Theta}, A \rangle_F - \log\det \boldsymbol{\Theta} + \lambda_n \sum_{i \neq j} w_{ij} \|Z_{ij}\|_F .$$

- ADMM solves the augmented Lagrangian in iterative manner with respect to $\boldsymbol{\Theta}, \boldsymbol{Z}$, and $\boldsymbol{U}$ where $\boldsymbol{U}$ is the scaled dual variables.
- At kth iteration, all of updates have the following the closed form:
$\boldsymbol{\Theta}$ Update. $\boldsymbol{\Theta}^{k+1} := 1/2\eta Q\left(\Lambda + \sqrt{\Lambda^2 + 4\eta I}\right)Q^T$ where $\eta = \rho/n$ and $Q\Lambda Q^T$ is the eigendecomposition of $\eta(\boldsymbol{Z}^k - \boldsymbol{U}^k) - \overline{\mathbf{B}_{aug}(\mathbf{x})}$ [5].
$\boldsymbol{Z}$-Update. For $i, j \in \{1, \ldots, p\}$, compute $Z_{ij, i \neq j} = \left(1 - \frac{\lambda_n w_{ij}}{\rho \|\Theta_{ij}^{k+1} + U_{ij}^k\|_F}\right)\left(\Theta_{ij}^{k+1} + U_{ij}^k\right)$ if $\left\|\Theta_{ij}^{k+1} + U_{ij}^k\right\|_F \geq \lambda_n w_{ij}/\rho$ or 0 otherwise. And $\boldsymbol{Z}^{k+1} := \boldsymbol{\Theta}^{k+1} + \boldsymbol{U}^k$ for the rest elements.
$\boldsymbol{U}$-Update. $\boldsymbol{U}^{k+1} := \boldsymbol{U}^k + \boldsymbol{\Theta}^{k+1} - \boldsymbol{Z}^{k+1}$.

Edge Recovery Consistency

- Assume a PE-MRF X satisfies the graph condition (Incoherence condition on the Hessian matrix [3] and graph structure) and boundness condition on sufficient statistics.

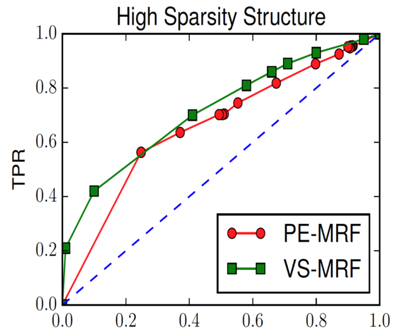
Theorem. For $\lambda_n \geq \kappa_1 \sqrt{\frac{\log(m_{max}p)}{n}}$, let $\hat{\boldsymbol{\Theta}}$ be the (unique) solution of *group graphical lasso*. If the number of samples $n \geq \kappa_2 \log m_{max} p$, then the estimated edge $E(\hat{\boldsymbol{\Theta}}) = \{(s,t) \mid \left\|\hat{\Theta}_{ij}\right\|_2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

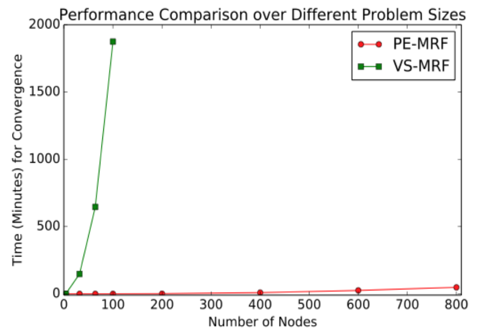
- Here κ_1, κ_2 and κ_3 depend on the $\{m_r\}, \{w_{st}\}$ and other parameters defined in assumptions, and c is some universal constant.

Experiments

Synthetic Data.

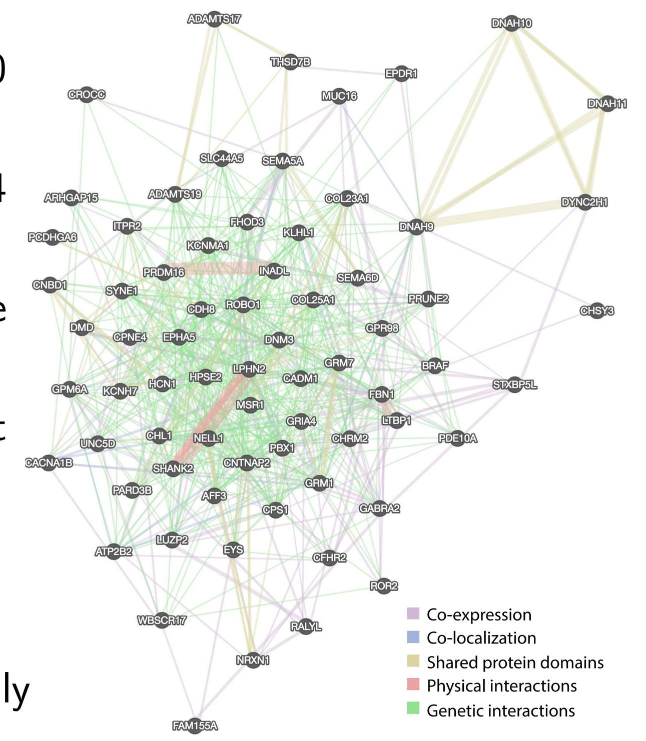
- Experiment on (synthetic) heterogeneous network with 32 nodes: 8 Bernoulli, 8 gamma, 8 Gaussian, and 8 three-dimensional Dirichlet.
- Plot the ROC curves for edge recovery percentage, compared with VS-MRF [4].
- For inferring the 100-node Markov network, our PE-MRF solver can takes **under 30 seconds** (whereas VS-MRF takes **over 31 hours**).





Heterogeneous Genomic Networks.

- Level III public data from The Cancer Genome Atlas for 290 breast cancer patients.
- Contains expression profiles for 500 genes (Gaussian) and 314 miRNAs (Poisson).
- Consider a 65-core of the gene-gene subnetwork to demonstrate the its utility.
- For validation, observe how many gold standard edges exist between 65-core genes from external data.



Conclusion

- Propose a PE-MRF model, a subclass of the exponential family that is well-suited for heterogeneous multivariate distributions.
- Formulate an approximated maximum likelihood problem by deriving upper bound on the log partition function.
- Develop an $O(p^3)$ ADMM algorithm with closed-form updates.
- The estimator guarantees to recovery underlying Markov structure consistently.
- Our results, as well as the widespread applications with heterogeneous data sources, lead to many extensions of this work. For example, instead of inferring a single PE-MRF network, we could use the time-stamped observations to estimate a time-varying network because it is possible that Markov structure changes over time.

References.

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

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

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

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

[5] D. M. Witten and R. Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636, 2009.