Learning Markov Networks for Mixed Big Data:
Applications to Cancer Genomics

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.

March 27, 2015
Motivation: Genomic Networks from High-Throughput Data
Motivation: Big, Mixed Genomics Data

TCGA Genomics Data:
- SNPs / Copy Number Variation.
  - DNA sequence changes.
- Gene Expression.
  - Gene transcription.
- Methylation.
  - Epigenetic changes.
- MicroRNA expression.
  - Gene inhibitors.
Motivation: Big, Mixed Genomics Data

TCGA Genomics Data:
- SNPs & Copy Number Variation ($\approx 10,000,000$).
  - Binary or discrete.
- Gene Expression (via RNA Sequencing - $\approx 10,000$).
  - Counts.
- Methylation ($\approx 2,000,000$).
  - Bounded continuous.
- MicroRNA expression ($\approx 1,000$).
  - Continuous.
Review: Markov Networks

Graphical Models via Exponential Families
- Motivation
- Graphical Models via Exponential Families
- Results

Mixed Graphical Models
- Mixed MRFs
- Block-Directed MRFs
- Results
Markov Networks

- Represent dependencies in multivariate distributions.
- Undirected graphical models or pair-wise Markov Random Fields.
- Captures direct dependencies between random variables.
- $X = (X_1, X_2, \ldots X_p)$ is a random vector.
- Denote the graph $G$ by a set of vertices $V = 1, \ldots p$ and edges, $E \in V \times V$: $G = \{V, E\}$.
- Lack of edges denote conditional independence given all other nodes (pairwise conditional independence).

$$A \perp B \mid C$$
Markov Networks

- **Local Markov Property**: Conditional dependencies defined by node-neighborhoods, or the set of nodes connected to a given node via an edge.

- **Global Markov Property**: Pairwise conditional dependencies and neighborhoods jointly define the global dependence structure (formally defined by separators).

- **Hammersley-Clifford Theorem**: Density on graph factorizes according to sufficient statistics on cliques.
Review: Markov Networks

Graphical Models via Exponential Families
- Motivation
- Graphical Models via Exponential Families
- Results

Mixed Graphical Models
- Mixed MRFs
- Block-Directed MRFs
- 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. (Meinshausen & 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 Conditional Gaussian 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, ...

\[ P(Z) = \exp(\theta B(Z) + C(Z) - D(\theta)) \]

- \( \theta \) 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 $X = (X_1, X_2, \ldots X_p)$, suppose:

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

**Theorem**

Joint Density **necessarily** has the form:

$$P(X) = \exp \left\{ \sum_s \theta_s B(X_s) + \sum_{s \in V} \sum_{t \in N(s)} \theta_{st} B(X_s)B(X_t) ight. 
+ \sum_{s \in V} \sum_{t_2, \ldots, t_k \in N(s)} \theta_{s \ldots t_k} B(X_s) \prod_{j=2}^k B(X_{t_j}) + \sum_s C(X_s) - A(\theta) \right\}$$

$N(s)$ denotes the neighborhood of node $s$ & $A(\theta)$ is the log-normalization term. (**YRAL**, *NIPS*, 2012; **YRAL** *JMLR*, 2015)
Graphical Models via Exponential Families

Special Case:
- Pairwise 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 with response $X_s$ and predictors $X_{V \setminus s}$.
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.
  - \( \theta_{st} \leq 0 \) (only negative conditional dependencies!)

- Variations that permit negative & positive dependencies:
  - Truncation, Sub-linear, Quadratic, and approximations to these.

Results: TCGA Breast Cancer microRNA

MicroRNAs:
- Short RNAs ($\approx 22$ base pairs).
- Post-transcriptional regulators.
- Predominately inhibitors of gene expression.

Data:
- 544 breast cancer tumor samples, 524 microRNAs.
Results: TCGA Breast Cancer microRNA
Results: TCGA Breast Cancer microRNA

Selected Findings:

- Let-7c: Regulates breast cancer metastatic.
- MiR-10b: Over-expressed in triple negative (ER-, PR-, Her2-) breast cancer patients.
- MiR-375: Blocking slows down tumor progression in ER+ patients.
- MiR-10-b & MiR-375 estimated to be conditionally independent.

(AL, IEEE NB, 2013)
Comparisons: Poisson vs. Gaussian Graphical Models

Sub-Linear Poisson Graphical Model

Gaussian Graphical Model
1. Review: Markov Networks

2. Graphical Models via Exponential Families
   - Motivation
   - Graphical Models via Exponential Families
   - Results

3. Mixed Graphical Models
   - Mixed MRFs
   - Block-Directed MRFs
   - Results
Motivation: Big, Mixed Genomics Data

TCGA Genomics Data:
- SNPs & Copy Number Variation.
  - Binary or discrete.
- Gene Expression.
  - Counts.
- Methylation.
  - Bounded continuous.
- MicroRNA expression.
  - Continuous.

No general multivariate density that directly parameterizes dependencies for mixed variables exists!
Building Mixed MRFs:

- \( p \)-variate random vector: \( X = (X_1, X_2, \ldots, X_p) \) where \( X_s \in \mathcal{X}_s \) potentially all distinct types.

- Node-Conditional Distribution: \( P(X_s|X_{\setminus s}) \) a univariate exponential family.

\[
P(X_s|X_{\setminus s}) = \exp \left\{ g_s(X_{\setminus s}; \theta) B_s(X_s) + C_s(X_s) - D_s(X_{\setminus s}; \theta) \right\}
\]

- Canonical parameter, \( g_s(X_{\setminus s}; \theta) \), a function of other nodes (like GLMs).

- \( B_s(X_s) \) is the sufficient statistic.

- \( C_s(X_s) \) is the base measure.

- \( D_s(X_{\setminus s}; \theta) \) is the log-partition function.
Theorem

Joint Density necessarily has the form:

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

\(N(s)\) denotes the neighborhood of node \(s\) & \(A(\theta)\) is the log-normalization term.  
(YRAL, AISTATS, 2014)
Mixed Markov Random Fields

Special Case: Two partitions of variables \((X, Y)\) & pairwise graph.

\[
P(X, Y; \theta) = \exp \left\{ \sum_{s \in V_X} \theta_s B_X(X_s) + \sum_{s' \in V_Y} \theta_{s'} B_Y(Y_{s'}) + \sum_{(s,t) \in E_X} \theta_{st} B_X(X_s) B_X(X_t) + \sum_{(s',t') \in E_Y} \theta_{s't'} B_Y(Y_{s'}) B_Y(Y_{t'}) + \sum_{(s,s') \in E_{XY}} \theta_{ss'} B_X(X_s) B_Y(Y_{s'}) + \sum_{s \in V_X} C_X(X_s) + \sum_{s' \in V_Y} C_Y(Y_{s'}) \right\}
\]

\[
A(\theta) := \log \int_{\mathcal{X}^p} \exp \left\{ \sum_{s \in V_X} \theta_s B_X(X_s) + \sum_{s' \in V_Y} \theta_{s'} B_Y(Y_{s'}) + \sum_{s'' \in V_{XX}} C_{XX}(X_{s''}) \right\}
\]

\(B_X(.), C_X(.)\) sufficient stat. and base measure for node-conditional of \(X\)
\(B_Y(.), C_Y(.)\) sufficient stat. and base measure for node-conditional of \(Y\)
\(A(\theta)\) 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: $P(X, Y) = P(Y|X)P(X)$. 
Block-Directed MRFs: 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.

(YRABWL, 2014)
Block-Directed MRFs: Elementary Construction

Homogeneous Elementary Block-Directed MRF:

Heterogeneous Elementary Block-Directed MRF:
Elementary BDMRFs: Properties

Does this correspond to a Markov Network?

Theorem

If the covariate parameters (between block parameters) are functions solely of the (between block) node-neighborhoods, then the EBDMRF is a Markov Network.

Implication: Can interpret every edge via conditional dependence relationships.
Elementary BDMRFs: Properties

How does this compare to the Mixed MRF?

Example: \( P(Z) = P(Y|X)P(X) \) for pairwise graphs.

\[
P(X, Y; \theta) = \exp \left\{ \sum_{s \in V_X} \theta_s B_X(X_s) + \sum_{s' \in V_Y} \theta_{s'} B_Y(Y_{s'}) + \sum_{(s, t) \in E_X} \theta_{st} B_X(X_s) B_X(X_t) + \sum_{(s', t') \in E_Y} \theta_{s't'} B_Y(Y_{s'}) B_Y(Y_{t'}) + \sum_{(s, s') \in E_{XY}} \theta_{ss'} B_X(X_s) B_Y(Y_{s'}) + \sum_{s \in V_X} C_X(X_s) + \sum_{s' \in V_Y} C_Y(Y_{s'}) - A_{Y|X}(\theta) - A_X(\theta) \right\}
\]

\( A_{Y|X}(\theta) \) normalizability term for CRF.
\( A_X(\theta) \) normalizability term for MRF.
Elementary BDMRFs: Properties

What does this mean in terms of normalizability?

**Theorems**

1. EBDMRF normalizable if CRF and MRF are normalizable.
2. Normalizability conditions strictly weaker than mixed MRFs.

Implication: Dependencies permitted in EBDMRFs that aren’t in mixed MRFs.
Elementary BDMRFs: Properties

**Gaussian-Poisson EBDMRF**

Let $X$ be counts (Poisson) and $Y$ continuous (Gaussian).

- Mixed MRF does not permit dependencies between $X$ and $Y$.
- EBDMRF $P(X, Y) = P(Y|X)P(X)$ exists!
- EBDMRF $P(X, Y) = P(X|Y)P(Y)$ exists!

$$
\log P(X, Y) = \sum_{s \in V_Y} \left( \theta_s + \sum_{t \text{ s.t } (s,t) \in E_{XY}} \theta_{st} X_t \right) \frac{Y_s}{\sigma_s} + \sum_{(s,s') \in E_Y} \theta_{ss'} \frac{Y_s Y_{s'}}{\sigma_s \sigma_{s'}} - \sum_{s \in V_Y} \frac{Y_s^2}{2\sigma_s^2} \\
+ \sum_{X \in V_X} \theta_{X} X_t + \sum_{(t,t') \in E_X} \theta_{tt'} X_t X_{t'} - \sum_{t \in V_X} \log(X_t!) - A_{Y|X}(\theta(X)) - A_X(\theta).
$$
Block-Directed MRFs: Recursively Chained

Idea: Recursively apply chain rule to partitions of variables.

\[ P(X, Y, Z) = P(X|Y, Z)P(Y|Z)P(Z) \]
Block-Directed MRFs: Recursively Chained

When will this yield a Markov Network?

Partial ordering of blocks must be known:

- Blocked Directed Acyclic Graph (DAG):
  - Within Block: Undirected edges.
  - Between Blocks: Directed edges (no cycles!).

G. I. Allen (Rice & BCM)
Mixed Markov Networks
March 27, 2015 19 / 26
Block-Directed MRFs: Recursively Chained

### Theoretical Properties

- **Global Markov Property.**
  - Non-linear, arbitrary covariate functions.
  - Must be functions solely of node-neighbors across blocks.

- Normalizable if each CRF & MRF are normalizable
  - Significantly weaker conditions than Mixed MRFs.

### Implications:

- Most general, flexible class of MRF models.
- **First ever multivariate model that directly parameterizes dependencies over heterogeneous variables!**
Graph Selection and Estimation

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

- Density factors into products of CRFs $\Rightarrow$ Can learn each CRF separately.
- Each CRF factors by node-neighborhood $\Rightarrow$ Can learn each node neighborhood separately.
- Neighborhood Selection - For each node:
  - Maximize penalized conditional likelihood = Mixed, penalized GLMs!

$\ell_1$ regularized MLE (Gaussian node-conditional):

$$\min_{\theta_{xy}, \theta_{xx}} \|X_s - X/s\theta_{xx} - Y\theta_{xy}\|^2_2 + \lambda_{xx}\|\theta_{xx}\|_1 + \lambda_{xy}\|\theta_{xy}\|_1,$$
Theoretical Guarantees

Assumptions:

- Restricted Eigenvalue condition (on Fisher Information matrix).
- $\lambda \propto \sqrt{\log(p_x + p_y)/n }$ & $n \propto d^2\log(p)$ (degree squared).

Result:

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

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

(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
Simulation Study

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.

(YRABWL, 2014)
Case Study: Breast Cancer Genomics

Blue nodes: RNA-sequencing; Yellow 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.
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 & CCND1 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

### Graphical Models via Exponential Families
- **Stat:** Extends Markov Networks for (almost) any data type!
- **Application:** Poisson Graphical Models for RNA-Sequencing data.

### Mixed Graphical Models
- **Stat:** First ever direct multivariate density for mixed data types!
- **Stat:** Permits both directed AND undirected edges - flexible class of models.
- **Application:** Can model connections within and between multiple types of biomarkers.
Acknowledgments

Collaborators:

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

R & Bioconductor Package XMRF and TCGA2STAT coming soon.
Major References


