Nonparametric Graph Estimation

Han Liu

Department of Operations Research and Financial Engineering
Princeton University

Acknowledgement

Fang Han
JHU Biostats

John Lafferty
Chicago CS/Stats

Larry Wasserman
CMU Stats/ML

Tuo Zhao
JHU CS

http:// www.princeton.edu/~hanliu

High Dimensional Data Analysis

The dimensionality d increases with the sample size n

Approximation Error + $\boxed{\text{Estimation Error} + \text{Computing Error}}$

This talk

Well studied under linear and Gaussian models

A little nonparametricity goes a long way

Graph Estimation Problem

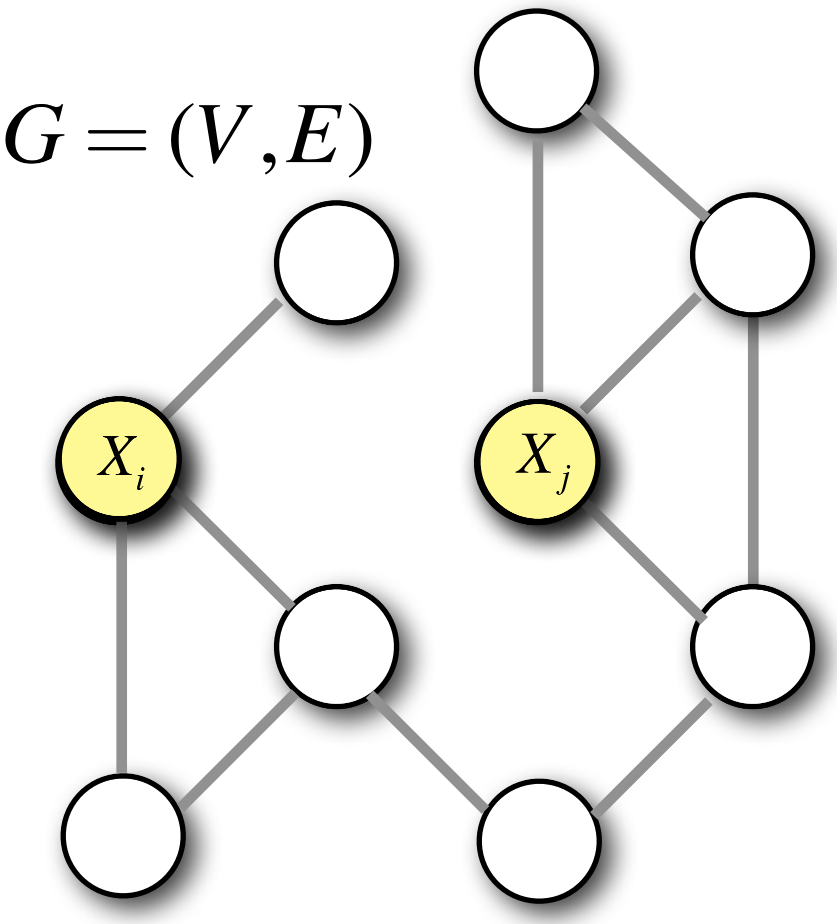
Infer conditional independence based on observational data

d variables $X_1,\dots,X_d$

$G=(V,E)$

n samples $x^1,\dots,x^n$

$$\begin{pmatrix} x_1^1 & \cdots & x_d^1 \\ \vdots & \ddots & \vdots \\ x_1^n & \cdots & x_d^n \end{pmatrix} \Rightarrow$$

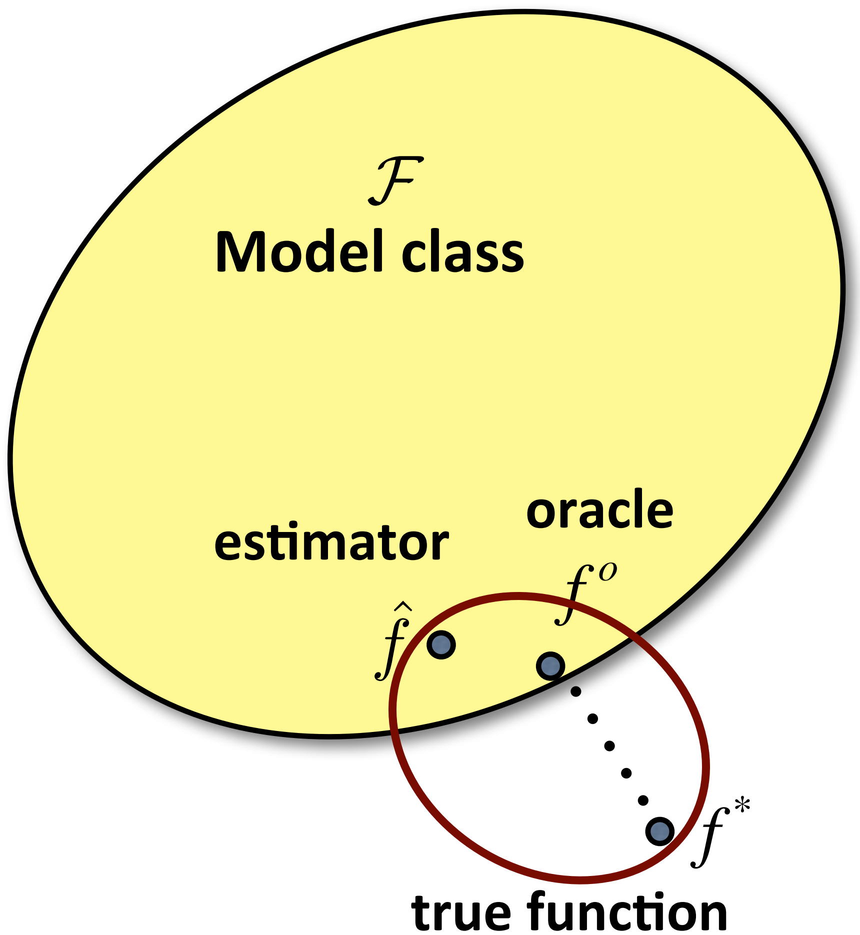


$(X_i, X_j) \notin E \Leftrightarrow X_i \perp X_j \mid$ **the rest**

Applications: density estimation, computing, visualization...

Desired Statistical Properties

Characterize the performance using different criteria



Persistency: $\text{Risk}(\hat{f}) - \text{Risk}(f^o) = o_P(1)$

Consistency: $\text{Distance}(\hat{f}, f^*) = o_P(1)$

Sparsistency: $\mathbb{P}\left(\text{graph}(\hat{f}) \neq \text{graph}(f^*)\right) = o(1)$

Minimax optimality

Outline

- **Nonparanormal**
- **Forest Density Estimation**
- **Summary**

Gaussian Graphical Models

$$X \sim N_d(\mu, \Sigma) \quad \Omega = \Sigma^{-1}$$

$\Omega_{jk} = 0 \Leftrightarrow X_j \perp X_k \mid$ the rest **(Lauritzen 96)**

glasso--Graphical Lasso (Yuan and Lin 06, Banerjee 08, Friedman et al. 08)

$$\min_{\Omega \succ 0} \{ \underbrace{\mathrm{tr}(\overset{\text{Sample covariance}}{\hat{S}}\Omega) - \log|\Omega|}_{\text{Negative Gaussian log-likelihood}} + \lambda \underbrace{\sum_{j,k} |\Omega_{jk}|}_{L_1\text{-regularization}} \}$$

Neighborhood selection (Meinshausen and Buhlmann 06)

Gaussian Graphical Models

CLIME -- Constrained L_1-Minimization Method (Cai et al. 2011)

$$\min_{\Omega} \sum_{j,k} \left|\Omega_{jk}\right| \text{ subject to } \| \hat{S}\Omega - \mathbf{I} \|_{\max} \leq \lambda$$

gDantzig -- Graphical Dantzig Selector (Yuan 2010)

Computation and Theory

Computing: scalable up to thousands of dimensions

glasso (Hastie et al.)

```
language:        Fortran
scalability:     d<3000
Speed:         very fast
```

huge (Zhao and Liu)

```
language:              C
scalability:      d<6000
Speed:       3 x faster
```

Theory: persistency, consistency, sparsistency, optimal rate,...

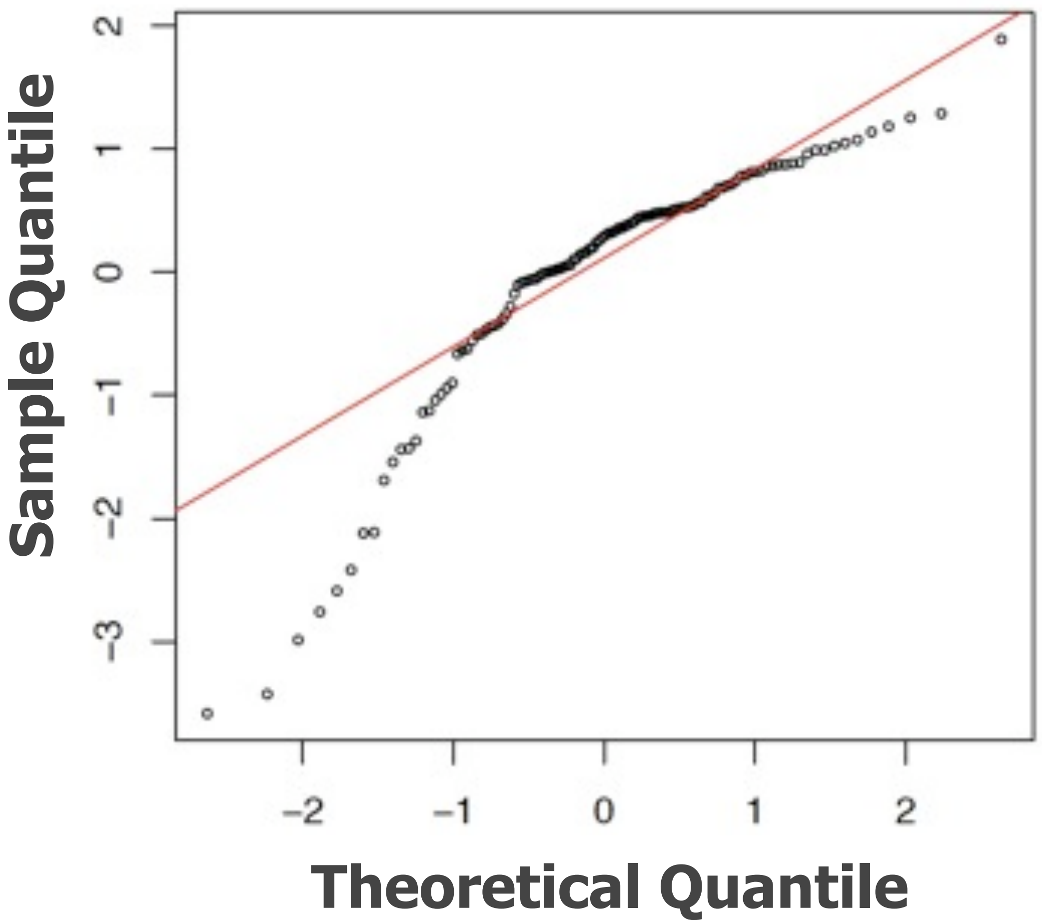
key result for analysis → $\| \hat{S} - \Sigma \|_{\max} = O_P\left(\sqrt{\frac{\log d}{n}}\right)$

sample covariance population covariance

Many Real Data are non-Gaussian

Arabidopsis Data (Wille et al. 04)
(n = 118, d=39)

Normal Q-Q plot of one typical gene



Relax the Gaussian assumption without losing statistical and computational efficiency?

The Nonparanormal

Gaussian $\Rightarrow$ Gaussian Copula

Nonparanormal Definition (Liu, Lafferty, Wasserman 09)

A random vector $X = (X_1, \ldots, X_d)$ is **nonparanormal**

$$X \sim NPN_d\left(\Sigma, \{f_j\}_{j=1}^d\right)$$

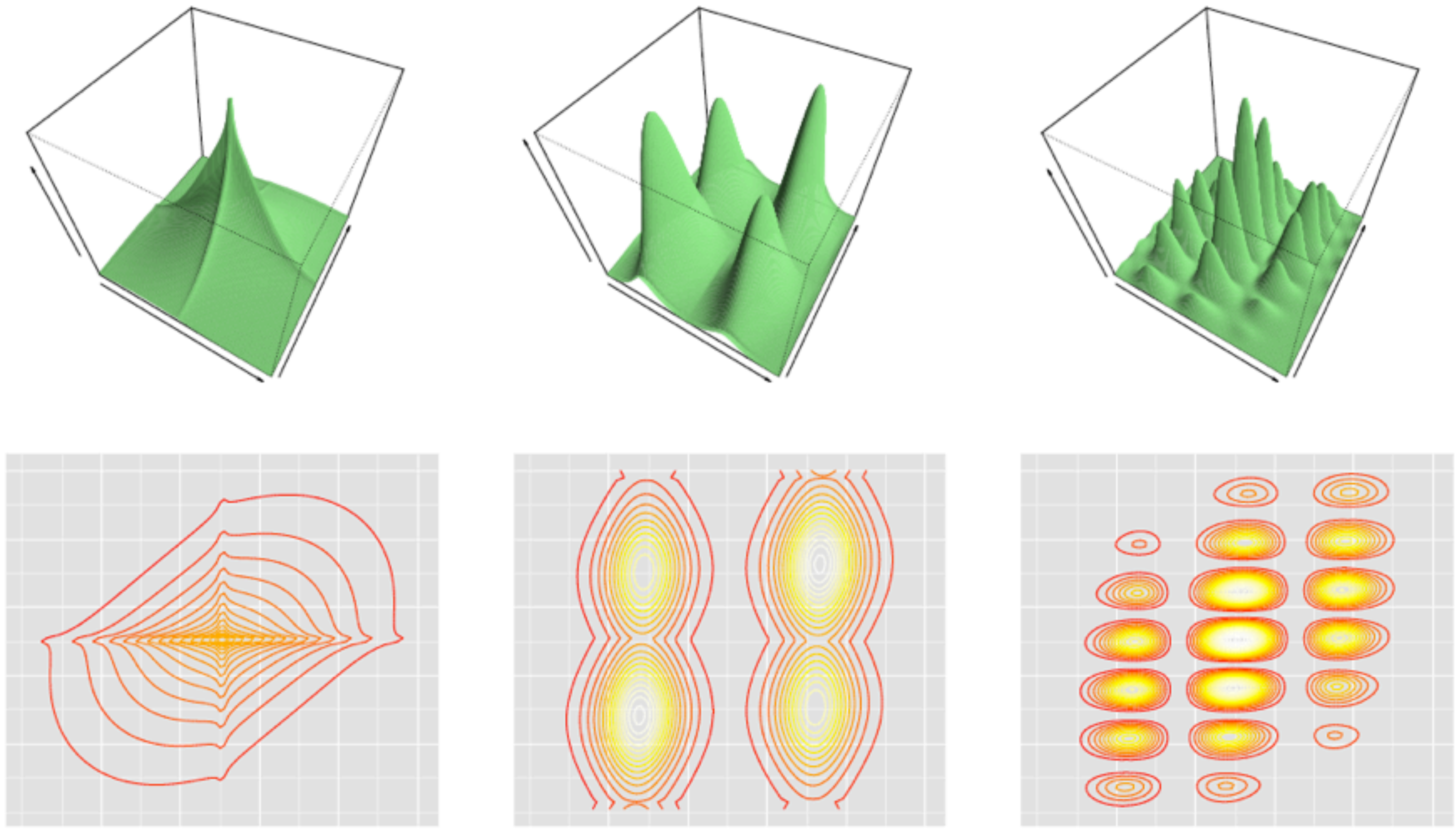
in case $f(X) = \left(f_1(X_1), \ldots, f_d(X_d)\right)$ is normal

$$f(X) \sim N_d\left(0, \Sigma\right).$$

Here f_j's are **strictly monotone** and $\mathrm{diag}(\Sigma) = 1$.

$f_j(t) = \dfrac{t - \mu_j}{\sigma_j} \implies$ **recover arbitrary Gaussian distributions**

Visualization



Bivariate nonparanormal densities with different transformations

Basic Properties

The graph is encoded in the inverse correlation matrix

Let $X \sim NPN_d\left(\Sigma, \{f_j\}_{j=1}^d\right)$ **and** $\Omega = \Sigma^{-1}$, **then**

$$p_X(x) = \frac{1}{(2\pi)^{d/2} \, |\Omega|^{-1/2}} \exp\left\{-\frac{1}{2} f(x)^T \Omega f(x)\right\} \prod_{j=1}^{d} \left| f_j'(x_j) \right|$$

$$\Downarrow$$

$\Omega_{ij} = 0 \Leftrightarrow X_i \perp X_j \mid$ the rest

Not jointly convex, how to estimate the parameters?

Estimating Transformation Functions

Directly estimate $\{f_j\}_{j=1}^{d}$ without worrying about Ω

CDF of X_j $\quad$ f_j strictly monotone $\quad$ $f_j(X_j) \sim N(0,1)$

$$F_j(t) = \mathbb{P}\left(X_j \le t\right) = \mathbb{P}\left(f_j(X_j) \le f_j(t)\right) = \Phi\left(f_j(t)\right)$$

$$\Downarrow$$

$$f_j(t) = \Phi^{-1}\left(F_j(t)\right)$$

Normal-score transformation

$$\hat{F}_j(t) = \frac{1}{n+1}\sum_{i=1}^{n} I(x_j^i \le t)$$

Estimating Inverse Correlation Matrix

Nonparanormal Algorithm (Liu, Han, Lafferty, Wasserman 12)

Step 1 : calculate the **Spearman's** rank correlation coefficient matrix $\hat{R}^{\rho}$

Step 2 : transform $\hat{R}^{\rho}$ into $\hat{\Sigma}^{\rho}$ according to

$$(*) \qquad \hat{\Sigma}^{\rho}_{jk} = 2 \cdot \sin\left(\frac{\pi}{6}\hat{R}^{\rho}_{jk}\right) \longleftarrow$$

$\hat{\Sigma}^{\rho}$ provides good estimate of Σ.

Step 3 : plug $\hat{\Sigma}^{\rho}$ into glasso / CLIME / gDantzig to get $\hat{\Omega}^{\rho}$ and the graph

The same procedure is independently proposed by (Xue and Zou 12)

Nonparanormal Theory

Theorem (Liu, Han, Lafferty, Wasserman 12)

Let $X \sim NPN_d(\Sigma, f)$ and $\Omega = \Sigma^{-1}$. Given whatever conditions on Σ and Ω that secure the **consistency and sparsistency** of glasso / CLIME / gDantzig under the Gaussian models, the nonparanormal is also consistent and sparsistent with exactly the same parametric rates of convergence.

$$\Downarrow$$

The nonparanormal is a safe replacement of the Gaussian model

Proof of the Theorem

Proof: The key is to show that $\| \hat{\Sigma}^{\rho} - \Sigma \|_{\max} = O_P\left(\sqrt{\frac{\log d}{n}}\right).$

For Gaussian distribution, **Kruskal (1948)** shows

monotone transformation invariant

Pearson's correlation coefficient

$$\boxed{\Sigma_{jk} = 2 \cdot \sin\left(\frac{\pi}{6} R^{\rho}_{jk}\right)}$$

Population Spearman's rank coefficient

Also true for the nonparanormal distribution

$$\| \hat{\Sigma}^{\rho} - \Sigma \|_{\max} \lesssim \| \hat{R}^{\rho} - R^{\rho} \|_{\max} = O_P\left(\sqrt{\frac{\log d}{n}}\right).$$

the theory of U - statistics. □

Empirical Results

For nonGaussian data, the nonparanormal >> glasso

Sample $x^i \sim NPN_d(\Sigma, f)$ **with** $n = 200,\ d = 40$ **and transformation** f_j

true graph **nonparanormal** **glasso**

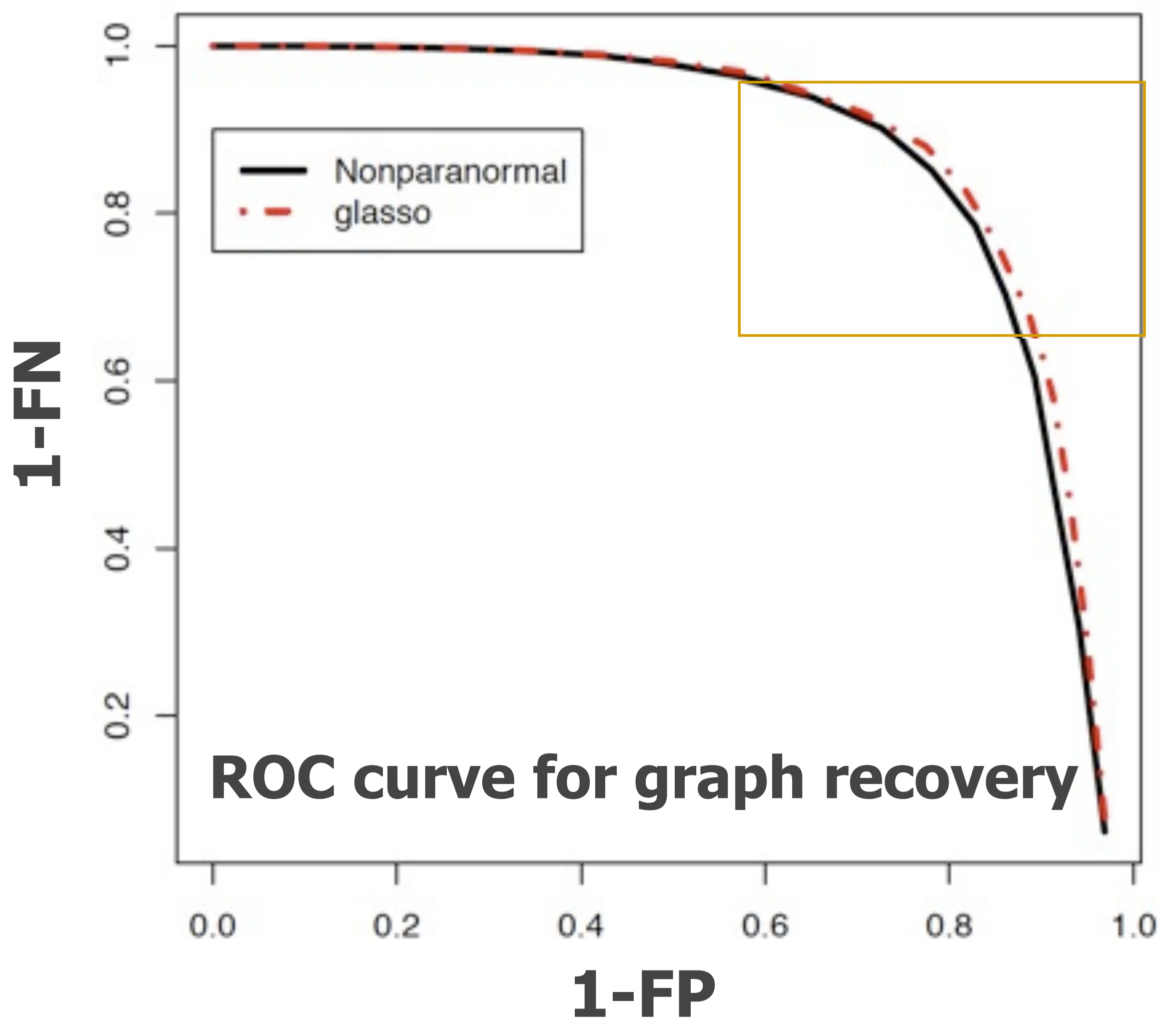
Oracle graph: **pick the best tuning parameter along the path**

Nonparanormal: Efficiency Loss

For Gaussian data, the nonparanormal almost loses no efficiency

Computationally -- no extra cost

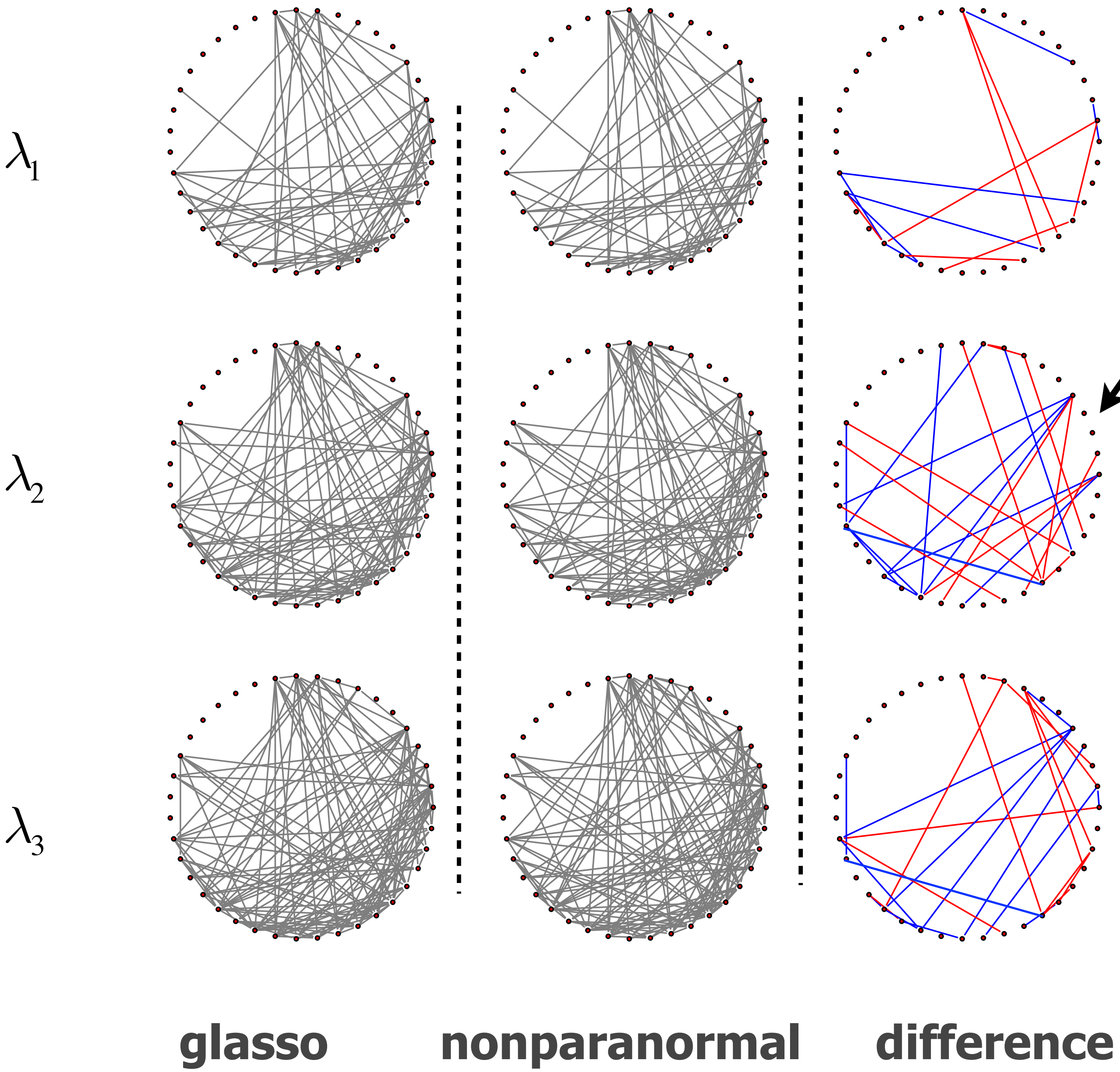
Statistically -- sample $x^1, \ldots, x^n \sim N_d(0, \Sigma)$ with $n = 80$ and $d = 100$



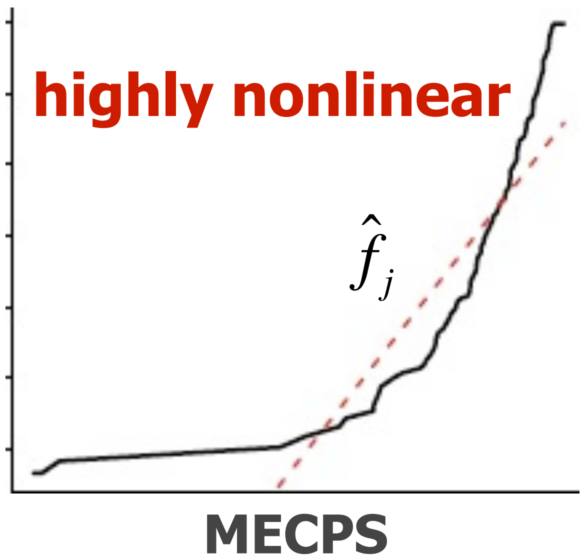
almost no efficiency loss

Arabidopsis Data

The nonparanormal behaves differently from glasso on the Arabidopsis data



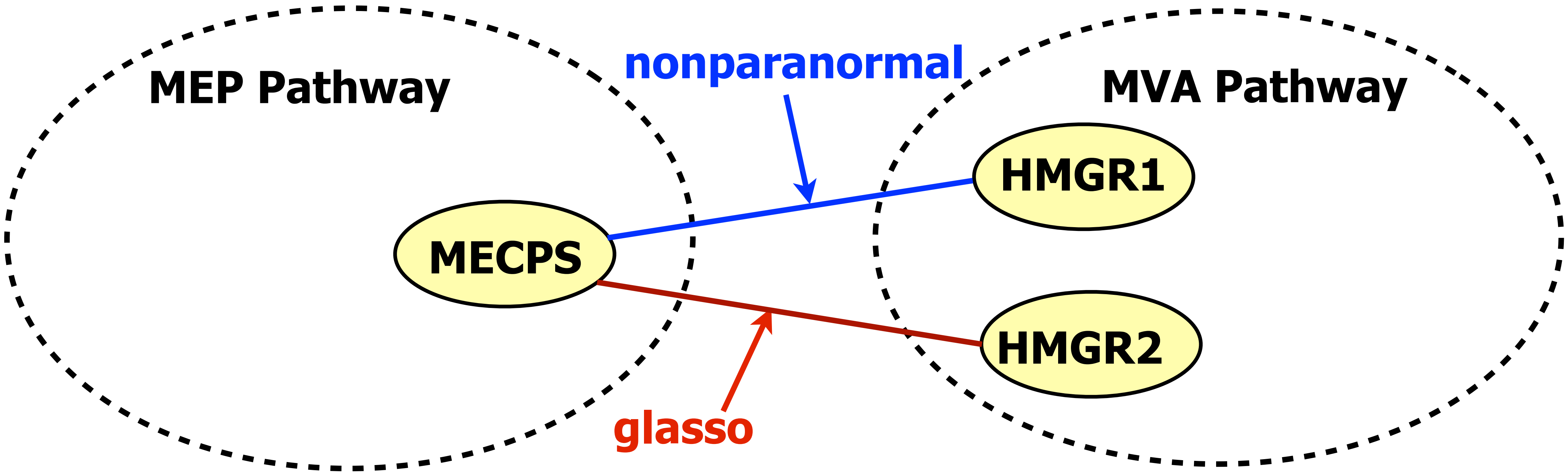
The paths are different



Nonlinear transformation causes graph difference

Scientific Implications

Cross-pathway interactions?



Still open in the current biological literature (Hou et al. 2010)

Tradeoff

Nonparanormal: **unrestricted graphs, more flexible distributions**

What if the true distribution is not nonparanormal?

Tradeoff structural flexibility for greater nonparametricity

Forest Densities

Gaussian Copula $\Rightarrow$ Fully nonparametric distribution

A forest $F=(V,E_F)$ is an acylic graph.

A distribution is supported on a forest $F=(V, E_F)$ if

$$p_F(x)=\prod_{(i,j)\in E_F}\frac{p(x_i,x_j)}{p(x_i)p(x_j)}\cdot\prod_{k\in V}p(x_k)$$

$\hat{F}=(V,E_{\hat{F}})$ $\quad \hat{p}(x_i,x_j)$, $\hat{p}(x_k)$ **Forest density estimator**

Advantages: visualization, computing, distributional flexibility, inference

Some Previous Work

Most existing work on forests are for discrete distributions

Chow and Liu (1968)

Bach and Jordan (2003)

Tan et al. (2010)

Chechetka and Guestrin (2007)

Our focus: statistical properties in high dimensions

Estimation

Find a forest $F^{(k)} = \arg\min_F \mathrm{KL}\left(p(x) \,\|\, p_F(x)\right)$ subject to $|E_F| \le k$

true density

projection of $p(x)$ onto F

Maximum weight forest problem (Kruskal 56)

$$F^{(k)} = \arg\max_F \sum_{(i,j)\in E_F} I(p_{ij}) \text{ subject to } |E_F| \le k$$

mutual information

$$I(p_{ij}) = \int p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)\,p(x_j)}\, dx_i\, dx_j$$

$\hat{p}(x_i, x_j),\ \ \hat{p}(x_k)$ **Clipped KDE**

Forest Density Estimation Algorithm

Forest Density Estimation Algorithm

1. Sort edges according to empirical mutual information $I(\hat{p}_{ij})$
2. Greedily pick a set of edges such that no cycles are formed
3. Output the obtained forest after k edges have been added

Assumptions for Forest Graph Estimation

(A1) Bivariate marginals $p(x_j, x_k) \in$ 2nd - order Hölder class

(A2) $p(x)$ has bounded support (e.g. $[0,1]^d$) and

$$\kappa_1 \leq \min_{j,k} \; p(x_j, x_k) \leq \max_{j,k} \; p(x_j, x_k) \leq \kappa_2$$

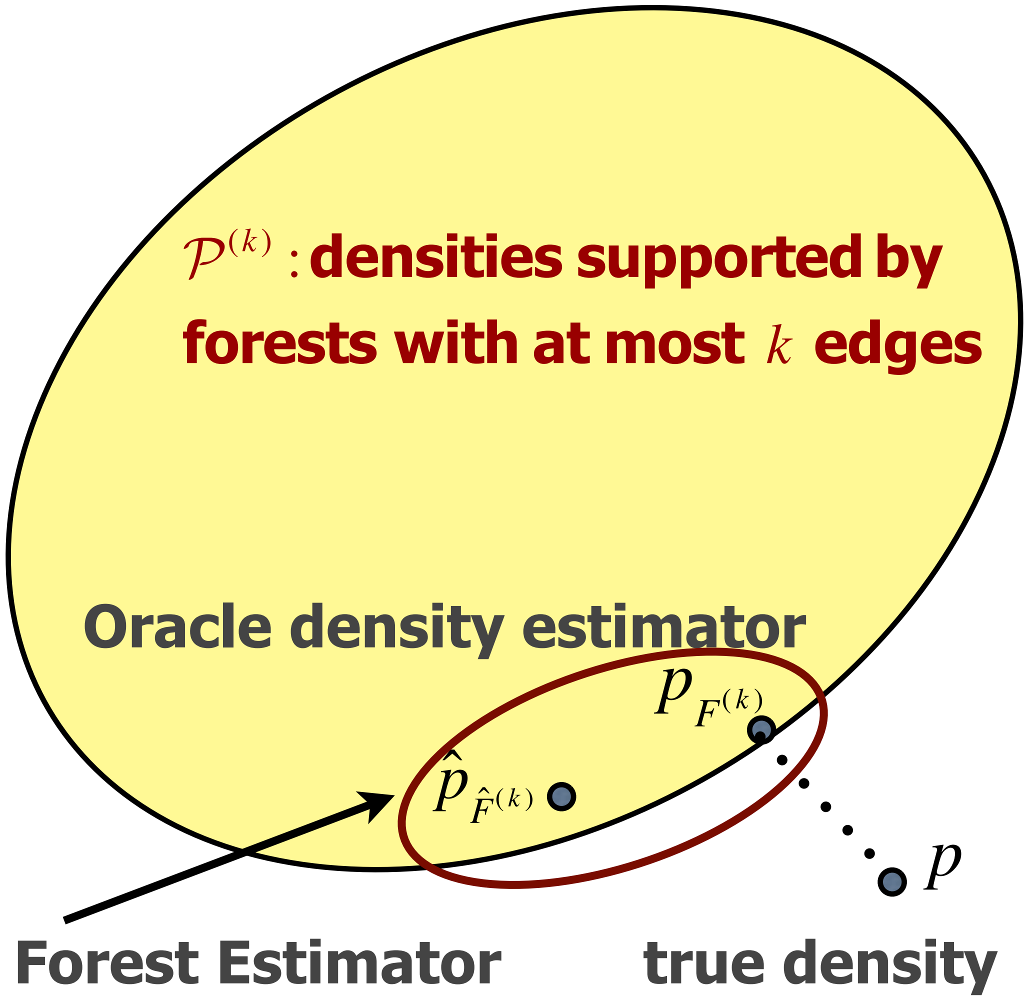
(A3) $p(x_j, x_k)$ has vanishing partial derivatives on boundaries

(A4) For a "crucial" set of edges, their mutual info. distinct enough from each other

To secure enough signal-to-noise-ratio for correct structure recovery (Tan, Anandkumar, Willsky 11)

Forest Density Estimation Theory

$$F^{(k)} = \arg\min_{F:\ |E_F| \le k} \mathrm{KL}\left(p(x) \,\|\, p_F(x)\right)$$



Theorem-Oracle Sparsistency (Liu et al. 12)

For graph estimation, let

$$\frac{\log d}{n} \to 0, \quad \longleftarrow \text{ parametric scaling}$$

and 1d and 2d KDEs use the same bandwidth

$$h \asymp n^{-1/4}, \quad \longleftarrow \text{ undersmooth}$$

we have $\sup_k \mathbb{P}\left(\hat{F}^{(k)} \neq F^{(k)}\right) = o(1)$.

Proof of the Sparsistency Result

Proof: The key is to bound

$$\underbrace{\left|I\left(\hat{p}_{jk}\right) - I(p_{jk})\right|}_{} \le \underbrace{\left|I\left(\hat{p}_{jk}\right) - \mathbb{E}I\left(\hat{p}_{jk}\right)\right|}_{\text{Stochastic}} + \underbrace{\left|\mathbb{E}I\left(\hat{p}_{jk}\right) - I(p_{jk})\right|}_{\text{Bias}}$$

$I(\hat{p}_{jk})$: estimated mutual info. $I(p_{jk})$: population mutual info.

$$\mathbb{P}\left(\textbf{Stochastic} \ge t\right) \le c_1 \exp\left(-c_2 n t^2\right) \longleftarrow \textbf{McDiarmid's inequality}$$

$$\textbf{Bias} \lesssim \sqrt{\underbrace{\int \left[\mathbb{E}\hat{p}_{jk}(x) - p_{jk}(x)\right]^2 dx}_{\sqrt{\textbf{IBias}(\hat{p}_{jk})}\ \lesssim\ h^2} + \underbrace{\int \mathbb{E}\left[\hat{p}_{jk}(x) - p_{jk}(x)\right]^2 dx}_{\textbf{IMSE}(\hat{p}_{jk})\ \lesssim\ h^4 + \frac{1}{nh^2}}} \qquad \square$$

Consistency

Theorem-Oracle Consistency (Liu et al. 12)

For density estimation, we set the bandwidths for the 1d and 2d KDE as

$$h_1 \asymp n^{-1/5} \text{ and } h_2 \asymp n^{-1/6}.$$

optimal rates for KDE

We have

$$\sup_p \mathbb{E} \| \hat{p}_{\hat{F}^{(k)}} - p_{F^{(k)}} \|_1 \leq C \cdot \sqrt{\frac{k}{n^{2/3}} + \frac{d}{n^{4/5}}}.$$

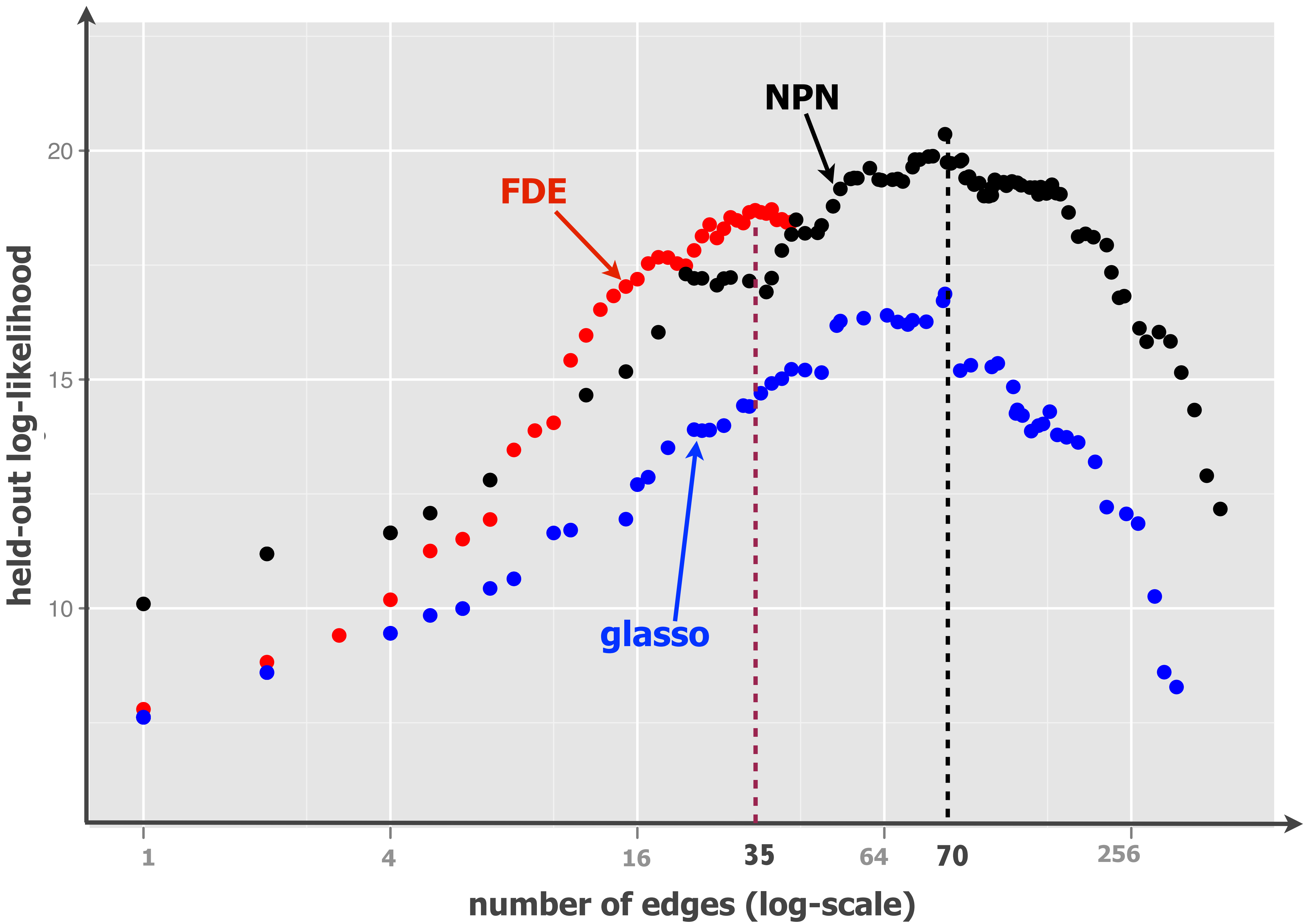
minimax optimal

bivariate KDE

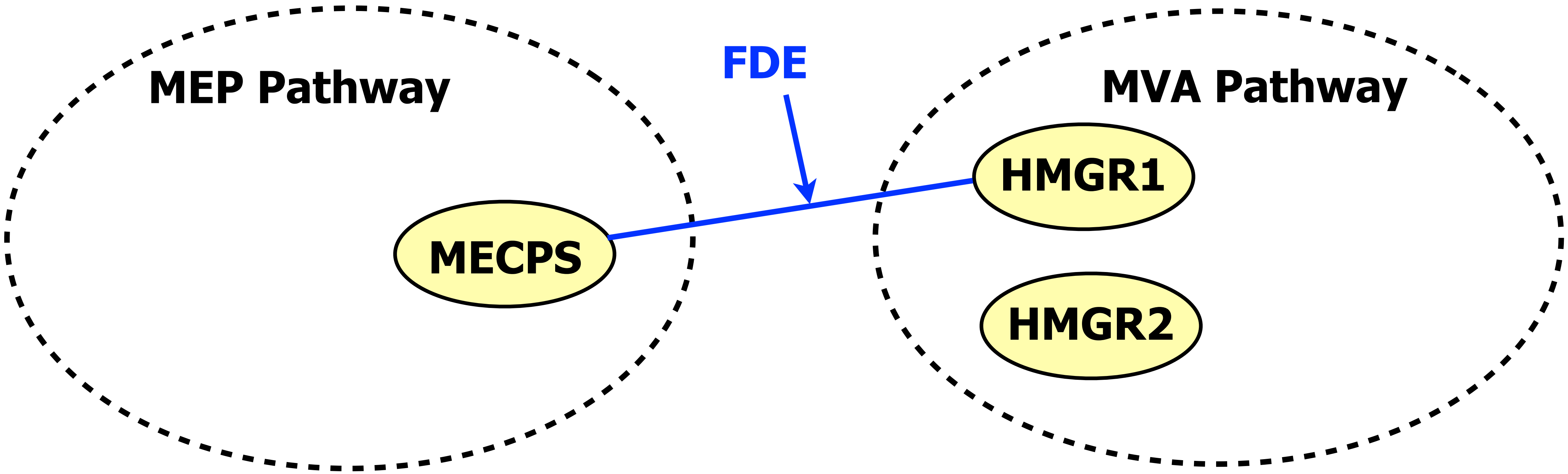
univariate KDE

Proof Pinsker's inequality and the decomposability of the forest density in terms of KL-divergence

Arabidopsis Data



Forest Graphs on the Arabadopsis Data



Forest density estimation is consistent with the nonparanormal

Nonparanormal vs. Forest Density Estimation

Second order log-density ANOVA models

$$\log p(x) = \alpha + \sum_{i=1}^{d} f_i(x_i) + \sum_{j<k} f_{jk}(x_j, x_k)$$

Nonparanormal :

$f_{jk}(x_j, x_k) = \Omega_{jk} f_j(x_j) f_k(x_k)$

and f_j, f_k are monotone.

Forest Density Estimation :

only involve at most $(d-1)$

interaction terms $f_{jk}(x_j, x_k)$.

Trade off structural complexity with distributional flexibility

Summary

Scalable nonparametric methods and high dimensional theory go together

Theory: nonparametric modeling with optimal parametric rates

Computing: as scalable as the best parametric implementation

Applications: potential to lead to nontrivial scientific insights

Software: "`huge`" and "`flare`" are available on CRAN