Regularized score matching for graphical models: Non-Gaussianity and missing data

Mathias Drton (with Lina Lin, Ali Shojaie)

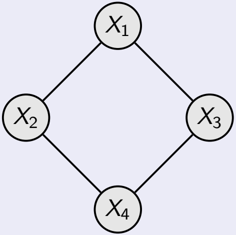
Department of Statistics
University of Washington

1. Conditional independence graphs (CIGs)

- $X = (X_1, \ldots, X_p)$: random vector with values in $\mathbb{R}^p$
- CIG of X : undirected graph G with $V(G) = \{X_1, \ldots, X_p\}$ and

  no edge between nodes X_j and X_k $\iff$ $X_j \perp\!\!\!\perp X_k \mid X_{\setminus\{j,k\}}$.

Example



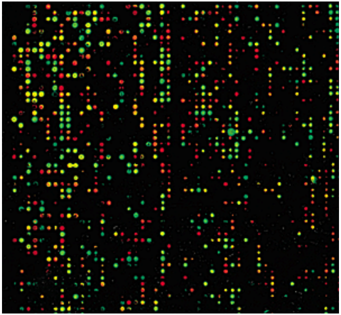
is CIG of X if

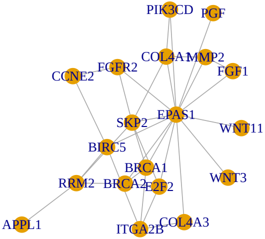
$$X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3,$$
$$X_2 \perp\!\!\!\perp X_3 \mid X_1, X_4$$

and no other full conditional independencies.

Motivation

Exploration of expression data to infer gene-gene interactions





number of genes $p > n$ number of samples

Gaussian graphical model

- Consider $X \sim \mathrm{N}_p(\mu, \mathbf{K}^{-1})$ with log-density:

$$\log f(x|\mu, \mathbf{K}) = -\frac{n}{2} \log \det(\mathbf{K}) - \frac{1}{2}(x-\mu)^T \mathbf{K}(x-\mu) + \text{const}$$

- CIG $\equiv$ sparsity pattern in precision matrix $\mathbf{K} = (\kappa_{jk})$:

$$X_j \perp\!\!\!\perp X_k \mid X_{\setminus\{j,k\}} \iff \kappa_{jk} = 0.$$

Many methods for high-dim. data: loss + regularizing penalty

Neighbourhood selection (Meinshausen and Bühlmann, 2006)

Graphical lasso/*glasso* (Yuan and Lin, 2007; Friedman et al., 2008)

...

Gaussian graphical model

- Consider $X \sim \mathrm{N}_p(\mu, \mathbf{K}^{-1})$ with log-density:

$$\log f(x|\mu, \mathbf{K}) = -\frac{n}{2} \log \det(\mathbf{K}) - \frac{1}{2}(x-\mu)^T \mathbf{K}(x-\mu) + \text{const}$$

- CIG $\equiv$ sparsity pattern in precision matrix $\mathbf{K} = (\kappa_{jk})$:

$$X_j \perp\!\!\!\perp X_k \mid X_{\setminus\{j,k\}} \iff \kappa_{jk} = 0.$$

Many methods for high-dim. data: loss + regularizing penalty

Neighbourhood selection (Meinshausen and Bühlmann, 2006)

Graphical lasso/*glasso* (Yuan and Lin, 2007; Friedman et al., 2008)

...

Non-Gaussian models: Pairwise interactions

- Log-densities of the form:

$$\log f(x|\theta) = \sum_{1 \leqslant j,k \leqslant p} \theta_{jk} t_{jk}(x_j, x_k) - \psi(\theta)$$

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{21} & \dots & \theta_{pp} \end{bmatrix} \qquad \theta_{jk} = \theta_{kj}, \quad j \neq k$$

$\psi(\theta)$: log-partition function.

- CIG $\equiv$ support of θ (Hammersley-Clifford):

$$X_j \perp\!\!\!\perp X_k \mid X_{\setminus\{j,k\}} \iff \theta_{jk} = 0.$$

- Gaussian special case (WLOG, $\mu = 0$):

$$\theta_{jk} = \kappa_{jk}, \quad t_{jk}(x_j, x_k) = x_j x_k, \quad \psi(\mathbf{K}) = -\frac{n}{2} \log\det(\mathbf{K}) + \text{const}$$

Non-Gaussian models: Pairwise interactions

- Log-densities of the form:

$$\log f(x|\theta) = \sum_{1 \leqslant j,k \leqslant p} \theta_{jk} t_{jk}(x_j, x_k) - \psi(\theta)$$

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{21} & \dots & \theta_{pp} \end{bmatrix} \qquad \theta_{jk} = \theta_{kj}, \quad j \neq k$$

$\psi(\theta)$: log-partition function.

- CIG $\equiv$ support of θ (Hammersley-Clifford):

$$X_j \perp\!\!\!\perp X_k \mid X_{\setminus\{j,k\}} \iff \theta_{jk} = 0.$$

- Gaussian special case (WLOG, $\mu = 0$):

$$\theta_{jk} = \kappa_{jk}, \quad t_{jk}(x_j, x_k) = x_j x_k, \quad \psi(\mathbf{K}) = -\frac{n}{2} \log \det(\mathbf{K}) + \text{const}$$

Non-Gaussian models: Pairwise interactions

- Log-densities of the form:

$$\log f(x|\theta) = \sum_{1 \leqslant j,k \leqslant p} \theta_{jk} t_{jk}(x_j, x_k) - \psi(\theta)$$

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{21} & \dots & \theta_{pp} \end{bmatrix} \qquad \theta_{jk} = \theta_{kj}, \quad j \neq k$$

$\psi(\theta)$: log-partition function.

- CIG $\equiv$ support of θ (Hammersley-Clifford):

$$X_j \perp\!\!\!\perp X_k \mid X_{\setminus\{j,k\}} \iff \theta_{jk} = 0.$$

- Gaussian special case (WLOG, $\mu = 0$):

$$\theta_{jk} = \kappa_{jk}, \quad t_{jk}(x_j, x_k) = x_j x_k, \quad \psi(\mathbf{K}) = -\frac{n}{2} \log\det(\mathbf{K}) + \text{const}$$

Different types of interactions: Example

- Model with densities:

$$f(x|A,B,C) \propto \underbrace{\exp\left\{-\frac{1}{2}\left[\sum_{j\leqslant k} A_{jk}x_j^2x_k^2 + \sum_{j\leqslant k} B_{jk}x_jx_k + \sum_j C_jx_j\right]\right\}}_{q(x|A,B,C)}$$

- Normal conditional distributions (Arnold et al., 2001)
- Dependence also through variance
- Intractable log-partition function

$$\psi(\theta) = \psi(A,B,C) = \log\int_{\mathbb{R}}\dots\int_{\mathbb{R}} q(x|A,B,C)dx_1\dots dx_p$$

Different types of interactions: Example

- Model with densities:

$$f(x|A,B,C) \propto \underbrace{\exp\left\{-\frac{1}{2}\left[\sum_{j\leqslant k} A_{jk}x_j^2x_k^2 + \sum_{j\leqslant k} B_{jk}x_jx_k + \sum_j C_jx_j\right]\right\}}_{q(x|A,B,C)}$$

- Normal conditional distributions (Arnold et al., 2001)
- Dependence also through variance
- Intractable log-partition function

$$\psi(\theta) = \psi(A,B,C) = \log \int_{\mathbb{R}} \dots \int_{\mathbb{R}} q(x|A,B,C)dx_1 \dots dx_p$$

Different types of interactions: Example

- Model with densities:

$$f(x|A,B,C) \propto \underbrace{\exp\left\{-\frac{1}{2}\left[\sum_{j\leqslant k} A_{jk}x_j^2x_k^2 + \sum_{j\leqslant k} B_{jk}x_jx_k + \sum_j C_jx_j\right]\right\}}_{q(x|A,B,C)}$$

- Normal conditional distributions (Arnold et al., 2001)
- Dependence also through variance
- Intractable log-partition function

$$\psi(\theta) = \psi(A,B,C) = \log\int_{\mathbb{R}}\dots\int_{\mathbb{R}} q(x|A,B,C)dx_1\dots dx_p$$

Different types of interactions: Example

- Model with densities:

$$f(x|A,B,C) \propto \underbrace{\exp\left\{-\frac{1}{2}\left[\sum_{j\leqslant k} A_{jk}x_j^2x_k^2 + \sum_{j\leqslant k} B_{jk}x_jx_k + \sum_j C_jx_j\right]\right\}}_{q(x|A,B,C)}$$

- Normal conditional distributions (Arnold et al., 2001)
- Dependence also through variance
- Intractable log-partition function

$$\psi(\theta) = \psi(A,B,C) = \log \int_{\mathbb{R}} \dots \int_{\mathbb{R}} q(x|A,B,C)dx_1 \dots dx_p$$

Approaches to inference

- Maximum likelihood

  Need to know partition function.

- Pseudo-likelihood

  Product of conditional likelihood functions

  e.g., neighbourhood selection
  (Meinshausen and Bühlmann, Ravikumar et al.)

  May need approximations of univariate log-partition functions.

  Need not be regression problem of standard GLM-type.

- Simpler option: Score matching

  Not new but was/is underused?

Approaches to inference

- Maximum likelihood

  Need to know partition function.

- Pseudo-likelihood

  Product of conditional likelihood functions

  e.g., neighbourhood selection
  (Meinshausen and Bühlmann, Ravikumar et al.)

  May need approximations of univariate log-partition functions.

  Need not be regression problem of standard GLM-type.

- Simpler option: Score matching

  Not new but was/is underused?

Approaches to inference

- Maximum likelihood

  Need to know partition function.

- Pseudo-likelihood

  Product of conditional likelihood functions

  e.g., neighbourhood selection
  (Meinshausen and Bühlmann, Ravikumar et al.)

  May need approximations of univariate log-partition functions.

  Need not be regression problem of standard GLM-type.

- Simpler option: Score matching

  Not new but was/is underused?

Approaches to inference

- Maximum likelihood

  Need to know partition function.

- Pseudo-likelihood

  Product of conditional likelihood functions

  e.g., neighbourhood selection
  (Meinshausen and Bühlmann, Ravikumar et al.)

  May need approximations of univariate log-partition functions.

  Need not be regression problem of standard GLM-type.

- Simpler option: Score matching

  Not new but was/is underused?

Approaches to inference

- Maximum likelihood

  Need to know partition function.

- Pseudo-likelihood

  Product of conditional likelihood functions

  e.g., neighbourhood selection
  (Meinshausen and Bühlmann, Ravikumar et al.)

  May need approximations of univariate log-partition functions.

  Need not be regression problem of standard GLM-type.

- Simpler option: Score matching

  Not new but was/is underused?

Approaches to inference

- Maximum likelihood

  Need to know partition function.

- Pseudo-likelihood

  Product of conditional likelihood functions

  e.g., neighbourhood selection
  (Meinshausen and Bühlmann, Ravikumar et al.)

  May need approximations of univariate log-partition functions.

  Need not be regression problem of standard GLM-type.

- Simpler option: Score matching

  Not new but was/is underused?

Approaches to inference

- Maximum likelihood

  Need to know partition function.

- Pseudo-likelihood

  Product of conditional likelihood functions

  e.g., neighbourhood selection
  (Meinshausen and Bühlmann, Ravikumar et al.)

  May need approximations of univariate log-partition functions.

  Need not be regression problem of standard GLM-type.

- Simpler option: Score matching

  Not new but was/is underused?

Approaches to inference

- Maximum likelihood

  Need to know partition function.

- Pseudo-likelihood

  Product of conditional likelihood functions

  e.g., neighbourhood selection
  (Meinshausen and Bühlmann, Ravikumar et al.)

  May need approximations of univariate log-partition functions.

  Need not be regression problem of standard GLM-type.

- Simpler option: Score matching

  Not new but was/is underused?

Approaches to inference

- Maximum likelihood

  Need to know partition function.

- Pseudo-likelihood

  Product of conditional likelihood functions

  e.g., neighbourhood selection
  (Meinshausen and Bühlmann, Ravikumar et al.)

  May need approximations of univariate log-partition functions.

  Need not be regression problem of standard GLM-type.

- Simpler option: Score matching

  Not new but was/is underused?

2. Score matching (Hyvärinen, 2005)

- X : continuous observation with support $\mathcal{X} \subset \mathbb{R}^p$
- Density $f(x|\theta^*)$ from a parametric model $f(x|\theta)$, $\theta \in \Theta$.
- Idea: Avoid log-partition function by considering divergence

$$\mathcal{L}(\theta) = \frac{1}{2}\,\mathbb{E}_{\theta*}\Big[\| \underbrace{\nabla_x \log f(x|\theta) - \nabla_x \log f(x|\theta^*)}_{\text{“score matching”}} \|_2^2\Big]$$

- If support $\mathcal{X} = \mathbb{R}^p$, then under some mild conditions:

$$\mathcal{L}(\theta) = \mathbb{E}_{\theta*}\left[\Delta_x \log f(x|\theta) + \frac{1}{2}\|\nabla_x \log f(x|\theta)\|_2^2\right] + \text{const}$$

2. Score matching (Hyvärinen, 2005)

- X : continuous observation with support $\mathcal{X} \subset \mathbb{R}^p$
- Density $f(x|\theta^*)$ from a parametric model $f(x|\theta)$, $\theta \in \Theta$.
- Idea: Avoid log-partition function by considering divergence

$$\mathcal{L}(\theta) = \frac{1}{2}\,\mathbb{E}_{\theta*}\Big[\|\underbrace{\nabla_x \log f(x|\theta) - \nabla_x \log f(x|\theta^*)}_{\text{"score matching"}}\|_2^2\Big]$$

- If support $\mathcal{X} = \mathbb{R}^p$, then under some mild conditions:

$$\mathcal{L}(\theta) = \mathbb{E}_{\theta*}\left[\Delta_x \log f(x|\theta) + \frac{1}{2}\|\nabla_x \log f(x|\theta)\|_2^2\right] + \text{const}$$

2. Score matching (Hyvärinen, 2005)

- X : <u>continuous</u> observation with support $\mathcal{X} \subset \mathbb{R}^p$
- Density $f(x|\theta^*)$ from a parametric model $f(x|\theta)$, $\theta \in \Theta$.
- Idea: Avoid log-partition function by considering divergence

$$\mathcal{L}(\theta) = \frac{1}{2}\, \mathbb{E}_{\theta *}\big[\| \underbrace{\nabla_x \log f(x|\theta) - \nabla_x \log f(x|\theta^*)}_{\text{“score matching”}} \|_2^2\big]$$

- If support $\mathcal{X} = \mathbb{R}^p$, then under some mild conditions:

$$\mathcal{L}(\theta) = \mathbb{E}_{\theta *}\left[\Delta_x \log f(x|\theta) + \frac{1}{2}\|\nabla_x \log f(x|\theta)\|_2^2\right] + \text{const}$$

2. Score matching (Hyvärinen, 2005)

- X : <u>continuous</u> observation with support $\mathcal{X} \subset \mathbb{R}^p$
- Density $f(x|\theta^*)$ from a parametric model $f(x|\theta)$, $\theta \in \Theta$.
- Idea: Avoid log-partition function by considering divergence

$$\mathcal{L}(\theta) = \frac{1}{2}\,\mathbb{E}_{\theta*}\big[\|\underbrace{\nabla_x \log f(x|\theta) - \nabla_x \log f(x|\theta^*)}_{\text{“score matching”}}\|_2^2\big]$$

- If support $\mathcal{X} = \mathbb{R}^p$, then under some mild conditions:

$$\mathcal{L}(\theta) = \mathbb{E}_{\theta*}\left[\Delta_x \log f(x|\theta) + \frac{1}{2}\|\nabla_x \log f(x|\theta)\|_2^2\right] + \text{const}$$

Score matching

- $\mathcal{L}(\theta)$ minimized $(= 0)$ when $f(\cdot|\theta) = f(\cdot|\theta^*)$, so $\theta = \theta^*$ under identifiability.
- Estimate θ via

$$\hat{\theta} = \arg\min_{\theta} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\Delta_x \log f\left(x^i|\theta\right) + \frac{1}{2}\left\|\nabla_x \log f\left(x^i|\theta\right)\right\|_2^2\right)}_{\equiv \hat{\mathcal{L}}(\mathbf{x},\theta)}$$

- Derivatives $\frac{\partial}{\partial x} \implies$ No normalizing constant, no problems!
- Hyvärinen (2007) extends approach for $\mathcal{X} = \mathbb{R}_+^p$

  More on that later, for now denote that loss function $\hat{\mathcal{L}}_+(\mathbf{x},\theta)$.

Score matching

- $\mathcal{L}(\theta)$ minimized $(= 0)$ when $f(\cdot|\theta) = f(\cdot|\theta^*)$, so $\theta = \theta^*$ under identifiability.
- Estimate θ via

$$\hat{\theta} = \arg\min_{\theta} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\Delta_x \log f\left(x^i|\theta\right) + \frac{1}{2}\left\|\nabla_x \log f\left(x^i|\theta\right)\right\|_2^2\right)}_{\equiv \hat{\mathcal{L}}(\mathbf{x},\theta)}$$

- Derivatives $\frac{\partial}{\partial x} \implies$ No normalizing constant, no problems!
- Hyvärinen (2007) extends approach for $\mathcal{X} = \mathbb{R}_+^p$

  More on that later, for now denote that loss function $\hat{\mathcal{L}}_+(\mathbf{x}, \theta)$.

Score matching

- $\mathcal{L}(\theta)$ minimized $(= 0)$ when $f(\cdot|\theta) = f(\cdot|\theta^*)$, so $\theta = \theta^*$ under identifiability.

- Estimate θ via

$$\hat{\theta} = \arg\min_{\theta} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\Delta_x \log f\left(x^i|\theta\right) + \frac{1}{2}\left\|\nabla_x \log f\left(x^i|\theta\right)\right\|_2^2\right)}_{\equiv \hat{\mathcal{L}}(\mathbf{x},\theta)}$$

- Derivatives $\frac{\partial}{\partial x} \implies$ No normalizing constant, no problems!

- Hyvärinen (2007) extends approach for $\mathcal{X} = \mathbb{R}_+^p$

  More on that later, for now denote that loss function $\hat{\mathcal{L}}_+(\mathbf{x},\theta)$.

Score matching

- $\mathcal{L}(\theta)$ minimized $(= 0)$ when $f(\cdot|\theta) = f(\cdot|\theta^*)$, so $\theta = \theta^*$ under identifiability.
- Estimate θ via

$$\hat{\theta} = \arg\min_{\theta} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\Delta_x \log f\left(x^i|\theta\right) + \frac{1}{2}\left\|\nabla_x \log f\left(x^i|\theta\right)\right\|_2^2\right)}_{\equiv \hat{\mathcal{L}}(\mathbf{x},\theta)}$$

- Derivatives $\frac{\partial}{\partial x} \implies$ No normalizing constant, no problems!
- Hyvärinen (2007) extends approach for $\mathcal{X} = \mathbb{R}_+^p$

  More on that later, for now denote that loss function $\hat{\mathcal{L}}_+(\mathbf{x}, \theta)$.

Quadratic loss

- Pairwise interaction (PI) models,

$$\log f(x|\theta) = \sum_{1 \leqslant j \leqslant k \leqslant p} \theta_{jk} t_{jk}(x_j, x_k) + \psi(\theta),$$

  are exponential families.

- Then, $\hat{\mathcal{L}}(\mathbf{x}, \theta)$ and $\hat{\mathcal{L}}_+(\mathbf{x}, \theta)$ are semi-definite quadratic.

- Generically, the ℓ_1-regularized objective is

$$\boxed{\hat{\mathcal{L}}_{\lambda_n}(\mathbf{x}, \theta) = \frac{1}{2}\theta^T \mathbf{\Gamma}(\mathbf{x})\theta - \gamma(\mathbf{x})^T \theta + \cancel{c(\mathbf{x})} + \lambda_n \|\theta\|_1}$$

$$\mathbf{\Gamma}(\mathbf{x}) \geq 0 \text{ is } p^2 \times p^2 \text{ block-diagonal}$$

- Lasso-type objective: simple computation and theory!

Quadratic loss

- Pairwise interaction (PI) models,

$$\log f(x|\theta) = \sum_{1 \leqslant j \leqslant k \leqslant p} \theta_{jk} t_{jk}(x_j, x_k) + \psi(\theta),$$

  are exponential families.

- Then, $\hat{\mathcal{L}}(\mathbf{x}, \theta)$ and $\hat{\mathcal{L}}_+(\mathbf{x}, \theta)$ are semi-definite quadratic.

- Generically, the ℓ_1-regularized objective is

$$\boxed{\hat{\mathcal{L}}_{\lambda_n}(\mathbf{x}, \theta) = \frac{1}{2}\theta^T \mathbf{\Gamma}(\mathbf{x})\theta - \gamma(\mathbf{x})^T\theta + \cancel{c(\mathbf{x})} + \lambda_n \|\theta\|_1}$$

$$\mathbf{\Gamma}(\mathbf{x}) \geq 0 \text{ is } p^2 \times p^2 \text{ block-diagonal}$$

- Lasso-type objective: simple computation and theory!

Quadratic loss

- Pairwise interaction (PI) models,

$$\log f(x|\theta) = \sum_{1 \leqslant j \leqslant k \leqslant p} \theta_{jk} t_{jk}(x_j, x_k) + \psi(\theta),$$

  are exponential families.

- Then, $\hat{\mathcal{L}}(\mathbf{x}, \theta)$ and $\hat{\mathcal{L}}_+(\mathbf{x}, \theta)$ are semi-definite quadratic.

- Generically, the ℓ_1-regularized objective is

$$\boxed{\hat{\mathcal{L}}_{\lambda_n}(\mathbf{x}, \theta) = \frac{1}{2}\theta^T \boldsymbol{\Gamma}(\mathbf{x})\theta - \gamma(\mathbf{x})^T \theta + \cancel{c(\mathbf{x})} + \lambda_n \|\theta\|_1}$$

$\boldsymbol{\Gamma}(\mathbf{x}) \geq 0$ is $p^2 \times p^2$ block-diagonal

- Lasso-type objective: simple computation and theory!

Quadratic loss

- Pairwise interaction (PI) models,

$$\log f(x|\theta) = \sum_{1 \leqslant j \leqslant k \leqslant p} \theta_{jk} t_{jk}(x_j, x_k) + \psi(\theta),$$

  are exponential families.

- Then, $\hat{\mathcal{L}}(\mathbf{x}, \theta)$ and $\hat{\mathcal{L}}_+(\mathbf{x}, \theta)$ are semi-definite quadratic.

- Generically, the ℓ_1-regularized objective is

$$\boxed{\hat{\mathcal{L}}_{\lambda_n}(\mathbf{x}, \theta) = \frac{1}{2}\theta^T \mathbf{\Gamma}(\mathbf{x})\theta - \gamma(\mathbf{x})^T \theta + \cancel{c(\mathbf{x})} + \lambda_n \|\theta\|_1}$$

$$\mathbf{\Gamma}(\mathbf{x}) \succeq 0 \text{ is } p^2 \times p^2 \text{ block-diagonal}$$

- Lasso-type objective: simple computation and theory!

Quadratic loss

- Pairwise interaction (PI) models,

$$\log f(x|\theta) = \sum_{1 \leqslant j \leqslant k \leqslant p} \theta_{jk} t_{jk}(x_j, x_k) + \psi(\theta),$$

  are exponential families.

- Then, $\hat{\mathcal{L}}(\mathbf{x}, \theta)$ and $\hat{\mathcal{L}}_+(\mathbf{x}, \theta)$ are semi-definite quadratic.

- Generically, the ℓ_1-regularized objective is

$$\boxed{\hat{\mathcal{L}}_{\lambda_n}(\mathbf{x}, \theta) = \frac{1}{2}\theta^T \mathbf{\Gamma}(\mathbf{x})\theta - \gamma(\mathbf{x})^T \theta + \cancel{c(\mathbf{x})} + \lambda_n \|\theta\|_1}$$

$$\mathbf{\Gamma}(\mathbf{x}) \succeq 0 \text{ is } p^2 \times p^2 \text{ block-diagonal}$$

- Lasso-type objective: simple computation and theory!

Gaussian theory: CIG/support recovery

WLOG, consider $\mu = 0$. Define $\mathbf{W} = \frac{\mathbf{x}^T\mathbf{x}}{n}$ (sample covariance).

Objective:

$$\hat{\mathcal{L}}_{\lambda_n}(\mathbf{K}) = -\mathrm{tr}(\mathbf{K}) + \frac{1}{2}\mathrm{tr}(\mathbf{KKW}) + \lambda_n \|\mathbf{K}\|_1.$$

Taking $\theta = \mathrm{vec}(\mathbf{K})$, we have

$$\mathbf{\Gamma}(\mathbf{x}) = I_{p\times p} \otimes \mathbf{W}, \qquad \text{and} \qquad \gamma(\mathbf{x}) = \gamma = \mathrm{vec}(\mathbf{I}_{p\times p}).$$

Under irrepresentability and beta-min condition, CIG recovered w.h.p. if

$$\boxed{n \geq Cd^2 \log p}$$

where d is maximal node degree; $\lambda_n \asymp \sqrt{(\log p)/n}$

"State of the art"...

Gaussian theory: CIG/support recovery

WLOG, consider $\mu = 0$. Define $\mathbf{W} = \frac{\mathbf{x}^T\mathbf{x}}{n}$ (sample covariance).

Objective:

$$\hat{\mathcal{L}}_{\lambda_n}(\mathbf{K}) = -\mathrm{tr}(\mathbf{K}) + \frac{1}{2}\mathrm{tr}(\mathbf{KKW}) + \lambda_n\|\mathbf{K}\|_1.$$

Taking $\theta = \mathrm{vec}(\mathbf{K})$, we have

$$\mathbf{\Gamma}(\mathbf{x}) = I_{p\times p} \otimes \mathbf{W}, \qquad \text{and} \qquad \gamma(\mathbf{x}) = \gamma = \mathrm{vec}(\mathbf{I}_{p\times p}).$$

Under irrepresentability and beta-min condition, CIG recovered w.h.p. if

$$\boxed{n \geq Cd^2 \log p}$$

where d is maximal node degree; $\lambda_n \asymp \sqrt{(\log p)/n}$

"State of the art"...

Gaussian theory: CIG/support recovery

WLOG, consider $\mu = 0$. Define $\mathbf{W} = \frac{\mathbf{x}^T\mathbf{x}}{n}$ (sample covariance).

Objective:

$$\hat{\mathcal{L}}_{\lambda_n}(\mathbf{K}) = -\mathrm{tr}(\mathbf{K}) + \frac{1}{2}\mathrm{tr}(\mathbf{KKW}) + \lambda_n \|\mathbf{K}\|_1.$$

Taking $\theta = \mathrm{vec}(\mathbf{K})$, we have

$$\mathbf{\Gamma}(\mathbf{x}) = I_{p\times p} \otimes \mathbf{W}, \qquad \text{and} \qquad \gamma(\mathbf{x}) = \gamma = \mathrm{vec}(\mathbf{I}_{p\times p}).$$

Under irrepresentability and beta-min condition, CIG recovered w.h.p. if

$$\boxed{n \succeq Cd^2 \log p}$$

where d is maximal node degree; $\lambda_n \asymp \sqrt{(\log p)/n}$

"State of the art"...

Non-negative Gaussians

Gaussian truncated: $f(x|\mathbf{K}) \propto \exp\{-\frac{1}{2}x^T\mathbf{K}x\}, \quad x \in \mathbb{R}_+^p$

Objective: (may think log transform ...)

$$\hat{\mathcal{L}}_{+,\lambda_n}(\mathbf{K}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} 2x_{ij}x^{(i)T}\kappa_j - x_{ij}^2\kappa_{jj} + \frac{1}{2}\kappa_j^T\left(x_{ij}^2 x^{(i)}x^{(i)T}\right)\kappa_j + \lambda_n\|\mathbf{K}\|_1$$

Under irrepresentability and beta-min condition, CIG recovered w.h.p. if

$$\boxed{n \geq d^2(\log p)^8}$$

Rate *not* sharp; based on concentration inequality for log-concave densities

Non-negative Gaussians

Gaussian truncated: $f(x|\mathbf{K}) \propto \exp\{-\frac{1}{2}x^T\mathbf{K}x\}, \quad x \in \mathbb{R}_+^p$

Objective: (may think log transform ...)

$$\hat{\mathcal{L}}_{+,\lambda_n}(\mathbf{K}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} 2x_{ij}x^{(i)T}\kappa_j - x_{ij}^2\kappa_{jj} + \frac{1}{2}\kappa_j^T\left(x_{ij}^2 x^{(i)}x^{(i)T}\right)\kappa_j + \lambda_n\|\mathbf{K}\|_1$$

Under irrepresentability and beta-min condition, CIG recovered w.h.p. if

$$\boxed{n \succeq d^2(\log p)^8}$$

Rate *not* sharp; based on concentration inequality for log-concave densities

Irrepresentability condition

There exists an $\alpha \in (0,1]$ such that

$$\left|\!\left|\!\left| \boldsymbol{\Gamma}^*_{S^c S} (\boldsymbol{\Gamma}^*_{SS})^{-1} \right|\!\right|\!\right|_\infty \leqslant (1-\alpha).$$

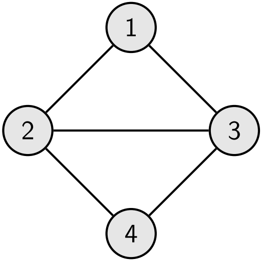
- Intuition:
  Regression coefficients for 'Noise' vs. 'Signal' not too large.
- Neighborhood selection: condition on covariance matrix
- glasso: condition on Hessian of log-determinant
- In Example from Meinshausen (2008) we have the implications

$$\text{glasso} \Rightarrow \text{Regularized score matching} \Rightarrow \text{MB}$$

Necessary conditions in a Gaussian example

Consider normal distribution with below covariance. Its CIG is the bottom-left graph.

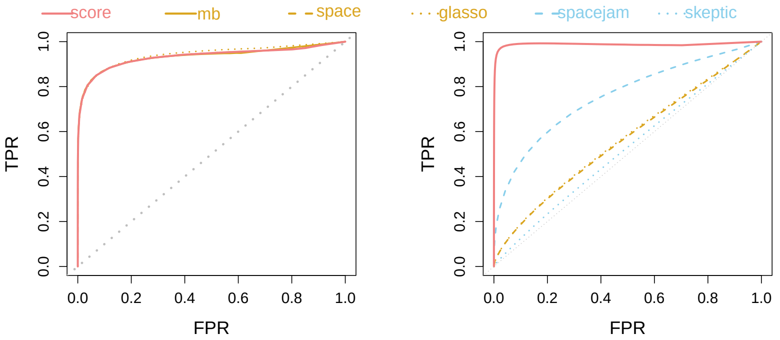
$$\Sigma = \begin{pmatrix} 1 & \rho & \rho & 2\rho^2 \\ \rho & 1 & 0 & \rho \\ \rho & 0 & 1 & \rho \\ 2\rho^2 & \rho & \rho & 1 \end{pmatrix}, \qquad \rho \geqslant 0.$$



Necessary for graph recovery:

- Reg. score matching: $\rho \leqslant 0.41$
- Neighborhood selection: $\rho \leqslant 0.5$
- glasso: $\rho \leqslant 0.23$

Simulation



$n = 600,\ p = 1000$ lattice

$n = 750,\ p = 625$ lattice

$$\exp\left\{-\frac{1}{2}x^T\mathbf{K}x\right\}$$

$$\exp\left\{-\frac{1}{2}\left[\sum_{j\leqslant k} A_{jk}x_j^2x_k^2 + \sum_j C_jx_j\right]\right\}$$

Illustration of analysis of RNAseq data using truncated normal models in paper. . .

3. Missing data: Problem setup

Suppose observations are missing completely-at-random.

We observe $\mathbf{z}$:

$$z_{ij} = x_{ij} \times \delta_{ij}$$
$$\delta_{ij} \sim \text{Bernoulli}(1-\rho), \quad \rho \in [0,1)$$

δ_{ij}'s represent the observed indicators.

Can also consider variable-dependent missingness:

$$\delta_{ij} \sim \text{Bernoulli}(1-\rho_j), \; \rho_j \in [0,1) \; \forall j$$

Question: how do we adjust for missing values?

3. Missing data: Problem setup

Suppose observations are missing completely-at-random.

We observe $\mathbf{z}$:

$$
\begin{aligned}
z_{ij} &= x_{ij} \times \delta_{ij} \\
\delta_{ij} &\sim \text{Bernoulli}(1-\rho), \quad \rho \in [0,1)
\end{aligned}
$$

δ_{ij}'s represent the observed indicators.

Can also consider variable-dependent missingness:

$$\delta_{ij} \sim \text{Bernoulli}(1-\rho_j),\ \rho_j \in [0,1)\ \forall j$$

Question: how do we adjust for missing values?

3. Missing data: Problem setup

Suppose observations are missing completely-at-random.

We observe $\mathbf{z}$:

$$z_{ij} = x_{ij} \times \delta_{ij}$$
$$\delta_{ij} \sim \text{Bernoulli}(1 - \rho), \quad \rho \in [0, 1)$$

δ_{ij}'s represent the observed indicators.

Can also consider variable-dependent missingness:

$$\delta_{ij} \sim \text{Bernoulli}(1 - \rho_j), \;\; \rho_j \in [0, 1) \; \forall j$$

Question: how do we adjust for missing values?

3. Missing data: Problem setup

Suppose observations are missing completely-at-random.

We observe $\mathbf{z}$:

$$z_{ij} = x_{ij} \times \delta_{ij}$$
$$\delta_{ij} \sim \text{Bernoulli}(1 - \rho), \quad \rho \in [0, 1)$$

δ_{ij}'s represent the observed indicators.

Can also consider variable-dependent missingness:

$$\delta_{ij} \sim \text{Bernoulli}(1 - \rho_j), \; \rho_j \in [0, 1) \; \forall j$$

Question: how do we adjust for missing values?

3. Missing data: Problem setup

Suppose observations are missing completely-at-random.

We observe $\mathbf{z}$:

$$
\begin{aligned}
z_{ij} &= x_{ij} \times \delta_{ij} \\
\delta_{ij} &\sim \text{Bernoulli}(1-\rho), \quad \rho \in [0,1)
\end{aligned}
$$

δ_{ij}'s represent the observed indicators.

Can also consider variable-dependent missingness:

$$\delta_{ij} \sim \text{Bernoulli}(1-\rho_j),\ \ \rho_j \in [0,1)\ \forall j$$

Question: how do we adjust for missing values?

Using surrogates

$$\boxed{\hat{\mathcal{L}}_{\lambda_n}(\mathbf{x}, \theta) = \frac{1}{2}\theta^T \mathbf{\Gamma}(\mathbf{x})\theta - \gamma(\mathbf{x})^T\theta + \cancel{c(\mathbf{x})} + \lambda_n \|\theta\|_1}$$

Idea: Use surrogates $\tilde{\mathbf{\Gamma}}(\mathbf{z})$ and $\tilde{\gamma}(\mathbf{z})$ in place of $\mathbf{\Gamma}(\mathbf{x})$ and $\gamma(\mathbf{x})$.

Criterion: Surrogates must be unbiased, i.e.,

$$\mathbb{E}_{\theta*}[\mathbf{\Gamma}(\mathbf{X})] = \mathbb{E}_{\theta*}[\tilde{\mathbf{\Gamma}}(\mathbf{Z})]$$
$$\mathbb{E}_{\theta*}[\gamma(\mathbf{X})] = \mathbb{E}_{\theta*}[\tilde{\gamma}(\mathbf{Z})]$$

We extend the ideas presented in:

- Loh and Wainwright (2012): multiplicative de-biasing
- Kolar and Xing (2012): use only complete tuples

Using surrogates

$$\boxed{\hat{\mathcal{L}}_{\lambda_n}(\mathbf{x}, \theta) = \frac{1}{2}\theta^T \mathbf{\Gamma}(\mathbf{x})\theta - \gamma(\mathbf{x})^T\theta + \cancel{c(\mathbf{x})} + \lambda_n\|\theta\|_1}$$

<u>Idea</u>: Use surrogates $\tilde{\mathbf{\Gamma}}(\mathbf{z})$ and $\tilde{\gamma}(\mathbf{z})$ in place of $\mathbf{\Gamma}(\mathbf{x})$ and $\gamma(\mathbf{x})$.

<u>Criterion</u>: Surrogates must be unbiased, i.e.,

$$\mathbb{E}_{\theta*}[\mathbf{\Gamma}(\mathbf{X})] = \mathbb{E}_{\theta*}[\tilde{\mathbf{\Gamma}}(\mathbf{Z})]$$
$$\mathbb{E}_{\theta*}[\gamma(\mathbf{X})] = \mathbb{E}_{\theta*}[\tilde{\gamma}(\mathbf{Z})]$$

We extend the ideas presented in:

- Loh and Wainwright (2012): multiplicative de-biasing
- Kolar and Xing (2012): use only complete tuples

Using surrogates

$$\boxed{\hat{\mathcal{L}}_{\lambda_n}(\mathbf{x}, \theta) = \frac{1}{2}\theta^T \boldsymbol{\Gamma}(\mathbf{x})\theta - \gamma(\mathbf{x})^T\theta + \cancel{c(\mathbf{x})} + \lambda_n\|\theta\|_1}$$

Idea: Use surrogates $\tilde{\boldsymbol{\Gamma}}(\mathbf{z})$ and $\tilde{\gamma}(\mathbf{z})$ in place of $\boldsymbol{\Gamma}(\mathbf{x})$ and $\gamma(\mathbf{x})$.

Criterion: Surrogates must be unbiased, i.e.,

$$\mathbb{E}_{\theta*}[\boldsymbol{\Gamma}(\mathbf{X})] = \mathbb{E}_{\theta*}[\tilde{\boldsymbol{\Gamma}}(\mathbf{Z})]$$
$$\mathbb{E}_{\theta*}[\gamma(\mathbf{X})] = \mathbb{E}_{\theta*}[\tilde{\gamma}(\mathbf{Z})]$$

We extend the ideas presented in:

- Loh and Wainwright (2012): multiplicative de-biasing
- Kolar and Xing (2012): use only complete tuples

Using surrogates

$$\boxed{\hat{\mathcal{L}}_{\lambda_n}(\mathbf{x},\theta) = \frac{1}{2}\theta^T\boldsymbol{\Gamma}(\mathbf{x})\theta - \gamma(\mathbf{x})^T\theta + \cancel{c(\mathbf{x})} + \lambda_n\|\theta\|_1}$$

Idea: Use surrogates $\tilde{\boldsymbol{\Gamma}}(\mathbf{z})$ and $\tilde{\gamma}(\mathbf{z})$ in place of $\boldsymbol{\Gamma}(\mathbf{x})$ and $\gamma(\mathbf{x})$.

Criterion: Surrogates must be unbiased, i.e.,

$$\mathbb{E}_{\theta*}[\boldsymbol{\Gamma}(\mathbf{X})] = \mathbb{E}_{\theta*}[\tilde{\boldsymbol{\Gamma}}(\mathbf{Z})]$$
$$\mathbb{E}_{\theta*}[\gamma(\mathbf{X})] = \mathbb{E}_{\theta*}[\tilde{\gamma}(\mathbf{Z})]$$

We extend the ideas presented in:

- Loh and Wainwright (2012): multiplicative de-biasing
- Kolar and Xing (2012): use only complete tuples

A demonstration (centered Gaussian)

Recall that:

$$\boldsymbol{\Gamma}(\mathbf{x}) = I_{p\times p} \otimes \mathbf{W}, \qquad \text{and} \qquad \gamma(\mathbf{x}) = \mathrm{vec}(\mathbf{I}_{p\times p}).$$

- Surrogates based on de-biasing:

$$\tilde{\boldsymbol{\Gamma}}(\mathbf{z}) = \boldsymbol{\Gamma}(\mathbf{z}) \div (\mathbf{I}_{p\times p} \otimes \mathbf{M}) \qquad \tilde{\gamma} = \gamma,$$

with $\mathbf{M} = \left(m_{jk}\right) \in \mathbb{R}^{p\times p}$ and

$$m_{jk} = \begin{cases} 1-\rho & \text{if } j = k \\ (1-\rho)^2 & \text{if } j \neq k \end{cases}.$$

- Surrogates based on complete tuples: straightforward

A demonstration (centered Gaussian)

Recall that:

$$\mathbf{\Gamma}(\mathbf{x}) = I_{p\times p} \otimes \mathbf{W}, \qquad \text{and} \qquad \gamma(\mathbf{x}) = \mathrm{vec}(\mathbf{I}_{p\times p}).$$

- Surrogates based on de-biasing:

$$\tilde{\mathbf{\Gamma}}(\mathbf{z}) = \mathbf{\Gamma}(\mathbf{z}) \div (\mathbf{I}_{p\times p} \otimes \mathbf{M}) \qquad \tilde{\gamma} = \gamma,$$

with $\mathbf{M} = \left(m_{jk}\right) \in \mathbb{R}^{p\times p}$ and

$$m_{jk} = \begin{cases} 1-\rho & \text{if } j = k \\ (1-\rho)^2 & \text{if } j \neq k \end{cases}.$$

- Surrogates based on complete tuples: straightforward

A demonstration (centered Gaussian)

Recall that:

$$\boldsymbol{\Gamma}(\mathbf{x}) = I_{p \times p} \otimes \mathbf{W}, \qquad \text{and} \qquad \gamma(\mathbf{x}) = \mathrm{vec}(\mathbf{I}_{p \times p}).$$

- Surrogates based on de-biasing:

$$\tilde{\boldsymbol{\Gamma}}(\mathbf{z}) = \boldsymbol{\Gamma}(\mathbf{z}) \oslash (\mathbf{I}_{p \times p} \otimes \mathbf{M}) \qquad \tilde{\gamma} = \gamma,$$

  with $\mathbf{M} = \left(m_{jk}\right) \in \mathbb{R}^{p \times p}$ and

$$m_{jk} = \begin{cases} 1 - \rho & \text{if } j = k \\ (1 - \rho)^2 & \text{if } j \neq k \end{cases}.$$

- Surrogates based on complete tuples: straightforward

Non-convex objective

- Surrogate-based loss need not be convex ($\tilde{\mathbf{\Gamma}}(\mathbf{z})$ not p.s.d.)
- Instead:

$$\hat{\theta} \in \arg \min_{\forall_j \|\theta_{\cdot j}\|_1 \leqslant R} \frac{1}{2}\theta^T \tilde{\mathbf{\Gamma}}(\mathbf{z})\theta - \tilde{\gamma}(\mathbf{z})^T \theta + \lambda_n \|\theta\|_1$$

  Two tuning parameters: R and λ_n.

- Paralleling/extending the complete data case, possible to get high-dimensional consistency/support recovery (see Sara's talk)

  Sample size scaling as in complete data case:

$$n \geq c(\rho) d^2 \log p \quad \text{(Gaussian)}$$
$$n \geq c(\rho) d^2 (\log p)^8 \quad \text{(Non-negative Gaussian)}$$

Non-convex objective

- Surrogate-based loss need not be convex ($\tilde{\mathbf{\Gamma}}(\mathbf{z})$ not p.s.d.)
- Instead:

$$\hat{\theta} \in \arg \min_{\forall_j \|\theta_{\cdot j}\|_1 \leqslant R} \frac{1}{2}\theta^T \tilde{\mathbf{\Gamma}}(\mathbf{z})\theta - \tilde{\gamma}(\mathbf{z})^T \theta + \lambda_n \|\theta\|_1$$

Two tuning parameters: R and λ_n.

- Paralleling/extending the complete data case, possible to get high-dimensional consistency/support recovery (see Sara's talk)

Sample size scaling as in complete data case:

$$n \geq c(\rho) d^2 \log p \quad \text{(Gaussian)}$$
$$n \geq c(\rho) d^2 (\log p)^8 \quad \text{(Non-negative Gaussian)}$$

Non-convex objective

- Surrogate-based loss need not be convex ($\tilde{\mathbf{\Gamma}}(\mathbf{z})$ not p.s.d.)
- Instead:

$$\hat{\theta} \in \arg \min_{\forall_j \|\theta_{\cdot j}\|_1 \leqslant R} \frac{1}{2}\theta^T \tilde{\mathbf{\Gamma}}(\mathbf{z})\theta - \tilde{\gamma}(\mathbf{z})^T\theta + \lambda_n \|\theta\|_1$$

  Two tuning parameters: R and λ_n.

- Paralleling/extending the complete data case, possible to get high-dimensional consistency/support recovery (see Sara's talk)

  Sample size scaling as in complete data case:

$$n \succeq c(\rho) d^2 \log p \quad \text{(Gaussian)}$$
$$n \succeq c(\rho) d^2 (\log p)^8 \quad \text{(Non-negative Gaussian)}$$

Numerical experiments ($p = 100$, $n = 1000$)

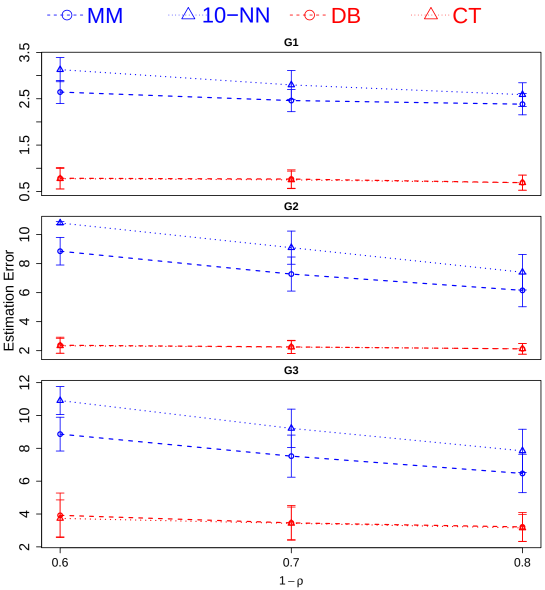
$$f(x|\mathbf{K}) \propto \exp\left\{\frac{1}{2}x^T\mathbf{K}x\right\},$$

$$x \in \mathbb{R}_+^p$$

G1: linear chain

G2: lattice

G3: Erdős-Rényi



Estimation error

$$= \max_j \|\hat{\theta}_{\cdot j} - \theta^*_{\cdot j}\|_1$$

4. Modification of non-negative score matching loss

- For the case of support equal to $\mathbb{R}_+^p$, Hyvärinen (2007) proposes

$$\mathcal{L}_+(f) = \int_{\mathbb{R}_+^p} f_0(x) \left[\left\| \nabla_x \log f(x) \circ x - \nabla_x \log f_0(x) \circ x \right\|_2^2 \right] dx,$$

- Under mild conditions,

$$\mathcal{L}_+(f) = \int_{\mathbb{R}_+^p} f_0(x) S_+(x, f)\, dx \; + \; \text{const}, \quad \text{with}$$

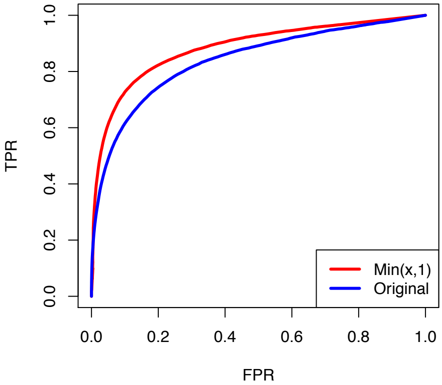
$$S_+(x, f) = \sum_{j=1}^{p} \left[2x_j \frac{\partial \log f(x)}{\partial x_j} + x_j^2 \frac{\partial^2 \log f(x)}{\partial x_j^2} + \frac{1}{2} x_j^2 \left(\frac{\partial \log f(x)}{\partial x_j} \right)^2 \right].$$

- Non-neg Gaussian example:

$$\hat{\mathcal{L}}_+(\mathbf{K}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} 2x_{ij} x^{(i)T} \kappa_j - x_{ij}^2 \kappa_{jj} + \frac{1}{2} \kappa_j^T \left(x_{ij}^2 x^{(i)} x^{(i)T} \right) \kappa_j$$

Ongoing work

- Idea: Replace "$\circ x$" by bounded function
- Improved performance and theoretical guarantees



$p = 100$, $n = 1000$, Erdos-Renyi graph with 0.03 edge density.

Conclusion

- No normalizing constants, no problems
- Quadratic loss also for non-Gaussian models
- Convenient computationally, tractable theoretically
- EJS paper: Lin et al. (2016)
- Related work:
    - Liu and Luo (2015): SCIO = Gaussian case
    - Zhang and Zou (2014): D-trace loss = Gaussian case
    - Forbes and Lauritzen (2015): Colored Gaussian graphical models
    - Janofsky (2015): exponential series models
    - Sun et al. (2015): infinite-dimensional exponential families
    - Yu et al. (2016): confidence intervals

Conclusion

- No normalizing constants, no problems
- Quadratic loss also for non-Gaussian models
- Convenient computationally, tractable theoretically
- EJS paper: Lin et al. (2016)
- Related work:
  - Liu and Luo (2015): SCIO = Gaussian case
  - Zhang and Zou (2014): D-trace loss = Gaussian case
  - Forbes and Lauritzen (2015): Colored Gaussian graphical models
  - Janofsky (2015): exponential series models
  - Sun et al. (2015): infinite-dimensional exponential families
  - Yu et al. (2016): confidence intervals

Conclusion

- No normalizing constants, no problems
- Quadratic loss also for non-Gaussian models
- Convenient computationally, tractable theoretically
- EJS paper: Lin et al. (2016)
- Related work:
  - Liu and Luo (2015): SCIO = Gaussian case
  - Zhang and Zou (2014): D-trace loss = Gaussian case
  - Forbes and Lauritzen (2015): Colored Gaussian graphical models
  - Janofsky (2015): exponential series models
  - Sun et al. (2015): infinite-dimensional exponential families
  - Yu et al. (2016): confidence intervals

Conclusion

- No normalizing constants, no problems
- Quadratic loss also for non-Gaussian models
- Convenient computationally, tractable theoretically
- EJS paper: Lin et al. (2016)
- Related work:
  - Liu and Luo (2015): SCIO = Gaussian case
  - Zhang and Zou (2014): D-trace loss = Gaussian case
  - Forbes and Lauritzen (2015): Colored Gaussian graphical models
  - Janofsky (2015): exponential series models
  - Sun et al. (2015): infinite-dimensional exponential families
  - Yu et al. (2016): confidence intervals

Conclusion

- No normalizing constants, no problems
- Quadratic loss also for non-Gaussian models
- Convenient computationally, tractable theoretically
- EJS paper: Lin et al. (2016)
- Related work:
  - Liu and Luo (2015): SCIO = Gaussian case
  - Zhang and Zou (2014): D-trace loss = Gaussian case
  - Forbes and Lauritzen (2015): Colored Gaussian graphical models
  - Janofsky (2015): exponential series models
  - Sun et al. (2015): infinite-dimensional exponential families
  - Yu et al. (2016): confidence intervals

References I

Forbes, P. G. M. and Lauritzen, S. (2015), "Linear estimating equations for exponential families with application to Gaussian linear concentration models," *Linear Algebra Appl.*, 473, 261–283.

Janofsky, E. (2015), "Exponential series approaches for nonparametric graphical models," Ph.D. thesis, The University of Chicago.

Lin, L., Drton, M., and Shojaie, A. (2016), "Estimation of high-dimensional graphical models using regularized score matching," *Electron. J. Stat.*, 10, 806–854.

Liu, W. and Luo, X. (2015), "Fast and adaptive sparse precision matrix estimation in high dimensions," *J. Multivariate Anal.*, 135, 153–162.

Sun, S., Kolar, M., and Xu, J. (2015), "Learning structured densities via infinite dimensional exponential families," in *NIPS*, pp. 2287–2295.

Yu, M., Kolar, M., and Gupta, V. (2016), "Statistical Inference for Pairwise Graphical Models Using Score Matching," in *NIPS*, pp. 2829–2837.

Zhang, T. and Zou, H. (2014), "Sparse precision matrix estimation via lasso penalized D-trace loss," *Biometrika*, 101, 103–120.