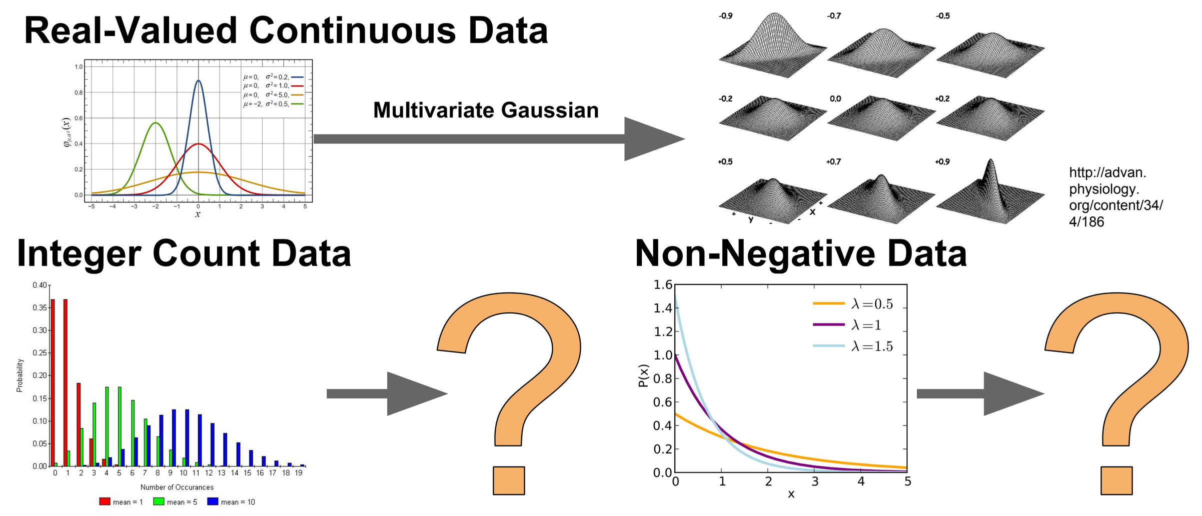
THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Computer Science
College of Natural Sciences

Square Root Graphical Models: Multivariate Generalizations of Univariate Exponential Families that Permit Positive Dependencies

David I. Inouye, Pradeep Ravikumar, Inderjit S. Dhillon

Center for Big Data Analytics
http://bigdata.ices.utexas.edu

Motivation 1: Non-Gaussian Data



- **Problem:** What is the appropriate graphical model for count data and non-negative skewed data?
- **Proposal:** **General** graphical models for *any* univariate exponential family with positive sufficient statistics

Motivation 2: Graphical Models from (Besag 1974, Yang et al. 2012)

- Generalizes *any* univariate exponential family (such as Poisson or exponential) with sufficient statistic $\mathrm{T}(x)$, base measure $\mathrm{B}(x)$ and domain $\mathcal{D}$

$$\mathrm{T}(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}^p = \text{Entry-wise application of } \mathrm{T}(x)$$

$$\Pr(\mathbf{x}|\boldsymbol{\theta},\Phi) \tag{1}$$

$$= \exp\left(\boldsymbol{\theta}^T\mathrm{T}(\mathbf{x}) + \mathrm{T}(\mathbf{x})^T\Phi\mathrm{T}(\mathbf{x}) + \sum_{s=1}^{p}\mathrm{B}(x_s) - \mathrm{A}(\boldsymbol{\theta},\Phi)\right)$$

$$\mathrm{A}(\boldsymbol{\theta},\Phi) \tag{2}$$

$$= \log\int_{\mathcal{D}}\exp\left(\boldsymbol{\theta}^T\mathrm{T}(\mathbf{x}) + \mathrm{T}(\mathbf{x})^T\Phi\mathrm{T}(\mathbf{x}) + \sum_{s=1}^{p}\mathrm{B}(x_s)\right)\mathrm{d}\mu(\mathbf{x})$$

$$\Phi = \Phi^T \in \mathbb{R}^{p\times p}, \phi_{ss} = 0 \ \forall s \text{ (no self-edges)}$$
$$\mu = \text{either Lebesgue or counting measure}$$

- Discrete and Gaussian graphical models are special cases
- **Problem: Only negative dependencies** for the Poisson and exponential models (informally because interaction is $O(\mathrm{T}(x)^2)$)
- **Proposal:** Graphical models that permit positive *and* negative dependencies with a simple but elegant modification

Square Root Graphical Model (SQR)

$$\Pr(\mathbf{x}\,|\,\boldsymbol{\theta},\Phi) \tag{3}$$

$$= \exp\left(\boldsymbol{\theta}^T\sqrt{\mathrm{T}(\mathbf{x})} + \sqrt{\mathrm{T}(\mathbf{x})}^T\Phi\sqrt{\mathrm{T}(\mathbf{x})} + \sum_s\mathrm{B}(x_s) - \mathrm{A}(\boldsymbol{\theta},\Phi)\right)$$

$$\mathrm{A}(\boldsymbol{\theta},\Phi) \tag{4}$$

$$= \log\int_{\mathcal{D}}\exp\left(\boldsymbol{\theta}^T\sqrt{\mathrm{T}(\mathbf{x})} + \sqrt{\mathrm{T}(\mathbf{x})}^T\Phi\sqrt{\mathrm{T}(\mathbf{x})} + \sum_s\mathrm{B}(x_s)\right)\mathrm{d}\mu(\mathbf{x})$$

$$\Phi = \Phi^T \in \mathbb{R}^{p\times p}, \phi_{ss} \in \mathbb{R} \ \forall s \text{ (self-edges allowed)}$$

$$\left(\sqrt{\mathrm{T}(\mathbf{x})}\right)_s \equiv \begin{cases} x_s & \text{if } \mathrm{T}(x_s) = x_s^2 \\ \sqrt{\mathrm{T}(x_s)} & \text{otherwise} \end{cases} \forall s$$

- Allows both positive and negative dependencies.
- Main intuition is that interaction term is $O(\mathrm{T}(x))$ so that $\mathrm{A}(\boldsymbol{\theta},\Phi) < \infty$ even with *positive* dependencies.

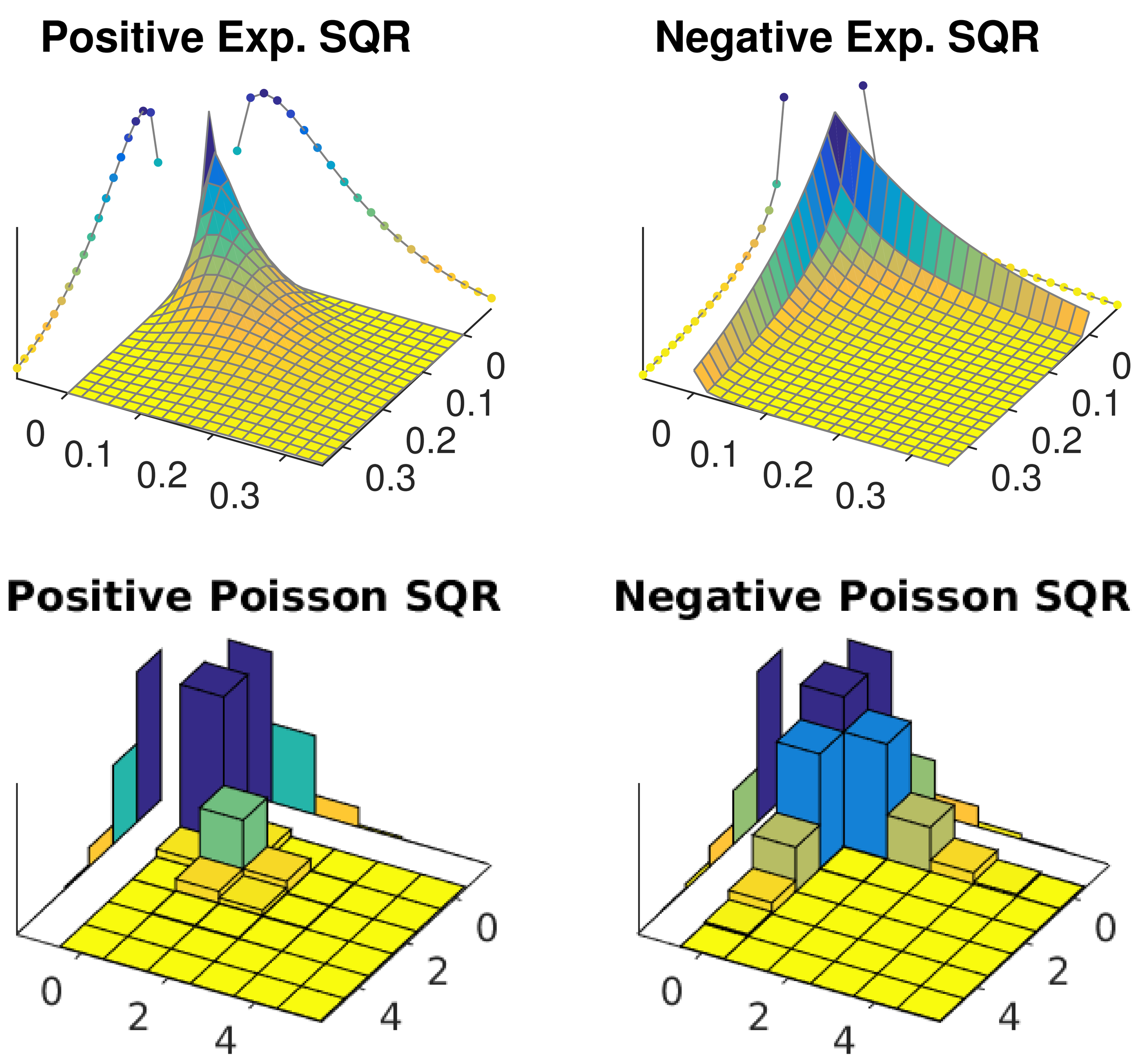
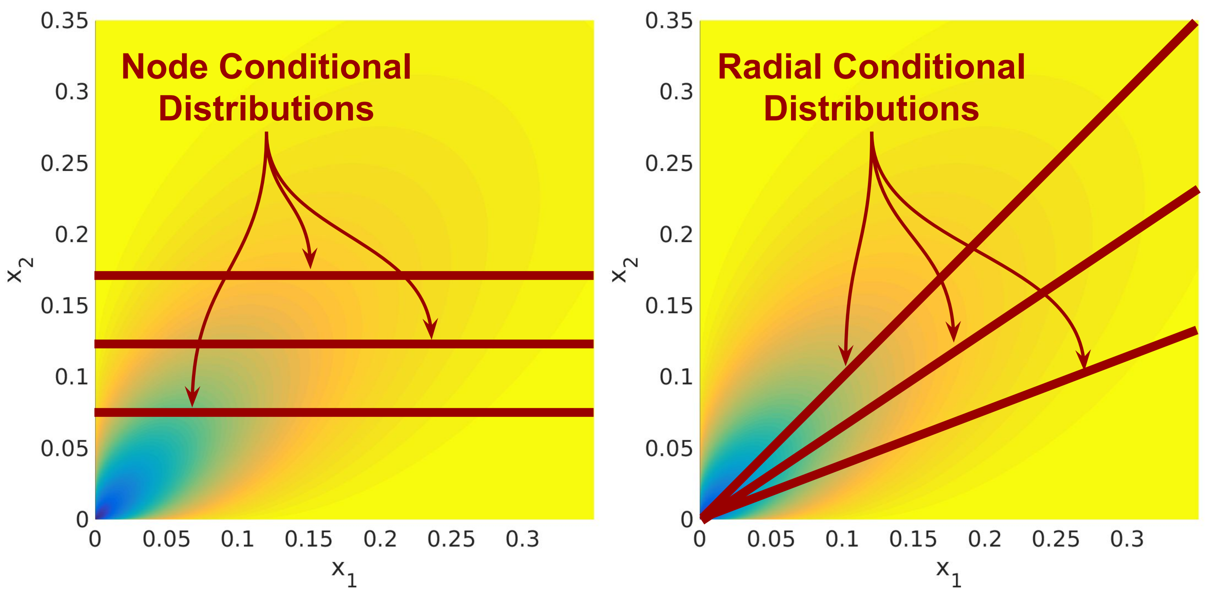


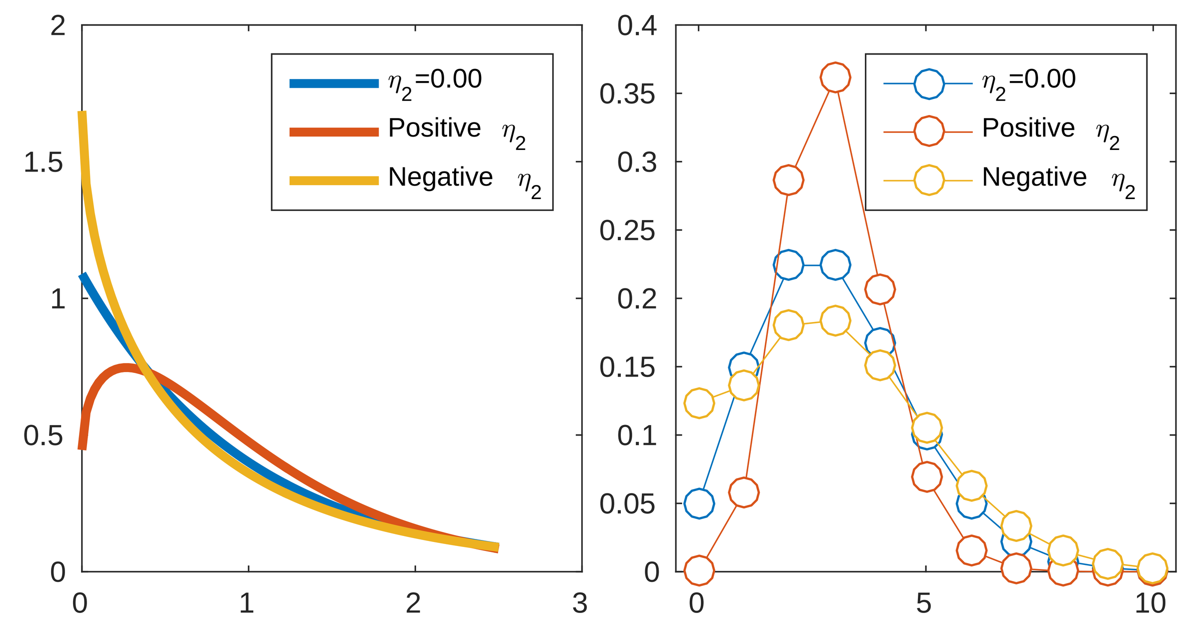
Figure : The SQR model class that can intuitively model *positive and negative* dependencies while having a simple parametric form.

SQR Conditional Distributions



- Node conditional: Distribution of one variable given the others
- Radial conditional: Distribution of scaling given the direction

SQR Node Condtionals



- *Two*-parameter exponential family:

$$\Pr(x_s\,|\,\mathbf{x}_{-s},\Phi) = \exp\Big(\underbrace{\phi_{ss}}_{\eta_1}\underbrace{\mathrm{T}(x_s)}_{\tilde{\mathrm{T}}_1(x_s)} + \underbrace{\left(\theta_s + 2\phi_{-s}^T\sqrt{\mathrm{T}(\mathbf{x}_{-s})}\right)}_{\eta_2}\underbrace{\sqrt{\mathrm{T}(x_s)}}_{\tilde{\mathrm{T}}_2(x_s)} + \mathrm{B}(x_s) - \underbrace{\mathrm{A}(\mathbf{x}_{-s},\Phi)}_{\mathrm{A}(\eta_1,\eta_2)}\Big)$$

Parameter Estimation using Node Conditionals

- As in [Ravikumar et al. 2010, Yang et al. 2015], we fit p *independent* ℓ_1-regularized node-wise regressions:

$$\arg\min_{\Phi} -\frac{1}{n}\sum_s\sum_i\Big(\eta_{1si}x_{si} + \eta_{2si}\sqrt{x_{si}} + \mathrm{B}(x_{si}) - \mathrm{A}_{\text{node}}(\eta_{1si},\eta_{2si})\Big) + \lambda\|\Phi\|_{1,\text{off}}, \tag{5}$$

$$\eta_{1si} = \phi_{ss}, \ \ \eta_{2si} = \theta + 2\phi_{-s}^T\sqrt{\mathrm{T}(\mathbf{x}_{-si})}, \ \ \|\Phi\|_{1,\text{off}} = \sum_{s\neq t}|\phi_{st}|$$

- **Main difficulty:** $\mathrm{A}(\eta_1,\eta_2)$ *usually* not known in closed-form
- However, $\mathrm{A}(\eta_1,\eta_2)$ for the exponential SQR is simply:

$$\mathrm{A}_{\text{Exp}}(\boldsymbol{\eta}) = \log\left(\frac{\sqrt{\pi}\eta_1\exp\left(\frac{-\eta_2^2}{4\eta_1}\right)\left(1-\mathrm{erf}\left(\frac{-\eta_2}{2\sqrt{-\eta_1}}\right)\right)}{-2(-\eta_1)^{\frac{3}{2}}} - \frac{1}{\eta_1}\right),$$

- Proximal gradient descent (though new arXiv paper uses Newton-like method)

SQR Radial Conditionals

- For simplicity, let $\mathrm{T}(\mathbf{x}) = \mathbf{x}$, $\mathbf{v} = \frac{\mathbf{x}}{\|\mathbf{x}\|_1}$, $z = \|\mathbf{x}\|_1$
- Assume direction $\mathbf{v}$ is given, but scaling z is unknown:

$$\Pr(\mathbf{x} = z\mathbf{v}\,|\,\mathbf{v},\boldsymbol{\theta},\Phi) = \exp\Big(\underbrace{\left(\sqrt{\mathbf{v}}^T\Phi\sqrt{\mathbf{v}}\right)}_{\tilde{\eta}_1}\underbrace{z}_{\tilde{\mathrm{T}}_1(z)} + \underbrace{\left(\boldsymbol{\theta}^T\mathbf{v}\right)}_{\tilde{\eta}_2}\underbrace{\sqrt{z}}_{\tilde{\mathrm{T}}_2(z)} + \underbrace{\textstyle\sum_{s=1}^{p}\mathrm{B}(zv_s)}_{\tilde{\mathrm{B}}_{\mathbf{v}}(z)} - \underbrace{\mathrm{A}_{\text{rad}}(\mathbf{v},\boldsymbol{\theta},\Phi)}_{\mathrm{A}(\tilde{\eta}_1,\tilde{\eta}_2)}\Big)$$

- Similar to node conditional form but with different parameters

Joint Normalization using Radial Conditionals

$$\mathrm{A}(\boldsymbol{\theta},\Phi) = \log\int_{\mathcal{V}}\underbrace{\int_{\mathcal{Z}(\mathbf{v})}\exp(\underbrace{\tilde{\eta}_1(\mathbf{v})z + \tilde{\eta}_2(\mathbf{v})\sqrt{z}}_{O(z)} + \underbrace{\tilde{B}_{\mathbf{v}}(z)}_{O(\mathrm{B}(z))})\mathrm{d}\mu(z)}_{\mathrm{A}_{\text{rad}}(\tilde{\eta}_1,\tilde{\eta}_2)}\,\mathrm{d}\mathbf{v},$$

$$\mathcal{V} = \{\mathbf{v} : \|\mathbf{v}\|_1 = 1, \mathbf{v} \in \mathbb{R}_+^p\}$$
$$\mathcal{Z}(\mathbf{v}) = \{z \in \mathbb{R}_+ : z\mathbf{v} \in \mathcal{D}\}.$$

- Because $\mathcal{V}$ is bounded, we merely need that the radial conditional distribution is normalizable (i.e. $\mathrm{A}_{\text{rad}}(\tilde{\eta}_1,\tilde{\eta}_2) < \infty$), which depends on the base univariate exponential family

Various Normalization Conditions

- **Gaussian SQR with $\mathrm{T}(x) = x^2$:** Standard positive-definite condition on covariance matrix
- **Poisson SQR:** **No restrictions** on parameter values (i.e. *any* positive or negative values in Φ permitted)
  - Informally, this is because the base measure is $O(-x\log x)$ while the other terms are $O(x)$ and thus any parameters are permitted
- **Exponential SQR:** Permits both **positive and negative dependencies** with the following condition akin to a negative-definite condition ($\mathcal{V}$ contains all *positive* vectors):

$$\Phi_{\text{Exp}} \in \{\Phi : \sqrt{\mathbf{v}}^T\Phi\sqrt{\mathbf{v}} < 0, \forall\mathbf{v} \in \mathcal{V}\}.$$

SQR Quantitative Results

- **Synthetic:** Constructed chain-like graphs where the edge strength was $0.9/k$ and $p = 30$. Higher k is much more difficult for two reasons:
  1. More edges (i.e. kp edges)
  2. Edges are weaker (i.e. $0.9/k$ per edge)

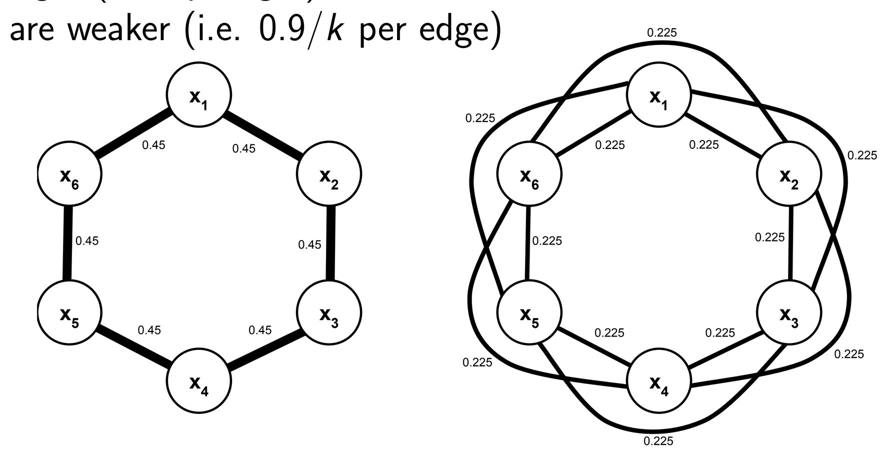


Figure : (Left) Synthetic chain graph with $k = 1$. (Right) Synthetic chain-like graph with $k = 2$ so that each node is connected to its first *and* second neighbors.

- **Airport delay dataset:** Average delay times per day for the largest 30 airports in 2014 ($p = 30, n = 365$)
  - Joint log partition function estimated using AIS sampling (Gibbs intermediate sampler with slice sampling for node conditionals)
  - Relative Likelihood Metric: $\exp((\mathcal{L}_{\text{SQR}} - \mathcal{L}_{\text{Ind}})/n)$, where $\mathcal{L}$ is the log likelihood.

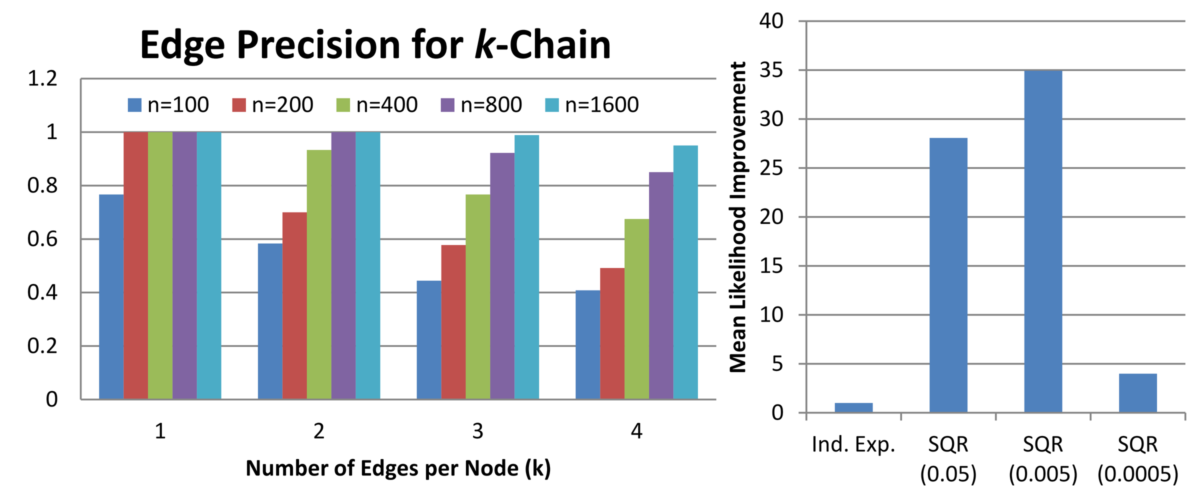


Figure : (Left) The top-kp edge precision for a synthetic circular chain-like graph. (Right) Relative log likehood results on the airport dataset of the exponential SQR compared to an independent exponential model.

Visualization of Top 50 SQR Edges

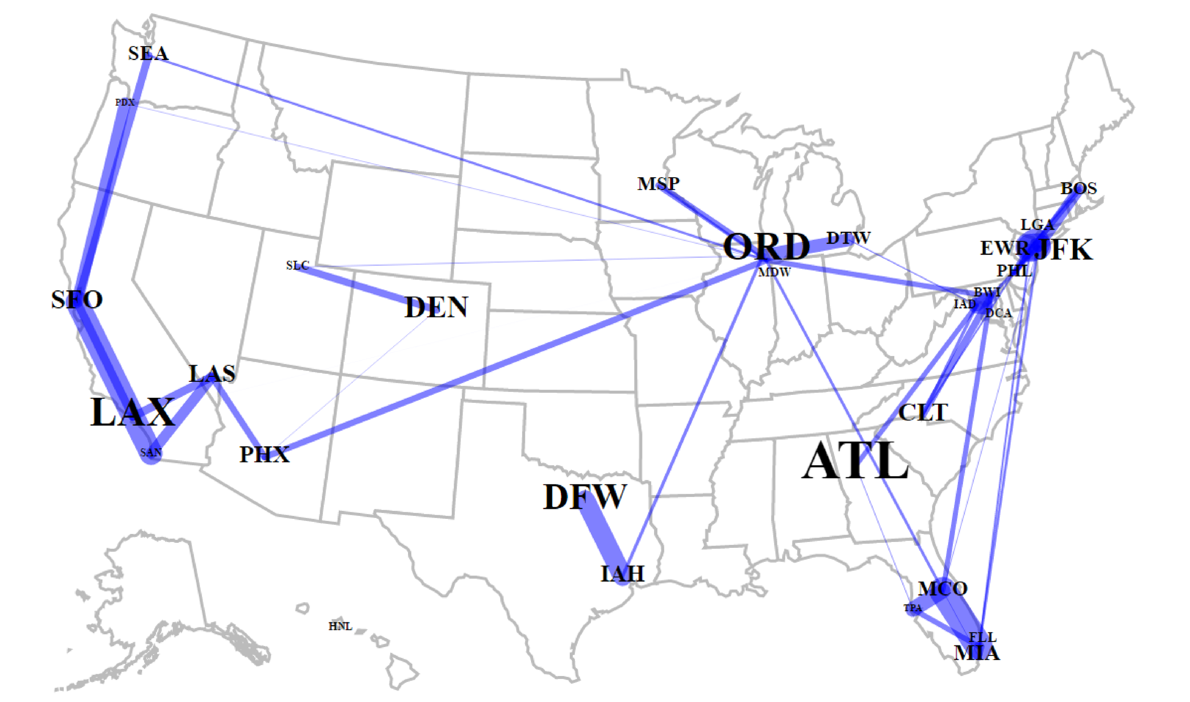


Figure : Visualizing the top 50 edges between airports shows the expected relationships likely due to weather or physical locality even though the algorithm was not provided any weather or location data.

Current Work

- **arXiv paper "Generalized Root Models"**
  - k-wise dependencies using k-th root sufficient statistics
  - Approximation algorithm for Poisson SQR
  - Fast Newton-like algorithm
- **Comparison to other multivariate models**
  - Poisson SQR allows strong *positive* and *negative* dependencies