Two-Sample Inference for High-Dimensional Markov Networks

Mladen Kolar

University of Chicago
Booth School of Business

Symposium on Data Science and Statistics
2019

This is joint work with

Byol Kim	at the University of Chicago,	and
Song Liu	at the University of Bristol	

Additional details can be found in

Byol Kim, Song Liu, and Mladen Kolar, *Two-sample inference for high-dimensional Markov networks*, arXiv e-prints (2019), arXiv:1905.00466.

Motivation: Alertness and motor control study

Experimental data from a pilot study, provided in processed form by Dr. Jade Thai and Dr. Christelle Langley at the University of Bristol.

Two time series of fMRI measurements at 0.906 s intervals from 116 ROIs.

A subject with multiple sclerosis and a healthy control were asked to squeeze and release a hand dynamometer while watching a sequence of visual stimuli:

1. (sensorimotor task) squeeze a hand dynamometer while watching the screen,
2. (intrinsic alertness task) squeeze a hand dynamometer after a trigger stimulus,
3. (extrinsic alertness task) the same as Task 2, but each trigger is preceded by another stimulus.

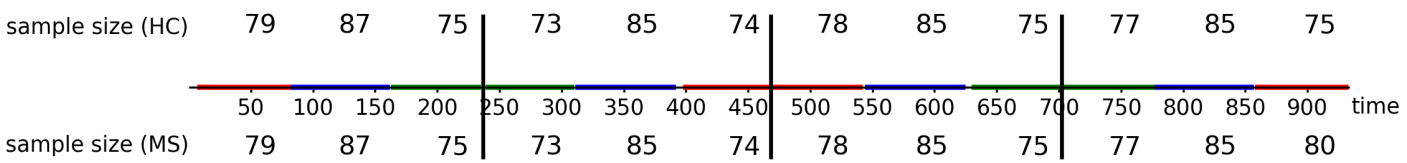


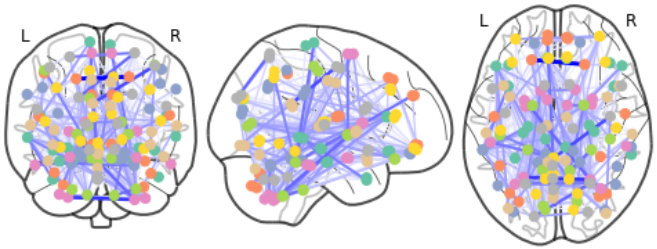
Figure: Task design (Task 1 - blue, Task 2 - green, Task 3 - red)

Motivation: Alertness and motor control study

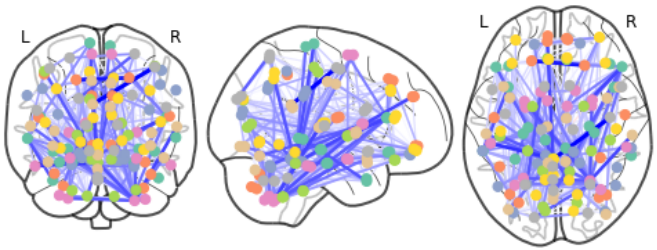
fMRI measurements
at 0.906 s intervals from 116 ROIs
measured during one of the three
tasks:

1. a sensorimotor task
   ($n_1 = 342$)
2. an intrinsic alertness task
   ($n_2 = 300$), and
3. an extrinsic alertness task
   ($n_3 = 306$).

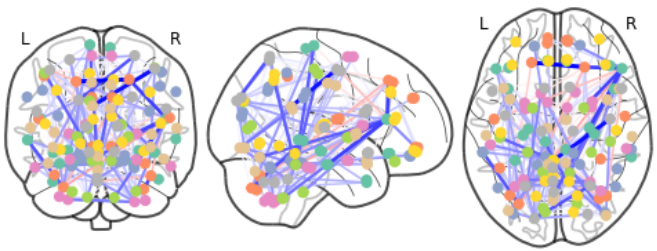
(a) Sensorimotor



(b) Intrinsic Alertness



(c) Extrinsic Alertness



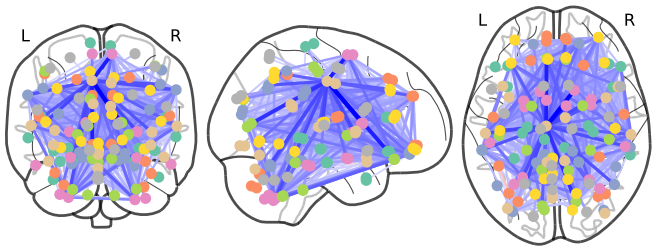
Motivation: Alertness and motor control study

fMRI measurements
at 0.906 s intervals from 116 ROIs
measured during one of the three
tasks:

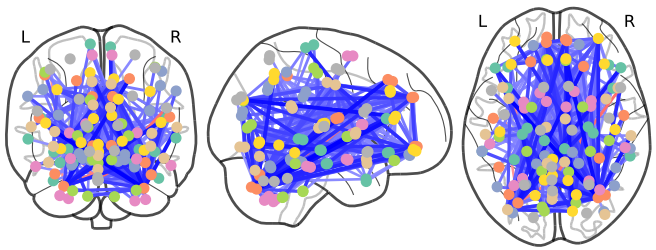
1. a sensorimotor task
   ($n_1 = 342$)
2. an intrinsic alertness task
   ($n_2 = 300$), and
3. an extrinsic alertness task
   ($n_3 = 306$).

(a) Sensorimotor vs Intrinsic Alertness

(b) Sensorimotor vs Extrinsic Alertness



(c) Intrinsic vs Extrinsic Alertness



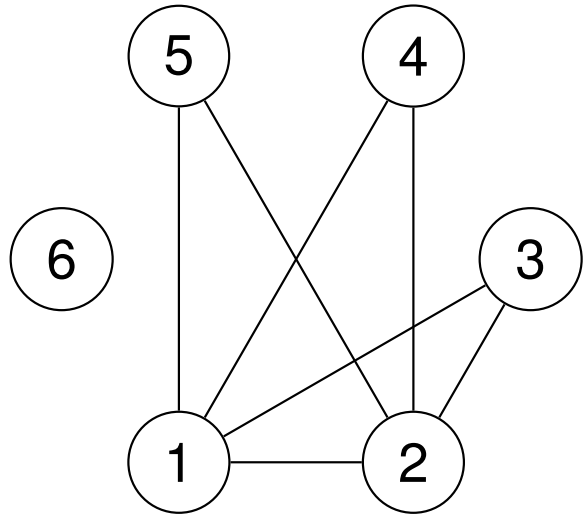
Probabilistic graphical models

A graphical model captures conditional independence relationships.

- Random vector $\mathbf{X} = (X_1, \ldots, X_m)$
- Graph $\mathbf{\Gamma} = (V, E)$ with $V = \{1, \ldots, m\}$

$$(u, v) \notin E \iff X_u \perp\!\!\!\perp X_v \mid X_{[m]\setminus\{u,v\}}$$

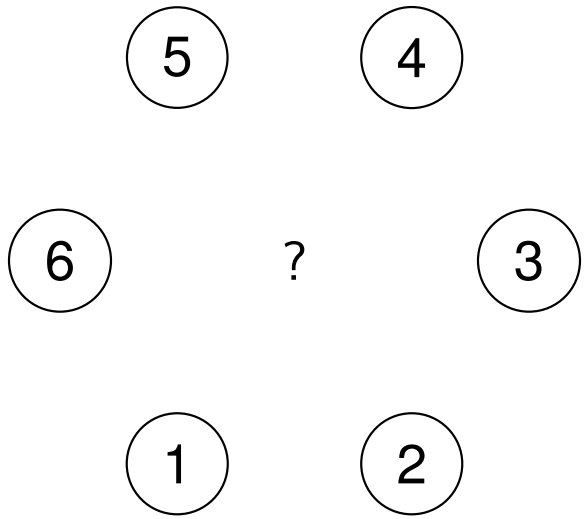
$$X_1 \perp\!\!\!\perp X_6 \mid X_2, \ldots, X_5$$



One-sample problem for graphical models

Given an i.i.d. sample $\mathbf{X}_n = \{\mathbf{X}^{(i)}\}_{i=1}^n$ from a distribution $\mathbb{P} \in \mathcal{P}$, learn the set of conditional independence relationships

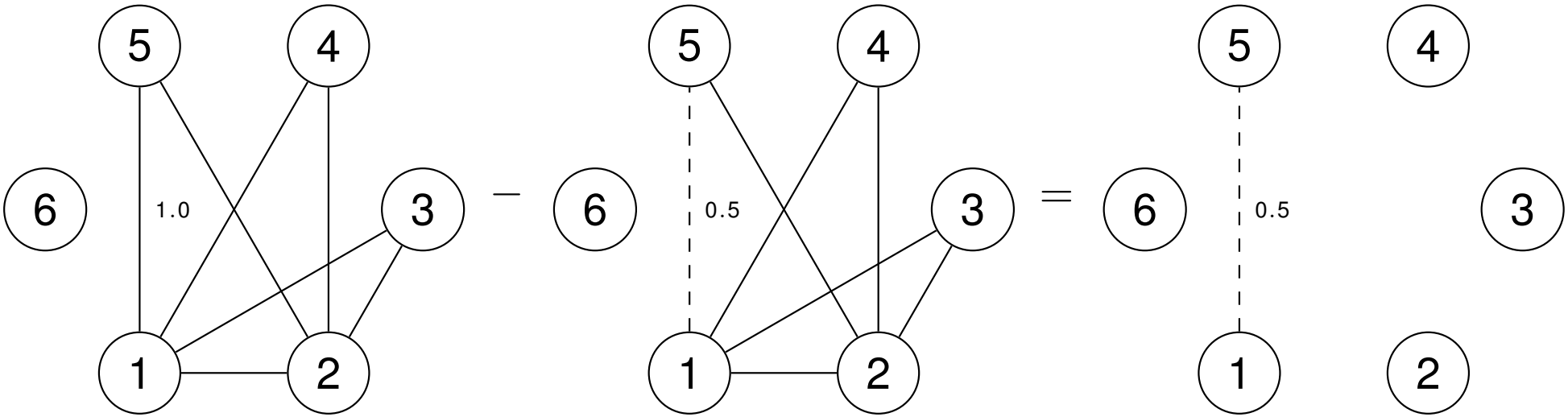
$$\hat{\Gamma} = \hat{\Gamma}(\mathbf{X}_n).$$



A plethora of literature: Banerjee et al. (2008); Friedman et al. (2007); Meinshausen and Bühlmann (2006); Ravikumar et al. (2010); Xue et al. (2012); Yang et al. (2012, 2013, 2014a,b, 2015); Yuan and Lin (2007), . . .

Two-sample problem for graphical models

Two graphs $\mathbf{\Gamma}_x$ and $\mathbf{\Gamma}_y$ for $\mathbf{X} \sim \mathbb{P}_x$ and $\mathbf{Y} \sim \mathbb{P}_y$, respectively.
We are interested in the difference, $\mathbf{\Theta} = \mathbf{\Gamma}_x - \mathbf{\Gamma}_y$.



Given two i.i.d. samples $\mathbf{X}_{n_x} = \{\mathbf{X}^{(i)} \sim \mathbb{P}_x\}_{i=1}^{n_x}$ and $\mathbf{Y}_{n_y} = \{\mathbf{Y}^{(j)} \sim \mathbb{P}_y\}_{j=1}^{n_y}$ where $\mathbb{P}_x$ and $\mathbb{P}_y \in \mathcal{P}$, learn the *difference* in conditional independence relationships

$$\hat{\mathbf{\Theta}} \approx \mathbf{\Theta} = \mathbf{\Gamma}_x - \mathbf{\Gamma}_y.$$

Some literature

Estimating the difference of Gaussian graphical models

1. Joint estimation, e.g. Danaher et al. (2014); Zhang and Wang (2012)
2. Direct difference estimation, e.g. Zhao et al. (2014)

$$\hat{\mathbf{\Theta}} = \arg\min_{\Theta} \|\mathbf{\Theta}\|_1 \quad \text{subject to} \quad \|\hat{\mathbf{\Sigma}}_x \mathbf{\Theta} \hat{\mathbf{\Sigma}}_y - (\hat{\mathbf{\Sigma}}_y - \hat{\mathbf{\Sigma}}_x)\|_\infty \leq \lambda.$$

Inference for graphical models

1. One-sample methods: Barber and Kolar (2018); Chen et al. (2014); Janková and van de Geer (2015, 2017); Lu et al. (2017); Ren et al. (2015); Wang and Kolar (2016); Yu et al. (2016).
2. Two-sample methods for Gaussian data: Xia et al. (2015) combine inference results for each precision matrix based on Ren et al. (2015).

What this talk is about

We construct an estimator that is not only consistent in high-dimensions, but also amenable to Gaussian approximation.

Gaussian approximation allows for easy uncertainty quantification and statistical inference.

In addition, our method

1. is applicable for general Markov networks, and
2. estimates the difference directly without estimating individual networks.

Formal set-up

Suppose we have $n = n_x + n_y$ observations

$$\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n_x)} \overset{\text{i.i.d.}}{\sim} f_x = f(\,\cdot\,; \gamma_x) \quad \perp\!\!\!\perp \quad \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n_y)} \overset{\text{i.i.d.}}{\sim} f_y = f(\,\cdot\,; \gamma_y),$$

for some $f_x = f(\,\cdot\,; \gamma_x)$ and $f_y = f(\,\cdot\,; \gamma_y)$ of the form

$$\begin{aligned} f(\mathbf{x}; \gamma) &\propto \exp\left(\gamma^\top \psi(\mathbf{x})\right) \\ &= \exp\left(\sum_{v=1}^{m} \gamma_v \psi_v(x_v) + \sum_{u=1}^{m} \sum_{v \geq u}^{m} \gamma_{uv} \psi_{uv}(x_u, x_v)\right) \qquad \forall\, \mathbf{x} \in \mathbb{X}. \end{aligned}$$

Goal: Given $(u, v) \in [m] \times [m]$, we want to construct a $\sqrt{n}$-consistent and asymptotically normal estimator of

$$\theta^*_{uv} = \gamma_{x,uv} - \gamma_{y,uv}$$

when $n_x, n_y \ll p$, where $p =$ number of parameters.

Kullback-Leibler importance estimation framework

For any $f = f(\cdot\,;\gamma) \in \mathcal{F}_\gamma$ and $\theta = \gamma - \gamma_y$, $f = r_\theta f_y$, where

$$r_\theta(\mathbf{x}) = \frac{f\ (\mathbf{x})}{f_y(\mathbf{x})} = \frac{Z(\gamma_y)\exp\left(\gamma^\top \psi(\mathbf{x})\right)}{Z(\gamma\ )\exp\left(\gamma_y^\top \psi(\mathbf{x})\right)} = \frac{\exp\left(\theta^\top \psi(\mathbf{x})\right)}{Z_y(\theta)}.$$

In particular, for $\theta^* = \gamma_x - \gamma_y$,

$$f_x = r_{\theta^*} f_y.$$

Thus, θ^* can be characterized as

$$\begin{aligned}\theta^* &= \operatorname*{argmin}_\theta D_{\mathrm{KL}}(f_x \| r_\theta f_y) \\ &= \operatorname*{argmin}_\theta -\mathbb{E}_x\left[\theta^\top \psi(\mathbf{x})\right] + \log \mathbb{E}_y\left[\exp\left(\theta^\top \psi(\mathbf{y})\right)\right]\end{aligned}$$

Kullback-Leibler importance estimation framework

For any $f = f(\cdot\,;\gamma) \in \mathcal{F}_\gamma$ and $\theta = \gamma - \gamma_y$, $f = r_\theta f_y$, where

$$r_\theta(\mathbf{x}) = \frac{f\ (\mathbf{x})}{f_y(\mathbf{x})} = \frac{Z(\gamma_y)\exp\left(\gamma^\top \psi(\mathbf{x})\right)}{Z(\gamma\ )\exp\left(\gamma_y^\top \psi(\mathbf{x})\right)} = \frac{\exp\left(\theta^\top \psi(\mathbf{x})\right)}{Z_y(\theta)}.$$

In particular, for $\theta^* = \gamma_x - \gamma_y$,

$$f_x = r_{\theta^*} f_y.$$

Thus, θ^* can be characterized as

$$\begin{aligned}
\theta^* &= \operatorname*{argmin}_\theta D_{\mathrm{KL}}(f_x \| r_\theta f_y) \\
&= \operatorname*{argmin}_\theta -\mathbb{E}_x\left[\theta^\top \psi(\mathbf{x})\right] + \log \mathbb{E}_y\left[\exp\left(\theta^\top \psi(\mathbf{y})\right)\right] \\
&\approx \operatorname*{argmin}_\theta -\frac{1}{n_x}\sum_{i=1}^{n_x} \theta^\top \psi(\mathbf{x}^{(i)}) + \log\left\{\frac{1}{n_y}\sum_{j=1}^{n_y} \exp\left(\theta^\top \psi(\mathbf{y}^{(j)})\right)\right\}.
\end{aligned}$$

Directly estimating the difference from the data

The *empirical KLIEP loss* is

$$\ell_{\text{KLIEP}}(\boldsymbol{\theta}; \mathbf{X}_{n_x}, \mathbf{Y}_{n_y}) = -\frac{1}{n_x}\sum_{i=1}^{n_x} \boldsymbol{\theta}^\top \boldsymbol{\psi}(\mathbf{x}^{(i)}) + \log\left\{\frac{1}{n_y}\sum_{j=1}^{n_y} \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\psi}(\mathbf{y}^{(j)})\right)\right\}.$$

When $n_y > p$,

$$\hat{\boldsymbol{\theta}}_{\text{KLIEP}} = \operatorname*{argmin}_{\theta} \ell_{\text{KLIEP}}(\boldsymbol{\theta}; \mathbf{X}_{n_x}, \mathbf{Y}_{n_y}).$$

$\hat{\boldsymbol{\theta}}_{\text{KLIEP}}$ is asymptotically normal and $\sqrt{n}$-consistent (Sugiyama et al., 2009).

When $n_y \leq p$,

$$\check{\boldsymbol{\theta}}_{\text{KLIEP}} = \operatorname*{argmin}_{\theta} \ell_{\text{KLIEP}}(\boldsymbol{\theta}; \mathbf{X}_{n_x}, \mathbf{Y}_{n_y}) + \lambda\|\boldsymbol{\theta}\|_1.$$

$\check{\boldsymbol{\theta}}_{\text{KLIEP}}$ recovers the support consistently (Liu et al., 2017).

SparKLIE+1: Sparse Kullback-Leibler Importance Estimation plus One-step Bias Correction

Inputs $\mathbf{X}_{n_x} = \{\mathbf{x}^{(i)}\}_{i=1}^{n_x}$, $\mathbf{Y}_{n_y} = \{\mathbf{y}^{(j)}\}_{j=1}^{n_Y}$, $\lambda_\theta > 0$, $\lambda_k > 0$.

Output The one-step estimate $\hat{\theta}_k^{1+}$ of the change in edge k.

Step 1. (sparse KLIEP)

$$\check{\boldsymbol{\theta}} \leftarrow \arg\min_{\boldsymbol{\theta}} \ \ell_{\text{KLIEP}}(\boldsymbol{\theta}; \mathbf{X}_{n_x}, \mathbf{Y}_{n_y}) + \lambda_\theta \|\boldsymbol{\theta}\|_1. \tag{1}$$

Bias correction with modified score

Suppose we are interested in θ_1^*, treating $\boldsymbol{\theta}_2^*$ as a nuisance parameter.

We study estimators of θ_1^* that arise as zeros of scores of the form

$$G(\theta_1; \check{\boldsymbol{\theta}}_2) = \boldsymbol{\omega}^\top \nabla \ell_{\text{KLIEP}}(\theta_1; \check{\boldsymbol{\theta}}_2), \quad \boldsymbol{\omega} \in \mathbb{R}^p,$$

where $\check{\boldsymbol{\theta}}_2$ is consistent, but not $\sqrt{n}$-consistent, estimator of $\boldsymbol{\theta}_2^*$.

Consider the first-order Taylor expansion of $G(\boldsymbol{\theta})$ about $\boldsymbol{\theta}^*$:

$$\begin{aligned}
&\boldsymbol{\omega}^\top \nabla \ell_{\text{KLIEP}}(\theta_1, \check{\boldsymbol{\theta}}_2) \\
&\qquad = \boldsymbol{\omega}^\top \nabla \ell_{\text{KLIEP}}(\boldsymbol{\theta}^*) + \boldsymbol{\omega}^\top \nabla^2 \ell_{\text{KLIEP}}(\boldsymbol{\theta}^*) \begin{bmatrix} \theta_1 - \boldsymbol{\theta}_1^* \\ \check{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^* \end{bmatrix} + o_{\mathbb{P}}(n^{-1/2}).
\end{aligned}$$

Bias correction with modified score

Naïve approach: $\omega = \mathbf{e}_1 \rightsquigarrow \hat{\theta}_1^{\mathsf{n}}$

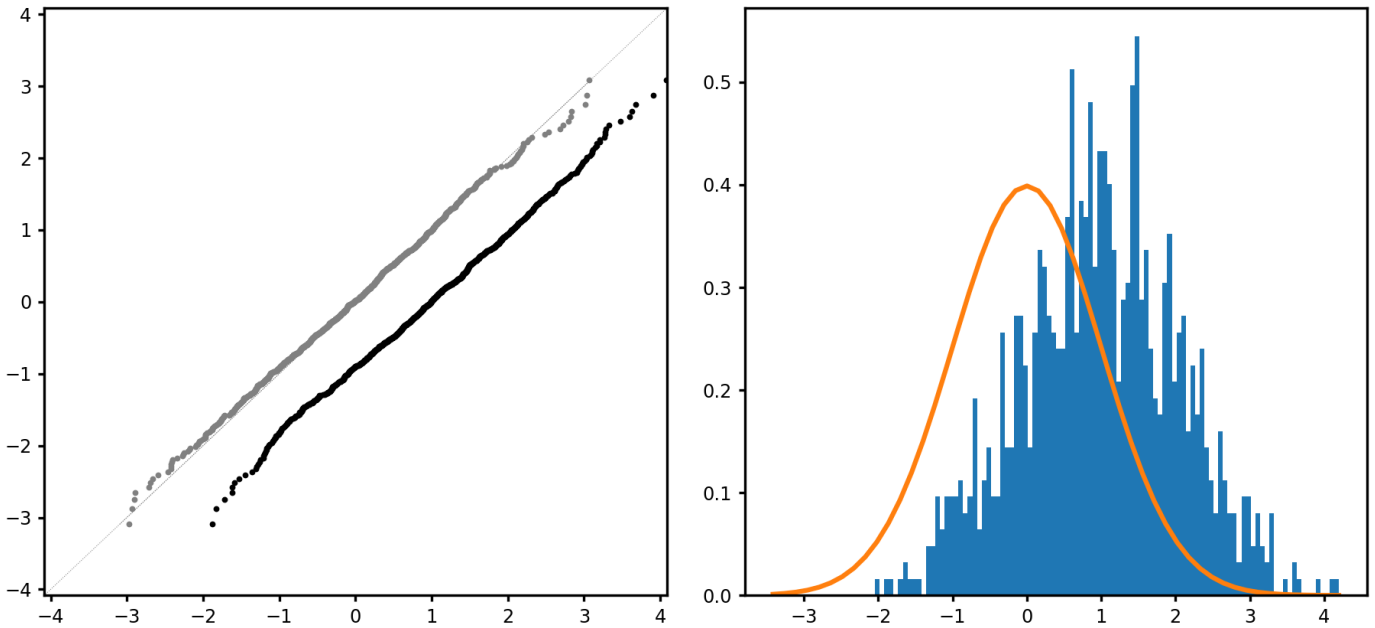
$$
\begin{aligned}
0 &\equiv \nabla_1 \ell_{\mathsf{KLIEP}}(\hat{\theta}_1^{\mathsf{n}}, \check{\boldsymbol{\theta}}_2) \\
&= \nabla_1 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*) + \begin{bmatrix} \nabla_{11}^2 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*) \\ \nabla_{21}^2 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*) \end{bmatrix}^\top \begin{bmatrix} \hat{\theta}_1^{\mathsf{n}} - \boldsymbol{\theta}_1^* \\ \check{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^* \end{bmatrix} + o_{\mathbb{P}}(n^{-1/2}).
\end{aligned}
$$

Bias correction with modified score

Naïve approach: $\omega = \mathbf{e}_1 \rightsquigarrow \hat{\theta}_1^{\mathsf{n}}$

$$
\begin{aligned}
0 &\equiv \nabla_1 \ell_{\mathsf{KLIEP}}(\hat{\theta}_1^{\mathsf{n}}, \check{\boldsymbol{\theta}}_2) \\
&= \nabla_1 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*) + \begin{bmatrix} \nabla_{11}^2 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*) \\ \nabla_{21}^2 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*) \end{bmatrix}^\top \begin{bmatrix} \hat{\theta}_1^{\mathsf{n}} - \boldsymbol{\theta}_1^* \\ \check{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^* \end{bmatrix} + o_{\mathbb{P}}(n^{-1/2}).
\end{aligned}
$$

BAD!!!



Bias correction with modified score

Naïve approach: $\omega = \mathbf{e}_1 \rightsquigarrow \hat{\theta}_1^{\mathsf{n}}$

$$
\begin{aligned}
0 &\equiv \nabla_1 \ell_{\mathsf{KLIEP}}(\hat{\theta}_1^{\mathsf{n}}, \check{\boldsymbol{\theta}}_2) \\
&= \nabla_1 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*) + \begin{bmatrix} \nabla_{11}^2 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*) \\ \nabla_{21}^2 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*) \end{bmatrix}^\top \begin{bmatrix} \hat{\theta}_1^{\mathsf{n}} - \boldsymbol{\theta}_1^* \\ \check{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^* \end{bmatrix} + o_{\mathbb{P}}(n^{-1/2}).
\end{aligned}
$$

BAD!!!

- $\nabla_{21}^2 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*)$ is typically nonzero,
- $\check{\boldsymbol{\theta}}_2$ is not $\sqrt{n}$-consistent,
- $\implies \sqrt{n}\, \nabla_{21}^2 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*)^\top (\check{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^*)$ is typically non-vanishing.

Bias correction with modified score

Naïve approach: $\omega = \mathbf{e}_1 \rightsquigarrow \hat{\theta}_1^{\mathsf{n}}$

$$
\begin{aligned}
0 &\equiv \nabla_1 \ell_{\mathsf{KLIEP}}(\hat{\theta}_1^{\mathsf{n}}, \check{\boldsymbol{\theta}}_2) \\
&= \nabla_1 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*) + \begin{bmatrix} \nabla_{11}^2 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*) \\ \nabla_{21}^2 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*) \end{bmatrix}^\top \begin{bmatrix} \hat{\theta}_1^{\mathsf{n}} - \boldsymbol{\theta}_1^* \\ \check{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^* \end{bmatrix} + o_{\mathbb{P}}(n^{-1/2}).
\end{aligned}
$$

BAD!!!

- $\nabla_{21}^2 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*)$ is typically nonzero,
- $\check{\boldsymbol{\theta}}_2$ is not $\sqrt{n}$-consistent,
- $\implies \sqrt{n} \nabla_{21}^2 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*)^\top (\check{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^*)$ is typically non-vanishing.

Solution: Force $(\nabla^2 \ell_{\mathsf{KLIEP}}(\boldsymbol{\theta}^*)\omega)_2$ to be (approximately) zero!

Bias correction with modified score

Smart approach: ω s.t. $\nabla^2 \ell_{\text{KLIEP}}(\boldsymbol{\theta}^*)\omega = \mathbf{e}_1 \rightsquigarrow \hat{\theta}_1^{\text{s}}$

$$
\begin{aligned}
0 &\equiv \omega^\top \nabla \ell_{\text{KLIEP}}(\hat{\theta}_1^{\text{s}}, \check{\boldsymbol{\theta}}_2) \\
&= \omega^\top \nabla \ell_{\text{KLIEP}}(\boldsymbol{\theta}^*) + \mathbf{e}_1^\top \begin{bmatrix} \hat{\theta}_1^{\text{s}} - \boldsymbol{\theta}_1^* \\ \check{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^* \end{bmatrix} + o_{\mathbb{P}}(n^{-1/2}),
\end{aligned}
$$

so that

$$
\sqrt{n}(\hat{\theta}_1^{\text{s}} - \theta_1^*) = -\sqrt{n}\omega^\top \nabla \ell_{\text{KLIEP}}(\boldsymbol{\theta}^*) + o_{\mathbb{P}}(n^{-1/2}).
$$

Moreover, $\omega^\top \nabla \ell_{\text{KLIEP}}(\boldsymbol{\theta}^*)$ is a difference of two sums of i.i.d. mean-zero random variable, so is well-approximated by a Gaussian distribution.

Bias correction with modified score

Smart approach: ω s.t. $\nabla^2 \ell_{\text{KLIEP}}(\boldsymbol{\theta}^*)\omega \approx \mathbf{e}_1 \;\rightsquigarrow\; \hat{\theta}_1^{\text{s}}$

$$
\begin{aligned}
0 &\equiv \omega^\top \nabla \ell_{\text{KLIEP}}(\hat{\theta}_1^{\text{s}}, \check{\boldsymbol{\theta}}_2) \\
&= \omega^\top \nabla \ell_{\text{KLIEP}}(\boldsymbol{\theta}^*) + \mathbf{e}_1^\top \begin{bmatrix} \hat{\theta}_1^{\text{s}} - \boldsymbol{\theta}_1^* \\ \check{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^* \end{bmatrix} + o_{\mathbb{P}}(n^{-1/2}),
\end{aligned}
$$

provided

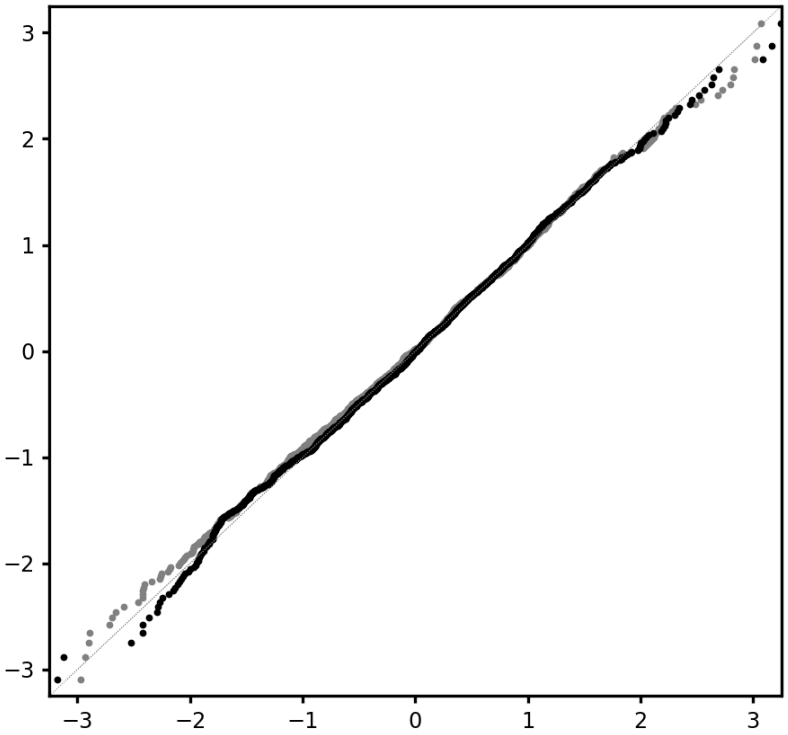
$$
\left\{ \nabla^2 \ell_{\text{KLIEP}}(\boldsymbol{\theta}^*)\omega - \mathbf{e}_1 \right\}^\top \begin{bmatrix} \hat{\theta}_1^{\text{s}} - \boldsymbol{\theta}_1^* \\ \check{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^* \end{bmatrix} = o_{\mathbb{P}}(n^{-1/2}),
$$

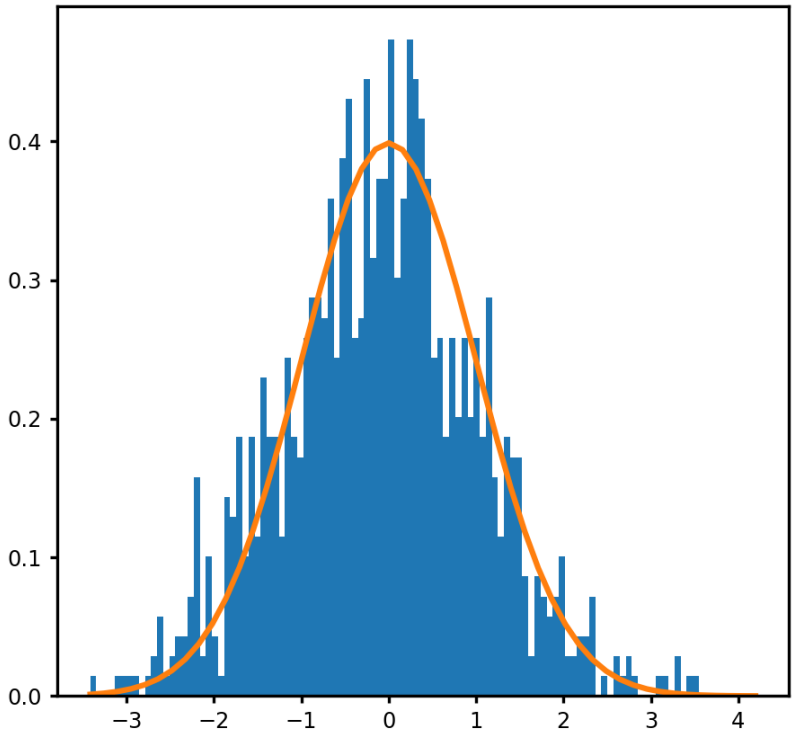
so that

$$
\sqrt{n}(\hat{\theta}_1^{\text{s}} - \theta_1^*) = -\sqrt{n}\,\omega^\top \nabla \ell_{\text{KLIEP}}(\boldsymbol{\theta}^*) + o_{\mathbb{P}}(n^{-1/2}).
$$

Moreover, $\omega^\top \nabla \ell_{\text{KLIEP}}(\boldsymbol{\theta}^*)$ is a difference of two sums of i.i.d. mean-zero random variable, so is well-approximated by a Gaussian distribution.

Bias correction with modified score





Bias correction with modified score

Smart approach: ω s.t. $\nabla^2 \ell_{\text{KLIEP}}(\theta^*)\omega \approx \mathbf{e}_1 \rightsquigarrow \hat{\theta}_1^{\text{s}}$

$$
\begin{aligned}
0 &\equiv \omega^\top \nabla \ell_{\text{KLIEP}}(\hat{\theta}_1^{\text{s}}, \check{\theta}_2) \\
&= \omega^\top \nabla \ell_{\text{KLIEP}}(\theta^*) + \mathbf{e}_1^\top \begin{bmatrix} \hat{\theta}_1^{\text{s}} - \theta_1^* \\ \check{\theta}_2 - \theta_2^* \end{bmatrix} + o_{\mathbb{P}}(n^{-1/2}),
\end{aligned}
$$

provided

$$
\left\{\nabla^2 \ell_{\text{KLIEP}}(\theta^*)\omega - \mathbf{e}_1\right\}^\top \begin{bmatrix} \hat{\theta}_1^{\text{s}} - \theta_1^* \\ \check{\theta}_2 - \theta_2^* \end{bmatrix} = o_{\mathbb{P}}(n^{-1/2}),
$$

so that

$$
\sqrt{n}(\hat{\theta}_1^{\text{s}} - \theta_1^*) = -\sqrt{n}\,\omega^\top \nabla \ell_{\text{KLIEP}}(\theta^*) + o_{\mathbb{P}}(n^{-1/2}).
$$

Caveat:

- ω must be approximated from the data.
- $\omega \approx \Sigma_\psi^{-1}\mathbf{e}_1$, where $\Sigma_\psi = \mathbb{E}\nabla^2 \ell_{\text{KLIEP}}(\theta^*)$.

SparKLIE+1: Sparse Kullback-Leibler Importance Estimation plus One-step Bias Correction

Inputs $\mathbf{X}_{n_x} = \{\mathbf{x}^{(i)}\}_{i=1}^{n_x}$, $\mathbf{Y}_{n_y} = \{\mathbf{y}^{(j)}\}_{j=1}^{n_Y}$, $\lambda_\theta > 0$, $\lambda_k > 0$.

Output The one-step estimate $\hat{\theta}_k^{1+}$ of the change in edge k.

Step 1. (sparse KLIEP)

$$\check{\boldsymbol{\theta}} \leftarrow \arg\min_{\boldsymbol{\theta}} \ \ell_{\text{KLIEP}}(\boldsymbol{\theta}; \mathbf{X}_{n_x}, \mathbf{Y}_{n_y}) + \lambda_\theta \|\boldsymbol{\theta}\|_1. \tag{1}$$

SparKLIE+1: Sparse Kullback-Leibler Importance Estimation plus One-step Bias Correction

Inputs $\mathbf{X}_{n_x} = \{\mathbf{x}^{(i)}\}_{i=1}^{n_x}$, $\mathbf{Y}_{n_y} = \{\mathbf{y}^{(j)}\}_{j=1}^{n_Y}$, $\lambda_\theta > 0$, $\lambda_k > 0$.

Output The one-step estimate $\hat{\theta}_k^{1+}$ of the change in edge k.

Step 1. (sparse KLIEP)

$$\check{\boldsymbol{\theta}} \leftarrow \arg\min_{\boldsymbol{\theta}} \ \ell_{\text{KLIEP}}(\boldsymbol{\theta}; \mathbf{X}_{n_x}, \mathbf{Y}_{n_y}) + \lambda_\theta \|\boldsymbol{\theta}\|_1. \tag{1}$$

Step 2. (sparse Hessian inversion)

$$\check{\boldsymbol{\omega}}_k \leftarrow \arg\min_{\boldsymbol{\omega}} \ \tfrac{1}{2}\boldsymbol{\omega}^\top \nabla^2 \ell_{\text{KLIEP}}(\check{\boldsymbol{\theta}})\boldsymbol{\omega} - \boldsymbol{\omega}^\top \mathbf{e}_k + \lambda_k \|\boldsymbol{\omega}\|_1. \tag{2}$$

SparKLIE+1: Sparse Kullback-Leibler Importance Estimation plus One-step Bias Correction

Inputs $\mathbf{X}_{n_x} = \{\mathbf{x}^{(i)}\}_{i=1}^{n_x}$, $\mathbf{Y}_{n_y} = \{\mathbf{y}^{(j)}\}_{j=1}^{n_Y}$, $\lambda_\theta > 0$, $\lambda_k > 0$.

Output The one-step estimate $\hat{\theta}_k^{1+}$ of the change in edge k.

Step 1. (sparse KLIEP)

$$\check{\boldsymbol{\theta}} \leftarrow \arg\min_{\boldsymbol{\theta}} \ \ell_{\text{KLIEP}}(\boldsymbol{\theta}; \mathbf{X}_{n_x}, \mathbf{Y}_{n_y}) + \lambda_\theta \|\boldsymbol{\theta}\|_1. \tag{1}$$

Step 2. (sparse Hessian inversion)

$$\check{\boldsymbol{\omega}}_k \leftarrow \arg\min_{\boldsymbol{\omega}} \ \tfrac{1}{2}\boldsymbol{\omega}^\top \nabla^2 \ell_{\text{KLIEP}}(\check{\boldsymbol{\theta}})\boldsymbol{\omega} - \boldsymbol{\omega}^\top \mathbf{e}_k + \lambda_k \|\boldsymbol{\omega}\|_1. \tag{2}$$

Step 3. (one-step correction)

$$\hat{\theta}_k \leftarrow \check{\theta}_k - \check{\boldsymbol{\omega}}_k^\top \nabla \ell_{\text{KLIEP}}(\check{\boldsymbol{\theta}}).$$

Conditions

Condition 1 (bounded density ratio model).

There exist $\varrho > 0$ such that

$$M_r^{-1} \leq r_\theta(\mathbf{y}) \leq M_r \text{ a.s. for all } \boldsymbol{\theta} \text{ with } \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 \leq \varrho$$

for some $M_r = M_r(\varrho) \geq 1$.

Condition 2 (bounded population eigenvalues).

There exist $0 < \underline{\kappa} \leq \bar{\kappa} < \infty$ such that

$$\underline{\kappa} \leq \min_{\|\mathbf{v}\| \leq 1, \mathbf{v} \neq \mathbf{0}} \mathbf{v}^\top \boldsymbol{\Sigma}_\psi \mathbf{v} \leq \max_{\|\mathbf{v}\| \leq 1, \mathbf{v} \neq \mathbf{0}} \mathbf{v}^\top \boldsymbol{\Sigma}_\psi \mathbf{v} \leq \bar{\kappa},$$

$$\underline{\kappa} \leq \min_{\|\mathbf{v}\| \leq 1, \mathbf{v} \neq \mathbf{0}} \mathbf{v}^\top \boldsymbol{\Sigma}_{\psi r} \mathbf{v} \leq \max_{\|\mathbf{v}\| \leq 1, \mathbf{v} \neq \mathbf{0}} \mathbf{v}^\top \boldsymbol{\Sigma}_{\psi r} \mathbf{v} \leq \bar{\kappa}.$$

Theoretical guarantee

Let $s_\theta = \|\theta^*\|_0$, $s_k = \|\omega_k^*\|_0$, and $s = s_\theta \vee s_k$.

Theorem 1 (Corollary 2 in Kim et al. (2019)).

Let $\varepsilon_{RSC,n} \searrow 0$. Under Conditions 1 and 2, provided that

$$n_y \gtrsim s \log^2(s) \log(p \vee n_y) \log(n_y) / \varepsilon_{RSC,n}^2,$$

the SparKLIE+1 estimator $\hat{\theta}_k$ with the regularization parameters set as

$$\lambda_\theta \asymp (\log p / n)^{1/2} \quad \text{and} \quad \lambda_k \asymp (s_k \log p / n)^{1/2}$$

obeys the Gaussian approximation bound

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left\{ \sqrt{n}(\hat{\theta}_k - \theta_k^*)/\hat{\sigma}_k \le t \right\} - \Phi(t) \right| \le O\left(\frac{s^{7/2} \log p}{\sqrt{n}} \right) + \varepsilon_{RSC,n} + c \exp\left(-c' \log p\right).$$

Low-dim comparisons via Theorem 1

Let z_α be the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$.

Approximate $100 \times (1-\alpha)$ **% confidence interval:**

$$\hat{\theta}_k \pm z_\alpha \times \hat{\sigma}_k/\sqrt{n}.$$

Level-α test for change in edge k:

$$\mathcal{H}_0 : \gamma_{x,k} = \gamma_{y,k} \quad \Longleftrightarrow \quad \mathcal{H}_0 : \theta_k^* = 0$$

Reject $\mathcal{H}_0$ whenever

$$|Z_{k,n_x,n_y}| = |\sqrt{n}\,\hat{\theta}_k/\hat{\sigma}_k| > z_\alpha.$$

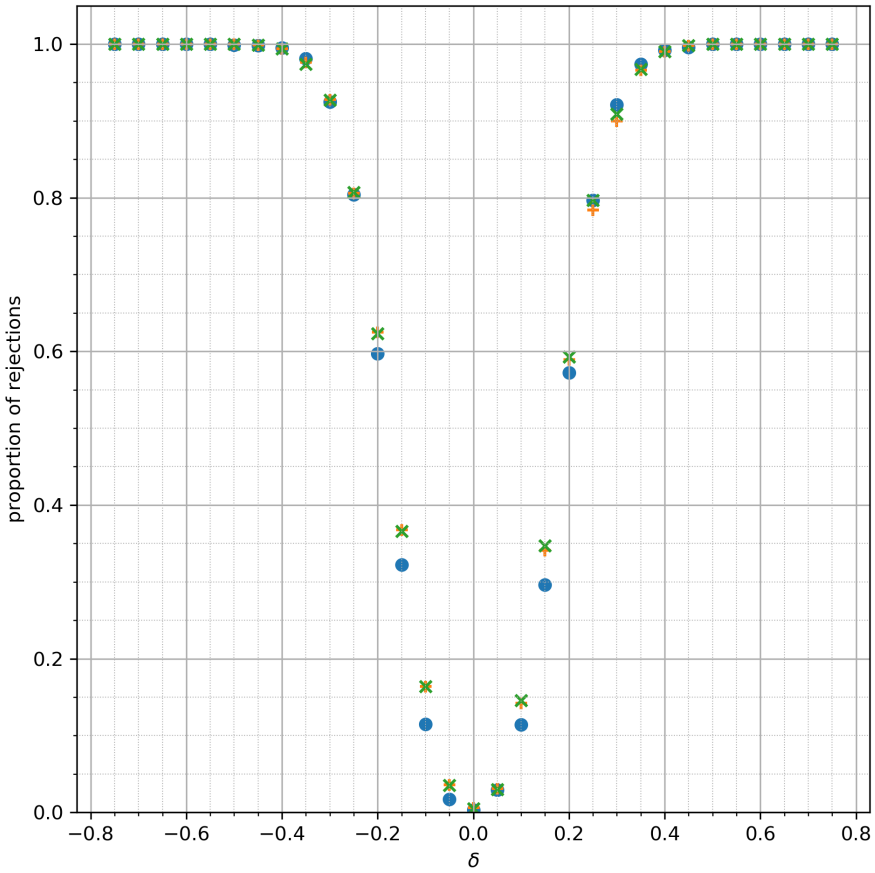
Checking coverage

Table: Proportion of successful coverage for 95% CIs

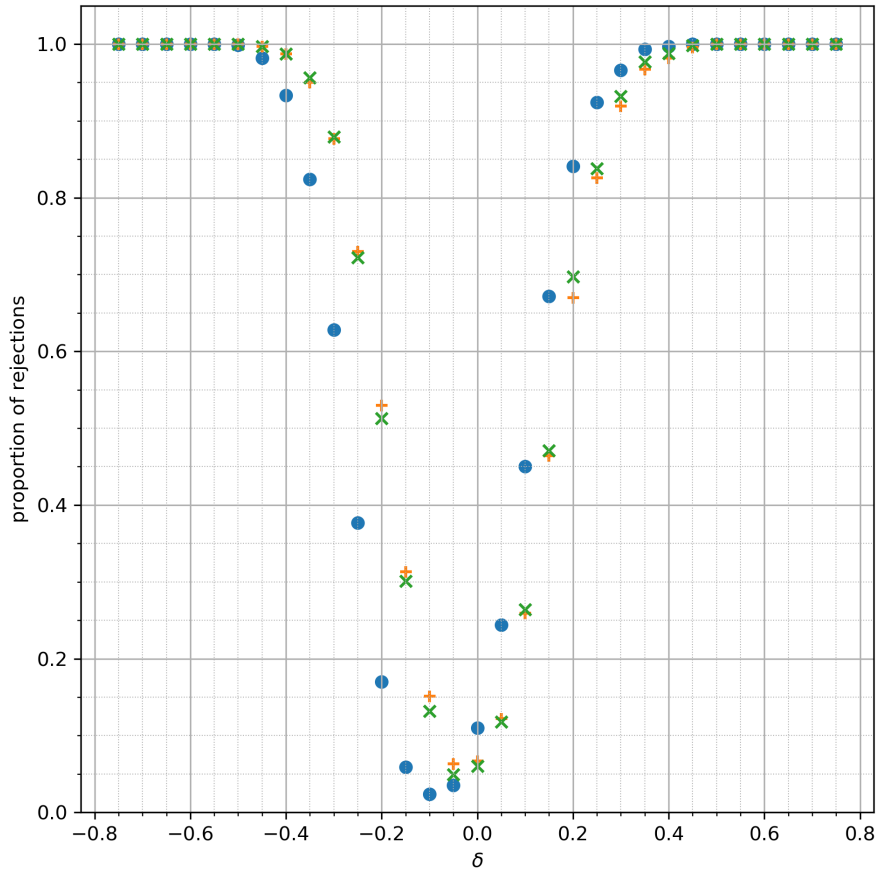
x-graph	**y**-graph	m	n_x	n_y	$\hat{\theta}^{\text{naïve}}$	$\hat{\theta}^{1+}$	$\hat{\theta}^{2+}$	$\hat{\theta}^{\text{oracle}}$
chain	(1)	25	150	300	0.882	0.952	0.952	0.952
		50	300	600	0.836	0.959	0.950	0.955
	(2)	25	150	300	0.888	0.934	0.942	0.944
		50	300	600	0.817	0.920	0.939	0.936
3-ary tree	(1)	25	150	300	0.893	0.956	0.941	0.940
		50	300	600	0.865	0.963	0.956	0.945
	(2)	25	150	300	0.919	0.963	0.941	0.951
		50	300	600	0.858	0.951	0.936	0.942

Looking at power

Figure: ● - naive, + - SparKLIE+1, x - SparKLIE+2



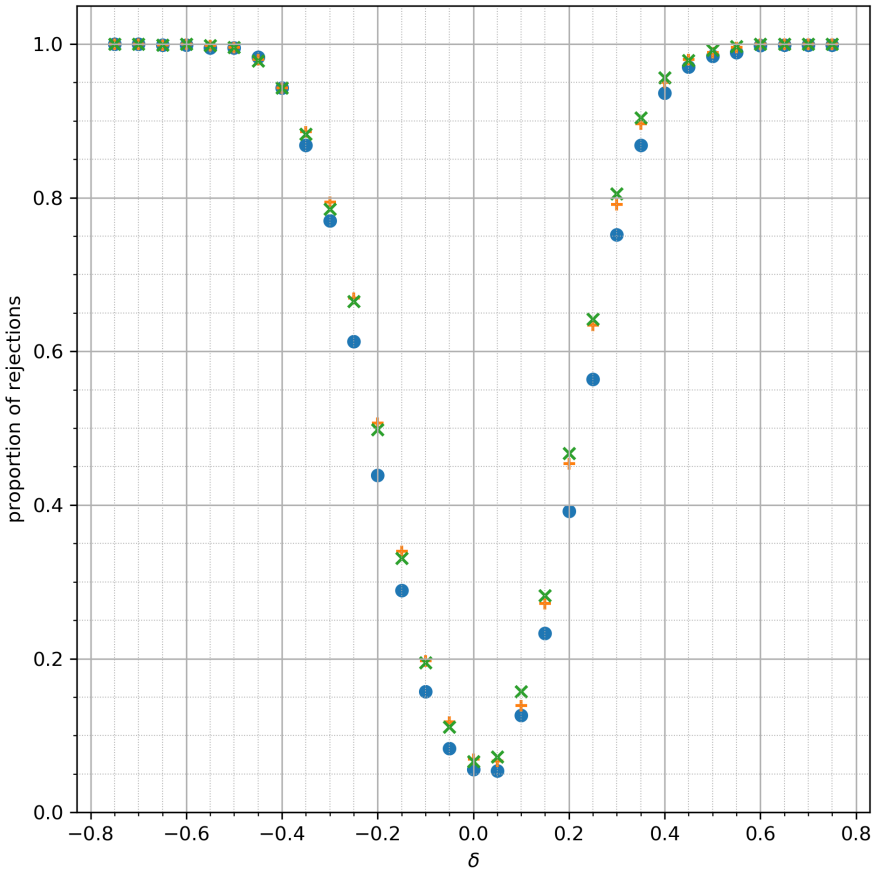
(a) none



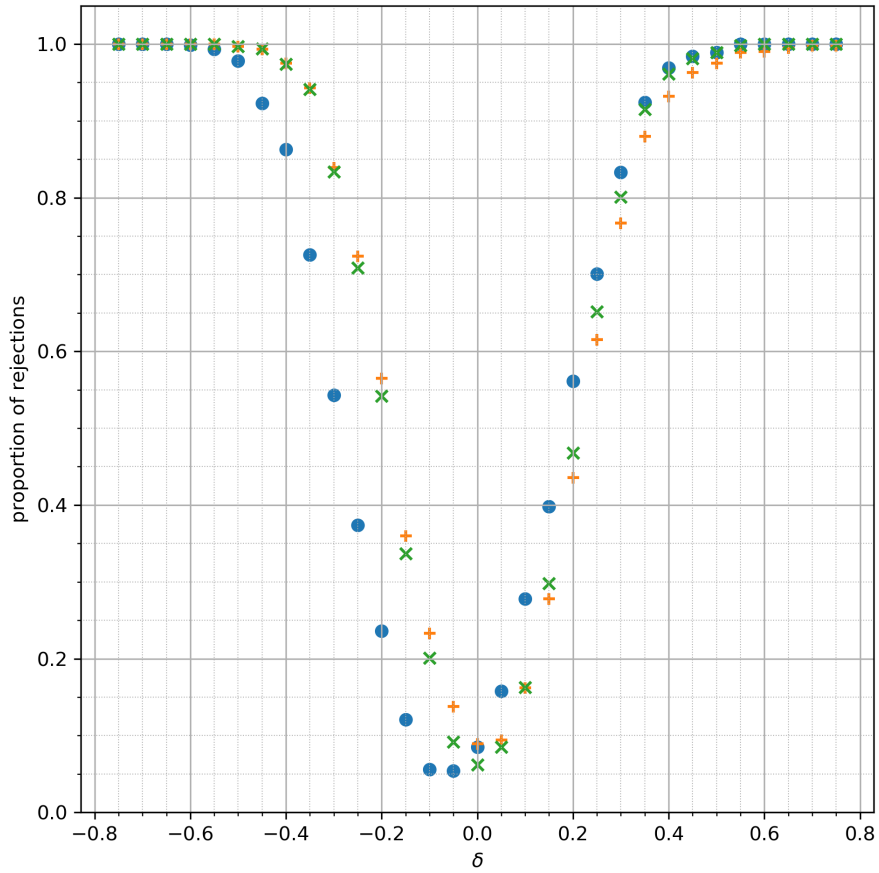
(b) weak

Looking at power

Figure: ● - naive, + - SparKLIE+1, x - SparKLIE+2



(a) strong



(b) mixed

A word about high-dim comparisons...

Let $\mathcal{I} \subseteq [p]$ be the indices of where you want to compare graphs.

- Previously, $|\mathcal{I}| = 1$.
- When $|\mathcal{I}| \ll p$, one can use SparKLIE+ estimators with multiple comparison correction.
- What about when $|\mathcal{I}| \approx p$?

Example (equal graph test):

$$\mathcal{H}_{\text{global}} : \gamma_x = \gamma_y \quad \Longleftrightarrow \quad \mathcal{H}_{\text{global}} : \boldsymbol{\theta}^* = \mathbf{0}.$$

Solution: Use max-type statistics

$$T_{n_x,n_y,\mathcal{I}} = \max_{k \in \mathcal{I}} \sqrt{n}\,|\hat{\theta}_k - \theta_k^*|.$$

Accurate estimates of the the quantiles of $T_{n_x,n_y,\mathcal{I}}$ can be obtained via bootstrap methods (empirical or multiplier).

High-dimensional Gaussian approximation intuition

$\sqrt{n}(\hat{\theta} - \theta^*)$ is approximately linear:

$$\sqrt{n}(\hat{\theta} - \theta^*) = L + R,$$

where

$$L = -\frac{1}{\sqrt{n}} \mathbf{\Omega}^{*\top} \left\{ \frac{n}{n_x} \sum_{i=1}^{n_x} \left(\psi(\mathbf{x}^{(i)}) - \boldsymbol{\mu}_\psi \right) - \frac{n}{n_y} \sum_{j=1}^{n_y} \left(\psi(\mathbf{y}^{(j)}) r_{\theta^*}(\mathbf{y}^{(j)}) - \boldsymbol{\mu}_\psi \right) \right\},$$

and $R = \sqrt{n}(\hat{\theta} - \theta^*) - L$.

Letting $\mathbf{\Sigma} = \mathrm{Cov}\, L$, and $\hat{\mathbf{\Sigma}}$, a consistent estimator of $\mathbf{\Sigma}$, consider the chain of approximations:

$$\sqrt{n}(\hat{\theta} - \theta^*) \overset{\mathcal{L}}{\approx} L \overset{\mathcal{L}}{\approx} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \overset{\mathcal{L}}{\approx} \mathcal{N}(\mathbf{0}, \hat{\mathbf{\Sigma}}).$$

The quantiles of $T = \max_k \sqrt{n}|\hat{\theta}_k - \theta^*_k|$ can be estimated by the quantiles of $\hat{T} = \max_k |\mathcal{N}(\mathbf{0}, \hat{\mathbf{\Sigma}})|$ as long as $\log^7 p = o(n)$ (Chernozhukov et al. 2013).

Sketching quantiles with Gaussian multiplier bootstrap

Inputs The data, consistent estimates $\check{\Omega}$ and $\hat{\theta}$.

Output An estimate $\hat{c}_{T,\alpha}$ of the $(1-\alpha)$-quantile of T.

Step 1. For $b = 1, \ldots, n_b$,
Draw $\xi_x^{(b,1)}, \ldots, \xi_x^{(b,n_x)}, \xi_y^{(b,1)}, \ldots, \xi_y^{(b,n_y)} \overset{iid}{\sim} \mathcal{N}(0,1)$.
Compute

$$\hat{T}^{(b)} = \max_k \frac{1}{\sqrt{n}} \left| \left\langle \check{\omega}_k, \frac{n}{n_x} \sum_{i=1}^{n_x} \left(\psi(\mathbf{x}^{(i)}) - \overline{\psi} \right) \xi_x^{(b,i)} - \frac{n}{n_y} \sum_{j=1}^{n_y} \left(\psi(\mathbf{y}^{(j)}) \hat{r}_{\hat{\theta}}(\mathbf{y}^{(j)}) - \hat{\mu}(\hat{\theta}) \right) \xi_y^{(b,j)} \right\rangle \right|.$$

Step 2. Find the $(1-\alpha)$-quantile as

$$\hat{c}_{T,\alpha} \leftarrow \lfloor (1-\alpha) n_b \rfloor \text{-th order statistic of } \{ \hat{T}^{(b)} : b = 1, \ldots, n_b \}.$$

Theoretical guarantee

Let $s_\theta = \|\theta^*\|_0$, $s_k = \|\omega_k^*\|_0$, $k = 1, \ldots, p$, and $s = \max\{s_\theta, s_k : k = 1, \ldots, p\}$.

Theorem 2 (Corollary 4.5 in Kim et al. (2019)).

Let $\varepsilon_{RSC,n} \searrow 0$. Under Conditions 1 and 2, provided that

$$n_y \gtrsim s \log^2(s) \log(p \vee n_y) \log(n_y) / \varepsilon_{RSC,n}^2,$$

T_{n_x,n_y} constructed with the regularization parameters set as

$$\lambda_\theta \asymp (\log p/n)^{1/2} \quad \text{and} \quad \lambda_k \asymp (s_k \log p/n)^{1/2}, \quad k = 1, \ldots, p,$$

and the Gaussian multiplier bootstrap sketched quantiles satisfy

$$\sup_{\alpha \in (0,1)} \left| \mathbb{P}\left\{ T_{n_x,n_y} \leq \hat{c}_{T,\alpha} \right\} - (1-\alpha) \right| = O\left(\left(\frac{s^{21} \log^7(pn)}{n} \right)^{1/6} + \varepsilon_{RSC,n} + c \exp\left(-c' \log p\right) \right)$$

with probability at least $1 - \varepsilon_{RSC,n} - c \exp(-c' \log p) - n^{-1}$.

Sketching quantiles with empirical bootstrap

Inputs $\mathbf{X}_{n_x} = \{\mathbf{x}^{(i)}\}_{i=1}^{n_x}$, $\mathbf{Y}_{n_y} = \{\mathbf{y}^{(j)}\}_{j=1}^{n_Y}$, $\lambda_\theta > 0$, $\hat{\boldsymbol{\theta}}$, $\check{\boldsymbol{\Omega}}$

Output An estimate $\hat{c}_{T,\alpha}$ of the $(1-\alpha)$-quantile of T.

Step 1. For $b = 1, \ldots, n_b$,
Draw $\mathbf{X}_{n_x}^{(b)} = \{\mathbf{x}^{(1,b)}, \ldots, \mathbf{x}^{(n_x,b)}\}$ and $\mathbf{Y}_{n_y}^{(b)} = \{\mathbf{y}^{(1,b)}, \ldots, \mathbf{y}^{(n_y,b)}\}$.
Compute

$$\check{\boldsymbol{\theta}}^{(b)} \leftarrow \text{minimize} \quad \ell_{\text{KLIEP}}(\boldsymbol{\theta}; \mathbf{X}_{n_x}^{(b)}, \mathbf{Y}_{n_y}^{(b)}) + \lambda_\theta \|\boldsymbol{\theta}\|_1$$

$$\hat{\boldsymbol{\theta}}^{(b)} \leftarrow \check{\boldsymbol{\theta}}^{(b)} - \check{\boldsymbol{\Omega}}^\top \nabla \ell_{\text{KLIEP}}(\check{\boldsymbol{\theta}}^{(b)}; \mathbf{X}_{n_x}^{(b)}, \mathbf{Y}_{n_y}^{(b)}) - \check{\boldsymbol{\Omega}}^\top \nabla \ell_{\text{KLIEP}}(\hat{\boldsymbol{\theta}}; \mathbf{X}_{n_x}, \mathbf{Y}_{n_y}).$$

Compute

$$\hat{T}^{(b)} \leftarrow \max_k \sqrt{n} \left| \hat{\theta}_k^{(b)} - \hat{\theta}_k \right|$$

Step 2. Find the $(1-\alpha)$-quantile as

$$\hat{c}_{T,\alpha} \leftarrow \lfloor (1-\alpha) n_b \rfloor \text{-th order statistic of } \{\hat{T}^{(b)} : b = 1, \ldots, n_b\}$$

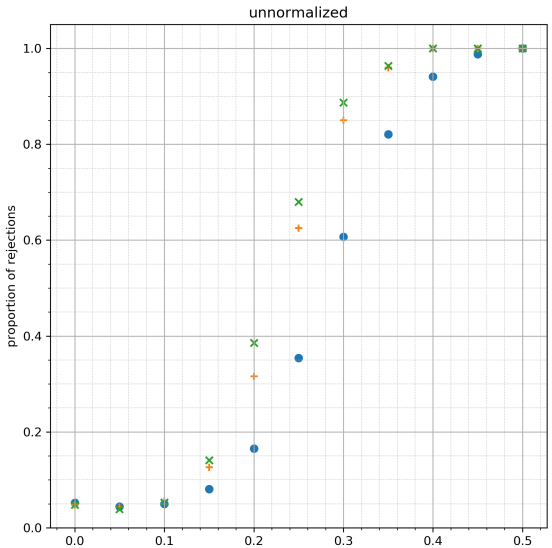
Checking size

Table: Type I error rate using empirical bootstrap sketched quantiles

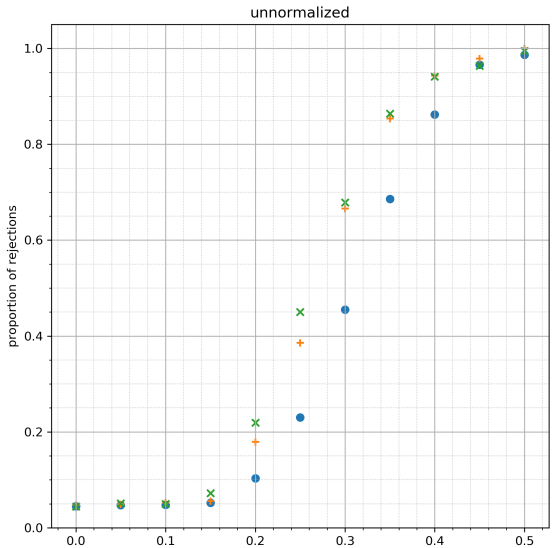
m	edge type	α					
		$\hat{c}_{T,\alpha}$			$\hat{c}_{W,\alpha}$		
		0.10	0.05	0.01	0.10	0.05	0.01
25	positive	0.093	0.049	0.015	0.100	0.051	0.014
	mixed	0.092	0.040	0.007	0.105	0.044	0.009
	negative	0.096	0.041	0.006	0.087	0.043	0.009
50	positive	0.105	0.054	0.009	0.086	0.051	0.012
	mixed	0.091	0.043	0.008	0.100	0.054	0.018
	negative	0.097	0.053	0.013	0.103	0.045	0.015
100	positive	0.092	0.042	0.011	0.105	0.046	0.009
	mixed	0.096	0.045	0.006	0.098	0.051	0.012
	negative	0.094	0.048	0.006	0.103	0.044	0.010

Looking at power

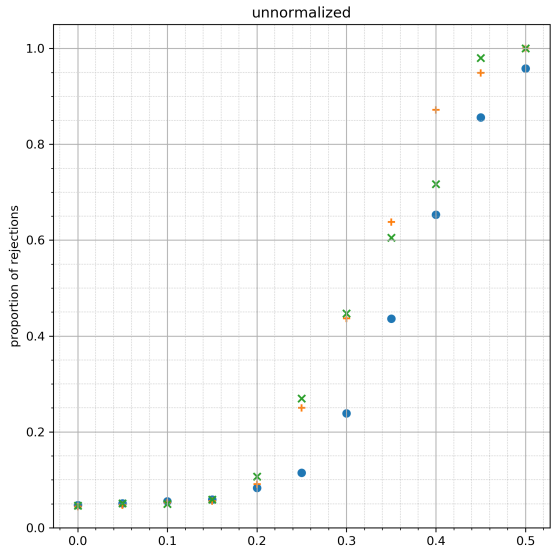
Figure: • - 1 change; + - 3 changes; x - 5 changes



(a) 25 nodes



(b) 50 nodes



(c) 100 nodes

Summary

SparKLIE+

- estimates each edge in the difference graph directly without estimating either graph,

Summary

SparKLIE+

- estimates each edge in the difference graph directly without estimating either graph,
- can be used to compare any two Markov networks with pairwise dependencies,

Summary

SparKLIE+

- estimates each edge in the difference graph directly without estimating either graph,
- can be used to compare any two Markov networks with pairwise dependencies,
- can be used with normal-theory-based statistical inference,

Summary

SparKLIE+

- estimates each edge in the difference graph directly without estimating either graph,
- can be used to compare any two Markov networks with pairwise dependencies,
- can be used with normal-theory-based statistical inference,
- with bootstrap, the max-type statistics can be used for high-dimensional comparisons with the desired error control.

Summary

SparKLIE+

- estimates each edge in the difference graph directly without estimating either graph,
- can be used to compare any two Markov networks with pairwise dependencies,
- can be used with normal-theory-based statistical inference,
- with bootstrap, the max-type statistics can be used for high-dimensional comparisons with the desired error control.

In terms of theory, our method

- puts the burden of the structural assumption on the *difference*,

Summary

SparKLIE+

- estimates each edge in the difference graph directly without estimating either graph,
- can be used to compare any two Markov networks with pairwise dependencies,
- can be used with normal-theory-based statistical inference,
- with bootstrap, the max-type statistics can be used for high-dimensional comparisons with the desired error control.

In terms of theory, our method

- puts the burden of the structural assumption on the *difference*,
- works under a weaker set of assumptions that the sparse KLIEP.

Summary

SparKLIE+

- estimates each edge in the difference graph directly without estimating either graph,
- can be used to compare any two Markov networks with pairwise dependencies,
- can be used with normal-theory-based statistical inference,
- with bootstrap, the max-type statistics can be used for high-dimensional comparisons with the desired error control.

In terms of theory, our method

- puts the burden of the structural assumption on the *difference*,
- works under a weaker set of assumptions that the sparse KLIEP.

Additional details may be found in

Byol Kim, Song Liu, and Mladen Kolar, *Two-sample inference for high-dimensional Markov networks*, arXiv e-prints (2019), arXiv:1905.00466.

Thank you!