A General Framework for Mixed Graphical Models

Genevera I. Allen

Dobelman Family Junior Chair,
Department of Statistics and Electrical and Computer Engineering, Rice University,
Department of Pediatrics-Neurology, Baylor College of Medicine,
Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital.

June 4, 2014

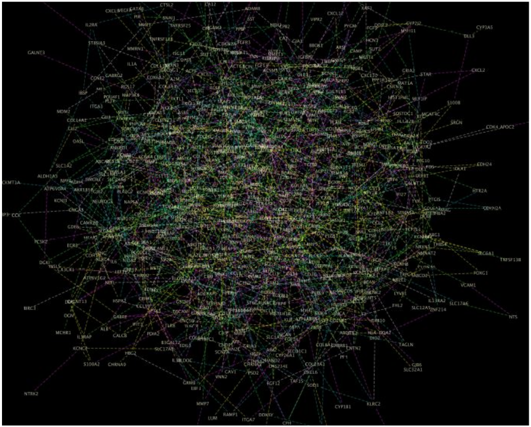
Joint work with Pradeep Ravikumar, Eunho Yang, Yulia Baker, Zhandong Liu and Ying-Wooi Wan.

Motivation: Mixed, Big Data

Mixed Data: Heterogeneous data types (e.g. continuous, skewed continuous, binary, categorical, counts, ordinal).

Examples:

- National Security.
- Internet Data and Advertising.
- Biomedical Imaging.
- Climate data.
- Genomics.



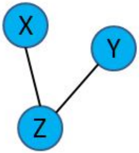
Visualization of mutations and functional genomic interactions in Glioblastoma

Markov Random Fields

- $X = (X_1, X_2, ..., X_p)$ a random vector.
- A graph G represented by a pair $(\mathbf{V},\mathbf{E})$.
  - $\mathbf{V}$: finite vertex set.
  - $\mathbf{E} \subset \mathbf{V} \times \mathbf{V}$: edge set.

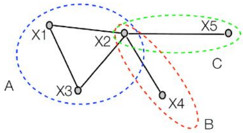
Undirected graphical models or pair-wise Markov Random Fields.

- Captures direct dependencies.
- No edge => conditional independence (pair-wise).



$$(X, Y) \notin E \iff X \perp\!\!\!\perp Y \mid \text{all other variables}$$

- Hammersley-Clifford Theorem: Density on graph factorizes according to sufficient statistics on cliques



$$p(X) = \frac{1}{Z}\psi_A(X_A)\psi_B(X_B)\psi_C(X_C)$$

1 Introduction

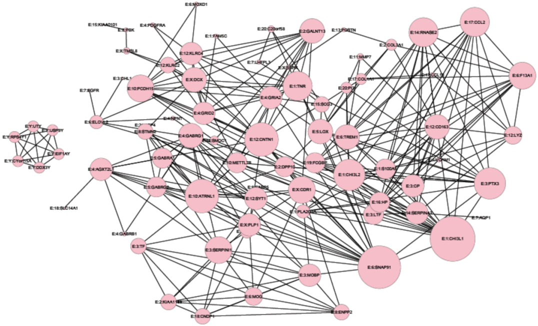
2 Graphical Models via Exponential Families

- Graphical Models via Exponential Families
- Mixed Graphical Models

3 Results

Motivation: Networks from RNA-Sequencing Data

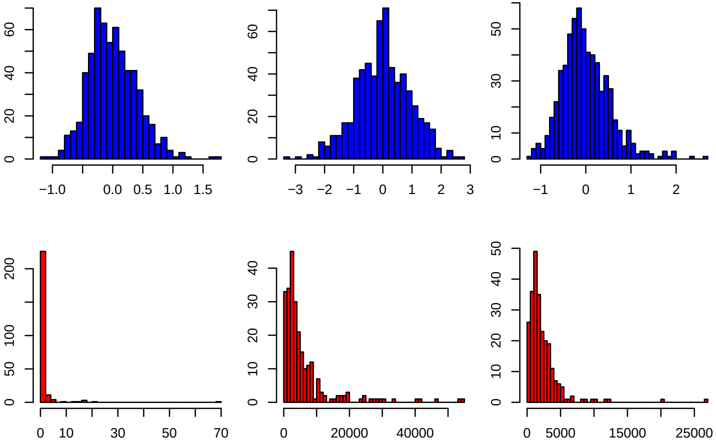
Gaussian Graphical Models have been widely used to infer genomic networks from microarray data:



Applications of Inferred Networks: Visualizing data, discovering biomarkers (hubs), regulatory pathways, potential drug targets.

Motivation: Networks from RNA-Sequencing Data

Next generation sequencing technology is rapidly replacing the microarray.



Gaussian Graphical Models not appropriate for next generation sequencing (RNA-seq) data!

Graphical Models from Count or Other Data Types?

1. Gaussian Graphical Model.
    - Conditional distributions are Gaussian, jointly multivariate Gaussian.
    - Sparse Graphical Model Estimation. (Meinshaussen & Buhlmann, 2006; Yuan & Lin, 2007; Banerjee *et al.*, 2008; Friedman *et al.*, 2008)
2. Ising & Potts Model.
    - Assumes node-conditional distributions are binomial / multinomial.
    - Sparse Graphical Model Estimation. (Ravikumar *et al.*, 2010)
3. Mixed Gaussian - Ising Model.
    - Graphical Models (Lauritzen (1996)).
        - ★ Continuous variables conditioned on all combos discrete variables are multivariate Gaussian.
        - ★ Scales exponentially.
    - Learning the Structure of Mixed Graphical Models (Lee and Hastie (2012)).
    - High-Dimensional Mixed Graphical Model (Cheng, Levina, Zhu (2013)).

Review: Univariate Exponential Families

Examples:

- Gaussian, Bernoulli, Poisson, Binomial, Negative Binomial, Exponential, ...

$$P(Z) = \exp\left(\theta\, B(Z) + C(Z) - D(\theta)\right)$$

- θ is the canonical parameter.
- $B(Z)$ is the sufficient statistic.
- $C(Z)$ is the base measure.
- $D(\theta)$ is the log-partition function.

Graphical Models via Exponential Families

For a random vector $X = (X_1, X_2, \dots X_p)$, suppose:

- Node-conditional distributions are univariate exponential family densities.
- Cliques are of order at most k.

Theorem

Joint Density necessarily has the form:

$$P(X) = \exp\Bigg\{ \sum_s \theta_s B(X_s) + \sum_{s\in V} \sum_{t \in N(s)} \theta_{st}\, B(X_s) B(X_t)$$

$$+ \sum_{s\in V} \sum_{t_2,\dots,t_k \in N(s)} \theta_{s\dots t_k}\, B(X_s) \prod_{j=2}^{k} B(X_{t_j}) + \sum_s C(X_s) - A(\theta) \Bigg\}$$

$N(s)$ denotes the neighborhood of node s & $A(\theta)$ is the log-normalization term.

Graphical Models via Exponential Families

Special Case:

- Cliques of order at most $k = 2$ (pair-wise interactions).
- Linear sufficient statistics $B(X_s) = X_s$.

Joint Density

$$P(X) = \exp\left\{ \sum_s \theta_s X_s + \sum_{(s,t)\in E} \theta_{st}\, X_s\, X_t + \sum_s C(X_s) - A(\theta) \right\}.$$

Node-Conditional Density

$$P(X_s | X_{V\setminus s}) \propto \exp\left\{ \left(\theta_s + \sum_{t\in N(s)} \theta_{st} X_t \right) X_s + C(X_s) \right\},$$

i.e. a Generalized Linear Model.

Graphical Models via Exponential Families

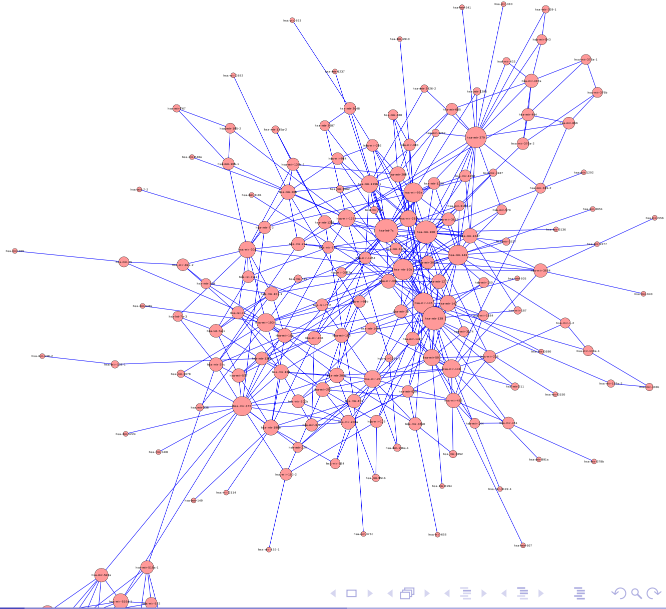
Example of Poisson Graphical Model (Count Data):

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

- Technical conditions needed to ensure proper densities.
- Other examples of novel graphical models:
  - Variations of Poisson case: Truncation, Sub-linear, Quadratic, and approximations to these.
  - Exponential, Gamma, Negative Binomial, etc.

Results: Breast Cancer microRNA Network

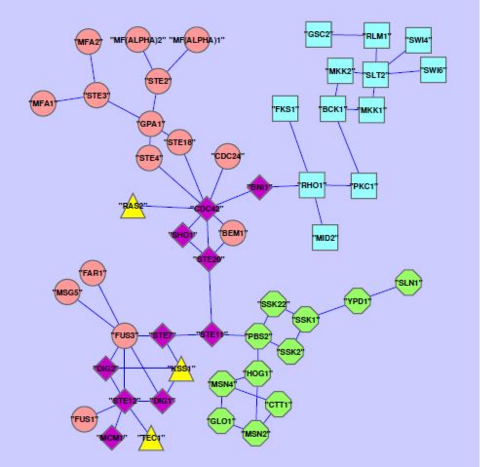
- The Cancer Genome Atlas (TCGA) Level III Data.
- 544 tumor samples, 524 miRNAs.
- miRNA-sequencing (counts).



Motivation: Big, Mixed Genomics Data

TCGA Genomics Data:

- SNPs / Copy Number Variation
  - binary or discrete data.
- Gene Expression (via RNA Sequencing)
  - count data.
- Methylation
  - continuous data.
- Other data types:
  - microRNA expression
  - Proteomics



No general multivariate density that directly parameterizes dependencies for mixed variables exists!

Mixed Graphical Models

Building Mixed MRFs:

- p-variate random response vector

$$X := (X_1, ..., X_p), X_r \in \mathcal{X}_r$$

- $\{\mathcal{X}_r\}_{r \in V}$ potentially all distinct data types.
- Node-Conditional Distribution $P(X_r | X_{V \setminus r})$ is specified via Univariate Exponential Family $\Longrightarrow$ consistent joint density

$$P(X_r | X_{V \setminus r}) = \exp\left(E_r(X_{V \setminus r}) B_r(X_r) + C_r(X_r) - \bar{D}_r(X_{V \setminus r})\right)$$

$E_r(X_{V \setminus r})$: function of the values at sites neighboring site r
$B_r(X_r)$: sufficient statistic
$C_r(X_r)$: base measure
$\bar{D}_r(X_{V \setminus r})$: log-partition function

Mixed Graphical Models

Clique Factors of Size at Most Two and Two Types of Variables

The joint distribution:

$$\begin{aligned} P(X, Y; \theta) = & \exp\Big\{ \sum_{r \in V_X} \theta_r B_X(X_r) + \sum_{r' \in V_Y} \theta_{r'} B_Y(Y_{r'}) \\ + & \sum_{(r,t) \in E_X} \theta_{rt}\, B_X(X_r)\, B_X(X_t) + \sum_{(r',t') \in E_Y} \theta_{r't'}\, B_Y(Y_{r'})\, B_Y(Y_{t'}) \\ + & \sum_{(r,r') \in E_{XY}} \theta_{rr'}\, B_X(X_r)\, B_Y(Y_{r'}) + \sum_{r \in V_X} C_X(X_r) + \sum_{r' \in V_Y} C_Y(Y_{r'}) - A(\theta) \Big\} \end{aligned}$$

$$A(\theta) := \log \int_{\mathcal{X}^p} \exp\Big\{ \sum_{r \in V_X} \theta_r B_X(X_r) + \sum_{r' \in V_Y} \theta_{r'} B_Y(Y_{r'}) + ... + \sum_{r' \in V_Y} C_Y(Y_{r'}) \Big\}$$

$B_X(.)$, $C_X(.)$ sufficient statistic and base measure for the node-cond distrib of X
$B_Y(.)$, $C_Y(.)$ sufficient statistic and base measure for the node-cond distrib of Y
$\theta_r = (\theta_r, \theta_{rt})$ set of parameters
$A(\theta)$ log-partition function

Mixed MRFs

Advantage:

- General mixed multivariate distribution exists!

Caveat:

- Stringent Normalizability Assumptions.
  - $A(\theta) < \infty$.
  - No distribution exists linking Poisson and Gaussian variables.

Mixed MRFs

Advantage:

- General mixed multivariate distribution exists!

Caveat:

- Stringent Normalizability Assumptions.
  - $A(\theta) < \infty$.
  - No distribution exists linking Poisson and Gaussian variables.

Solution:

- Chain rule of conditional probability: $P(X, Y) = P(Y|X)P(X)$.

Hydra Graphs: Elementary Construction

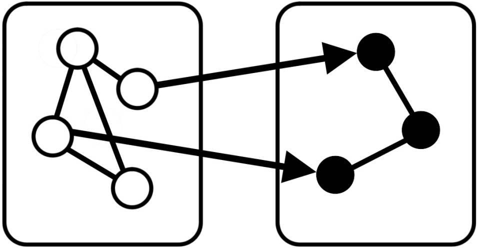
Partition p variables into two groups: $X = \{Y, Z\}$:

$$P(X) = P_1(Y|Z)P_2(Z)$$

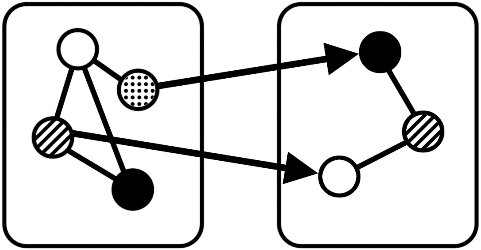
- P_1 is a **Conditional Markov Random Field** constructed via node-conditional exponential families.
  - Heterogeneous (Mixed).
  - Homogeneous.
- P_2 is a **Markov Random Field** constructed via node-conditional exponential families.
  - Heterogeneous (Mixed).
  - Homogeneous.

Hydra Graphs: Elementary Construction

Homogeneous Elementary Hydra Graphs:



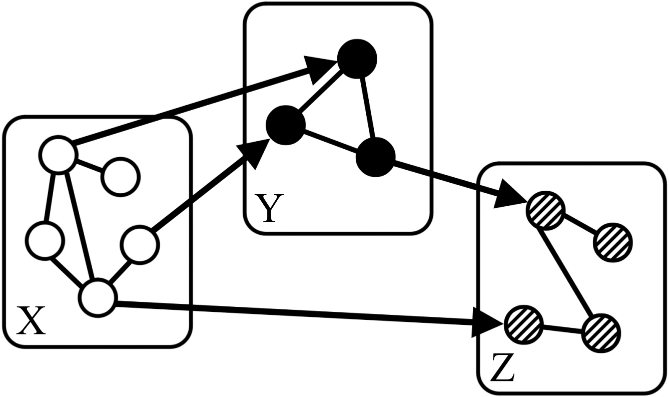
Heterogeneous Elementary Hydra Graphs:



Hydra Graphs: Recursively Chained

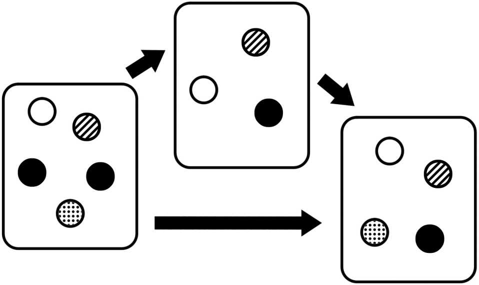
Idea: Recursively apply chain rule to partitions of variables.

$$P(X,Y,Z) = P(X|Y,Z)P(Y|Z)P(Z)$$



Directed edges: CRFs & **Undirected** edges: MRFs

Hydra Graphs: Recursively Chained



To yield a consistent joint density:

- Blocked Directed Acyclic Graph (DAG):
  - Within Block: Undirected edges.
  - Between Blocks: Directed edges (no cycles!).
- Each CRF / MRF component must be normalizable.
  - Much weaker conditions than Mixed MRFs.

Permits dependent Gaussian and Poisson distributions!

Graph Selection and Estimation

Objective: Given iid observations, seek to learn graph structure (selection) and parameters (estimation).

Node-Neighborhood Selection - For each node:

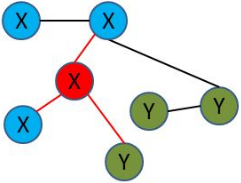
- Maximize penalized conditional likelihood = Mixed, penalized GLMs!

Theoretical Guarantees (under certain conditions):

- Unique solution.
- With high probability, exactly recover the true edge structure.
- Consistent parameter estimation.

ℓ_1 regularized M-estimator

$$-\,||X_r - X_{/r}\theta_{xx} - Y\theta_{xy}||_2^2 + \lambda_1||\theta_{xx}||_1 + \lambda_2||\theta_{xy}||_1,$$



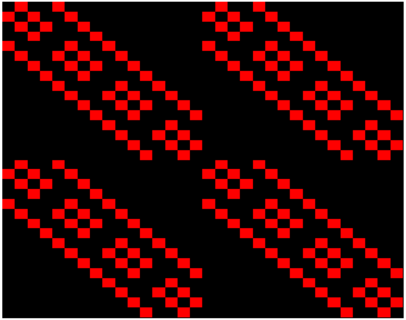
1. Introduction

2. Graphical Models via Exponential Families
   - Graphical Models via Exponential Families
   - Mixed Graphical Models

3. Results

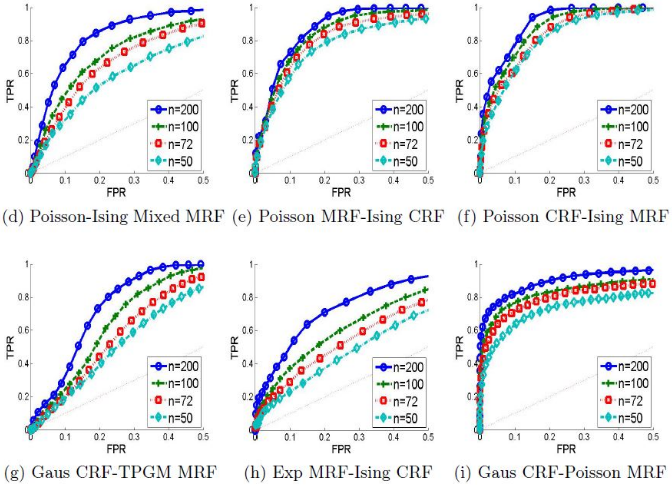
Simulation Study

- Samples generated via Gibbs sampling.
- Lattice structure



- $p = 72$: $p_Y = 36$, $p_Z = 36$
- Sample sizes: n=50, 72, 100 and 200.

Simulation Study

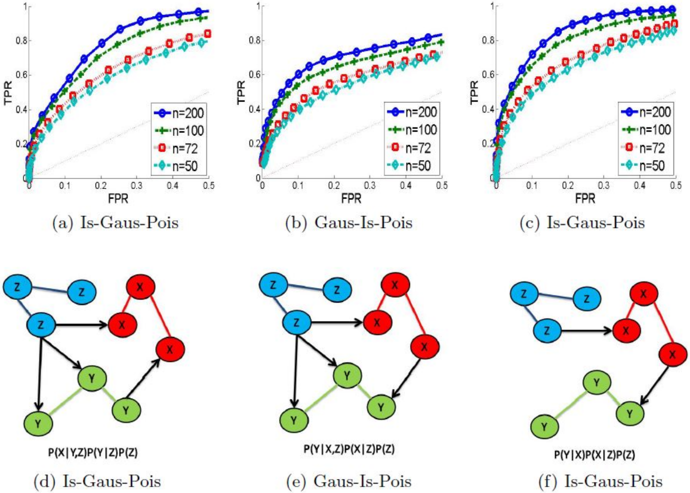


(d) Poisson-Ising Mixed MRF (e) Poisson MRF-Ising CRF (f) Poisson CRF-Ising MRF

(g) Gaus CRF-TPGM MRF (h) Exp MRF-Ising CRF (i) Gaus CRF-Poisson MRF

Figure: ROC curves for different types of models when $p_Y = 36$, $p_Z = 36$.

Simulation Study



(a) Is-Gaus-Pois (b) Gaus-Is-Pois (c) Is-Gaus-Pois

(d) Is-Gaus-Pois (e) Gaus-Is-Pois (f) Is-Gaus-Pois

Figure: ROC curves for 3 blocks of variables: binary (Ising, X), continuous (Gaussian, Y) and counts (Poisson, Z).

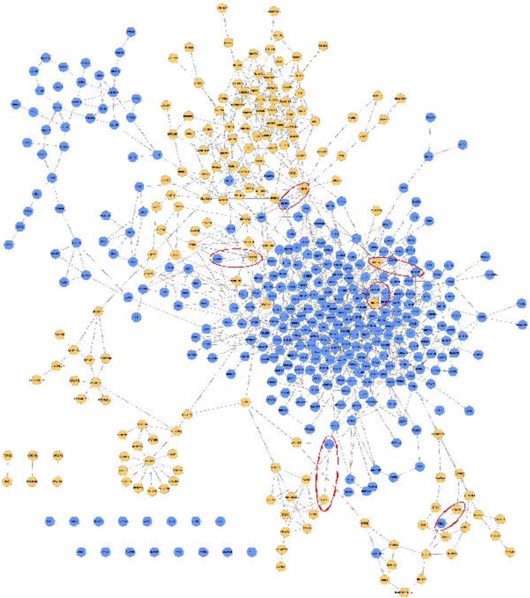
Case Study: Breast Cancer Genomics

Objective: Identify both between and within connections between mutation and expression biomarkers.

- Gene expression: TCGA Level III RNA-sequencing (counts).
- Mutations & Aberrations: Combination of TCGA Level II non-silent somatic mutation and TCGA Level III copy number variation (binary).
- 697 patients and 498 genes (329 expression biomarkers & 169 mutation biomarkers).
- Modeled via Poisson CRF- Ising MRF (mutations influence expression).
- Stability selection for model selection.

Case Study: Breast Cancer Genomics

Yellow nodes: RNA-sequencing; Blue nodes: genomic mutations



Case Study: Breast Cancer Genomics

Discovery of Previously Indicated Links:

- GATA3 mutation linked to SLC39A6 expression.
  - Ratio of gene expression levels used to defined breast cancer sub-types.
- FGFR1 mutation linked to PEG3 expression.
  - FGFR1 growth factors amplified in breast cancer work with PEG3 which modulates cancer progression.
- STAT3 mutation linked to ERBB2 expression.
  - Amplified in HERB2 sub-types and promotes cancer stem-cell proliferation.

Case Study: Breast Cancer Genomics

Novel Discoveries:

- TP53 mutation linked to ADAM6 expression.
  - TP53 a tumor suppressor gene & ADAM6 a long non-coding RNA over-expressed in breast cancer.
- FGF3 mutation linked to CCND1 expression.
  - FGF3 regulates estrogen expanding breast cancer stem cells & CCDN1 over-expression of hormone receptors in breast cancer.
- PIK3CA mutation linked to CLEC3A expression and NAT1 expression.
  - PIK3CA an oncogene, CLEC3A affects tumor metastasis, and NAT1 a potential marker for estrogen receptor positive sub-type.

Summary

Mixed Graphical Models

- Extends Markov Networks for (almost) any data type.
- First ever direct multivariate density for mixed data types!
- Hydra Graphs: Flexible models.
- Can be used to model connections both *within* and *between* multiple types of biomarkers.

`R` & `Bioconductor` Package & Matlab Toolbox `expMRF` coming soon.

Acknowledgments

Collaborators:

- Zhandong Liu, Baylor College of Medicine.
- Pradeep Ravikumar, University of Texas, Austin.
- Eunho Yang, University of Texas, Austin.
- Yulia Baker, PhD Candidate, Statistics, Rice University.
- Matthew Anderson, Baylor College of Medicine.
- Ying-Wooi Wan, Baylor College of Medicine.

Funding:

National Science Foundation, Division of Mathematical Sciences 1264058.

Major References

E. Yang, Y. Baker, P. Ravikumar, G. I. Allen, Y. W. Wan, and Z. Liu, ”A General Framework for Mixed Graphical Models”, (Coming Soon), 2014.

E. Yang, Y. Baker, P. Ravikumar, G. I. Allen, and Z. Liu, ”Mixed Graphical Models via Exponential Families”, In *Artificial Intelligence and Statistics* (AISTATS), 2014.

G. I. Allen and Z. Liu, “A Local Poisson Graphical Model for Inferring Networks from Next Generation Sequencing Data”, *IEEE Transactions on NanoBiosciences*, 2013.

G. I. Allen and Zhandong Liu, ”A Log-Linear Graphical Model for Inferring Genetic Networks from High-Throughput Sequencing Data”, In *IEEE International Conference on Bioinformatics and Biomedicine* (BIBM), 2012.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu, ”Graphical Models via Generalized Linear Models”, In *Neural Information Processing Systems* (NIPS), 2012.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu, ”On Graphical Models via Univariate Exponential Families”, (arXiv:1301.4183), 2013.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu, ”On Poisson Graphical Models”, In *Neural Information Processing Systems* (NIPS), 2013.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu, ”Conditional Random Fields via Univariate Exponential Families”, In *Neural Information Processing Systems* (NIPS), 2013.