Mixed Graphical Models via Exponential Families

Pradeep Ravikumar

Joint with E. Yang, Y. Baker, G. Allen, Z. Liu

University of Texas at Austin

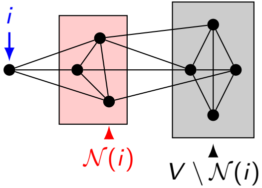
Conference on Applied Statistics in Defense 2014

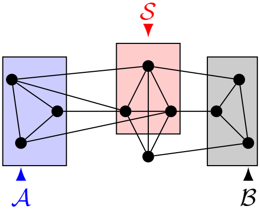
Graphical Models : Introduction

- A **multivariate distribution** over a large number of variables can be represented using a graph $G = (V, E)$ (with $|V| = p$)
- Graph Nodes $i \in V$ correspond to **random variables** X_i
- Graph Edges E encode
  - Correlations?
  - Causations?
  - **Markov Independence** Relationships!

Graphical Models : Introduction

- A **multivariate distribution** over a large number of variables can be represented using a graph $G = (V, E)$ (with $|V| = p$)
- Graph Nodes $i \in V$ correspond to **random variables** X_i
- Graph Edges E encode
  - Correlations?
  - Causations?
  - **Markov Independence** Relationships!



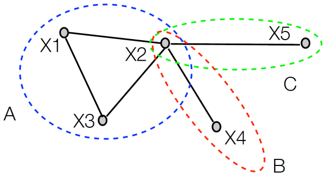


$X_i \perp\!\!\!\perp X_{V \setminus \mathcal{N}(i)} \mid X_{\mathcal{N}(i)}$ $\quad$ $X_{\mathcal{A}} \perp\!\!\!\perp X_{\mathcal{B}} \mid X_{\mathcal{S}}$

Graphical Models: Factorization

- (Hammersley-Clifford theorem) Joint dist. is the product of **local factors**: each of which depends only on a clique

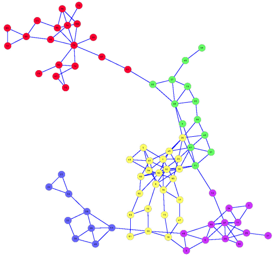
$$P(X) = \frac{1}{Z}\Psi_A(X_A)\Psi_B(X_B)\Psi_C(X_C)$$



MRFs for Categorical Data

- **Categorical data:** $X_s \in \{0, 1, ..., K\}$
- **Potts Model:**

$$\mathbb{P}(X) \propto \exp\left\{ \sum_{(s,t)\in E} \theta_{st} \, \mathbb{I}(X_s = X_t) \right\}$$

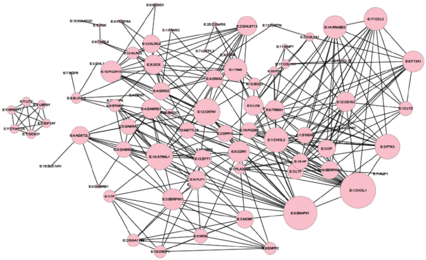


- Other Discrete Data MRFs: Ising MRF, Overcomplete Discrete MRF
- Applications: Internet data, Genomics data, Image processing, Marketing, Statistical physics, ...

MRFs for Thin-tailed Continuous Data

- **(Thin-tailed) Continuous data:** $X_s \in \mathbb{R}$
- **Gaussian MRFs:**

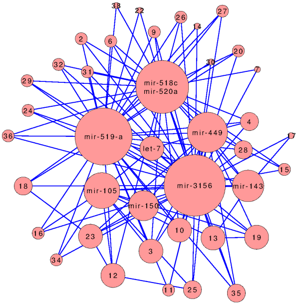
$$\mathbb{P}(X) \propto \exp\left\{ -\frac{1}{2}\langle\langle \Theta, XX^T \rangle\rangle + \langle \theta, X \rangle \right\}$$



MRFs for Count-valued Data

- **Count data:** $X_s \in \{0, 1, \ldots\}$
- **Poisson MRFs:** (Yang et al., 2012,13)

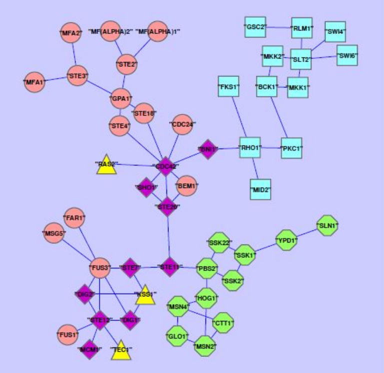
$$\mathbb{P}(X) \propto \exp\left\{ \sum_s \theta_s X_s + \sum_{(s,t)\in E} \theta_{st}\, X_s\, X_t - \sum_s \log(X_s!) \right\}$$



Mixed Data

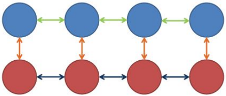
- What if we jointly observe heterogeneous/mixed variables of many **different types?**

  - SNPs: discrete data
  - Gene Expression: continuous data
  - RNA sequencing: count data
  - Genomic Mutations: binary data



- Need multivariate MRFs that permit **dependencies over mixed variables**!

Existing Models for Mixed Data Types



- Limited to **Gaussian-Discrete** case: a continuous random vector, conditioned on a discrete random vector, distributed as multivariate Gaussian.
  - Formulated by Lauritzen (1996), where they called these "conditional Gaussian MRFs"
  - Structure can be learnt tractably (Lee and Hastie (2012))
  - Extension to three-way interactions (Cheng, Levina, Zhu (2012))

Extending Heterogenous Univariate Distributions to Multivariate Graphical Models

- Need a general class of mixed graphical models
- **(KEY QUESTION)** Can we go systematically **from heterogenous univariate dist.** to **multivariate mixed MRFs**?

Review: Exponential Families

- Most common **univariate** distributions: Gaussian, Exponential, Bernoulli, Binomial, Poisson, Negative binomial, ...
- A broad class of distributions sharing a certain form:

$$P(X;\theta) = \exp\left\{\sum_{i\in\mathcal{I}} \theta_i B_i(X) + C(X) - A(\theta)\right\}$$

- Ingredients:

$\theta = \{\theta_i\}_{i\in\mathcal{I}}$	Parameters
$B(X) = \{B_i(X)\}_{i\in\mathcal{I}}$	Sufficient statistics
$C(X)$	Base measure
$A(\theta) = \log\left\{\sum_X \exp\langle\theta, B(X)\rangle + C(X)\right\}$	Log-partition function

Heterogeneous Univariate Exponential Families $\rightarrow$ Mixed Graphical Models

- We know
  - Gaussian graphical models: each X_s is Gaussian dist. given all neighbors
  - Ising models: each X_s is Bernoulli dist. given all neighbors
  - Poisson models: each X_s is Poisson dist. given all neighbors
- Introduce a new class of graphical models:
  - Given $X_{V \setminus s}$, each variable X_s follows a **potentially different** univariate exponential family distribution of choice
  - Dependencies between $(X_1, ..., X_p)$ modeled via graphical model structure.

Heterogeneous Univariate Exponential Families → Mixed Graphical Models

- We know
  - Gaussian graphical models: each X_s is Gaussian dist. given all neighbors
  - Ising models: each X_s is Bernoulli dist. given all neighbors
  - Poisson models: each X_s is Poisson dist. given all neighbors
- Introduce a new class of graphical models:
  - Given $X_{V \setminus s}$, each variable X_s follows a **potentially different** univariate exponential family distribution of choice
  - Dependencies between $(X_1, ..., X_p)$ modeled via graphical model structure.

What form would such a joint distribution take — if any?

Mixed Exponential Family Markov Random Fields

- The most **general** assumptions:
  - Allow **heterogenous** node-conditional distributions:

$P(X_s \mid X_{V\setminus s})$	$=$	$\exp\{E_s(X_{V\setminus s})\, B_s(X_s) + C_s(X_s) - \bar{A}_s(X_{V\setminus s})\}$
$E_s(X_{V\setminus s})$		Parameters
$B_s(X)$		Sufficient statistics
$C_s(X)$		Base measure
$\bar{A}_s(\theta)$		Log-partition function

  - Joint distribution over X given by a graphical model with **factors of size** $\leq k$

$$P(X) \propto \prod_{c \in \mathcal{C}} \Psi_c(X_c).$$

Mixed Exponential Family Markov Random Fields

Theorem

Under the previous general conditions that (a) the node conditional distributions belong to exponential families and (b) the joint is a graphical model with factors of size at most k,

Joint dist. **necessarily** *has the following form:*

$$P(X) = \exp\Bigg\{ \sum_{s} \theta_s B_s(X_s) + \sum_{s\in V} \sum_{t\in N(s)} \theta_{st} \, B_s(X_s) B_t(X_t) + \ldots$$
$$+ \sum_{s\in V} \sum_{t_2,\ldots,t_k \in N(s)} \theta_{s\ldots t_k} \, B_s(X_s) \prod_{j=2}^{k} B_{t_j}(X_{t_j}) + \sum_{s} C_s(X_s) - A(\theta) \Bigg\}$$

Multivariate Graphical Models for Different Types

- Given multiple variables of varied types, only need to specify k, $\{B_s(X)\}_{s \in V}$ and $\{C_s(X)\}_{s \in V}$
- Some could be **time interval** data: time spent on website, networks call time, etc., yet others could be **count** data: incident reports, websites visit counts, next generation genomic data based on RNA fragment counts, etc.), and so on
- The mixed MRF would provide a **joint distribution** over all of these heterogeneous random variables.

Examples: Mixed MRFs with Two Data Types

- Gaussian - Ising Graphical Models

$$P(Y,Z) \propto \exp\Big\{ \sum_{s \in V_Y} \frac{\theta_s^Y}{\sigma_r} Y_s + \sum_{s' \in V_Z} \theta_{s'}^z Z_{s'} + \sum_{(s,t) \in E_Y} \frac{\theta_{st}^{yy}}{\sigma_r \sigma_t} Y_s Y_t + \sum_{(s',t') \in E_Z} \theta_{s't'}^{zz} Z_{s'} Z_{t'} + \sum_{(s,s') \in E_{YZ}} \frac{\theta_{ss'}^{yz}}{\sigma_r} Y_s Z_{s'} - \sum_{s \in V_Y} \frac{Y_s^2}{2\sigma_r^2} \Big\}.$$

- Poisson - Ising Graphical Models

$$P(Y,Z) \propto \exp\Big\{ \sum_{s \in V_Y} \theta_s^Y Y_s + \sum_{s' \in V_Z} \theta_{s'}^z Z_{s'} + \sum_{(s,t) \in E_Y} \theta_{st}^{yy} Y_s Y_t + \sum_{(s',t') \in E_Z} \theta_{s't'}^{zz} Z_{s'} Z_{t'} + \sum_{(s,s') \in E_{YZ}} \theta_{ss'}^{yz} Y_s Z_{s'} - \sum_{s \in V_Y} \log(Y_s!) \Big\}. \quad (1)$$

Examples: Mixed MRFs with Two Data Types

- Gaussian - Poisson Graphical Models ?

$$P(Y,Z) \propto \exp\Bigg\{ \sum_{s \in V_Y} \frac{\theta_s^Y}{\sigma_r} Y_s + \sum_{s' \in V_Z} \theta_{s'}^z Z_{s'} + \sum_{(s,t)\in E_Y} \frac{\theta_{st}^{yy}}{\sigma_r \sigma_t} Y_s Y_t$$

$$+ \sum_{(s',t')\in E_Z} \theta_{s't'}^{zz} Z_{s'} Z_{t'} + \sum_{(s,s')\in E_{YZ}} \frac{\theta_{ss'}^{yz}}{\sigma_r} Y_s Z_{s'} - \sum_{s \in V_Y} \frac{Y_s^2}{2\sigma_r^2} - \sum_{s' \in V_Z} \log(Z_{s'}!) \Bigg\}.$$

Examples: Mixed MRFs with Two Data Types

- Gaussian - Poisson Graphical Models ?

$$P(Y,Z) \propto \exp\Bigg\{ \sum_{s\in V_Y} \frac{\theta_s^Y}{\sigma_r} Y_s + \sum_{s'\in V_Z} \theta_{s'}^z Z_{s'} + \sum_{(s,t)\in E_Y} \frac{\theta_{st}^{yy}}{\sigma_r \sigma_t} Y_s Y_t$$

$$+ \sum_{(s',t')\in E_Z} \theta_{s't'}^{zz} Z_{s'} Z_{t'} + \sum_{(s,s')\in E_{YZ}} \frac{\theta_{ss'}^{yz}}{\sigma_r} Y_s Z_{s'} - \sum_{s\in V_Y} \frac{Y_s^2}{2\sigma_r^2} - \sum_{s'\in V_Z} \log(Z_{s'}!) \Bigg\}.$$

Corollary

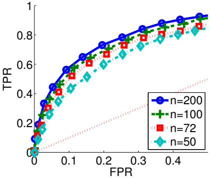
The Gaussian-Poisson distribution is **not normalizable** *unless $\theta_{st} = 0$ for all $(s,t) \in E_{YZ}$.*

Normalizability Conditions for Manichean Graphical Models

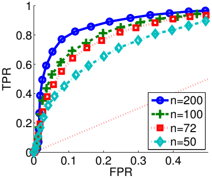
We provide conditions **characterizing normalizability** of general mixed graphical models.

Experiments: Simulated Data

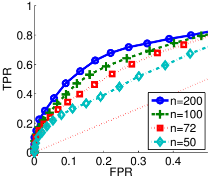
- Lattice graphs, with $p = 72$: $p_Y = 36$ and $p_Z = 36$



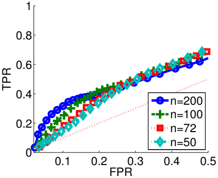
(a) Poisson-Ising



(b) Gaussian-Ising



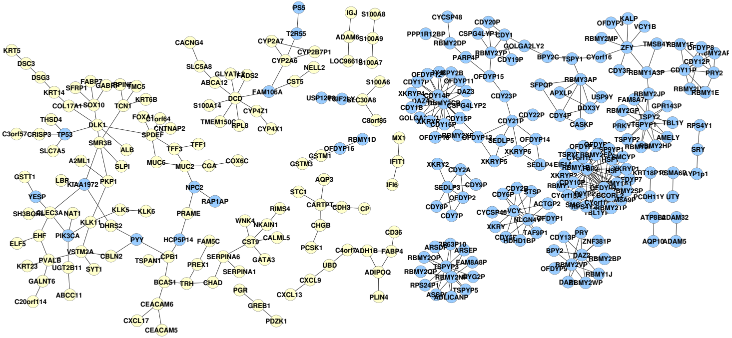
(c) TPGM-Ising



(d) TPGM-Gaussian

Experiments: Cancer Genomic and Transcriptomic Data

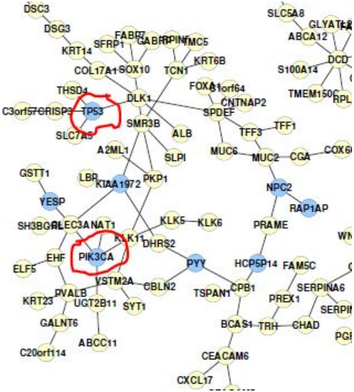
- Combine 'Level III RNA-sequencing' data and 'Level II non-silent somatic mutation and level III copy number variation data' for 697 breast cancer patients.



- TPGM - Ising graphical model
- (Yellow) Gene expression via RNA-sequencing, count-valued
- (Blue) Genomic mutation, binary mutation status
- Well known components: (DLK1, THSD4) - (TP53)

Experiments: Cancer Genomic and Transcriptomic Data

- Combine ‘Level III RNA-sequencing’ data and ‘Level II non-silent somatic mutation and level III copy number variation data’ for 697 breast cancer patients.



- TPGM - Ising graphical model
- (Yellow) Gene expression via RNA-sequencing, count-valued
- (Blue) Genomic mutation, binary mutation status
- Well known components: (DLK1, THSD4) - (TP53)

Summary

- **Broadens the class of off-the-shelf graphical models**, and provides a flexible multivariate modeling toolkit for **mixed data**
  - Univariate exp. family $\rightarrow$ multivariate **Mixed** MRFs
- Allows us to use graphical model machinery to model dependencies for a broader range of data, where each variable may belong to a **potentially different type**
- Can **estimate such graphical models/networks** under standard regularity conditions
- Software (R package) coming soon!

Thank you!

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

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

M.S. Kaiser and N. Cressie. Modeling poisson variables with positive spatial dependence. *Statistics & Probability Letters*, 35(4):423–432, 1997.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Dirty statistical models. In *Neur. Info. Proc. Sys. (NIPS)*, 26, 2013.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. In *48th Annual Allerton Conference on Communication, Control and Computing*, 2010.