Variations on Nonparametric Additive Models: Computational and Statistical Aspects

John Lafferty

Department of Statistics &
Department of Computer Science
University of Chicago

Collaborators

Sivaraman Balakrishnan (CMU)
Mathias Drton (Chicago)
Rina Foygel (Chicago)
Michael Horrell (Chicago)
Han Liu (Princeton)
Kriti Puniyani (CMU)
Pradeep Ravikumar (Univ. Texas, Austin)
Larry Wasserman (CMU)

Perspective

Even the simplest models can be interesting, challenging, and useful for large, high-dimensional data.

Motivation

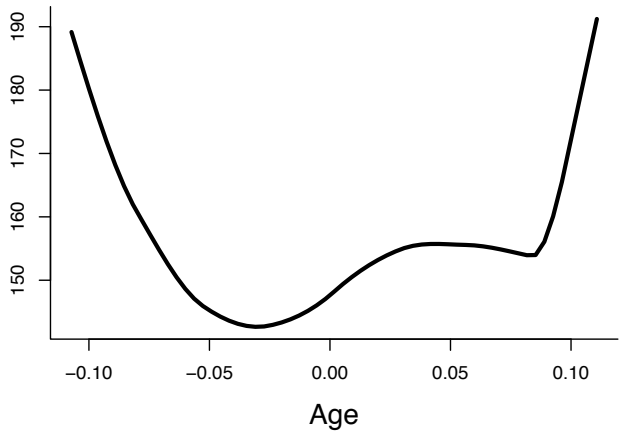
Great progress has been made on understanding sparsity for high dimensional linear models

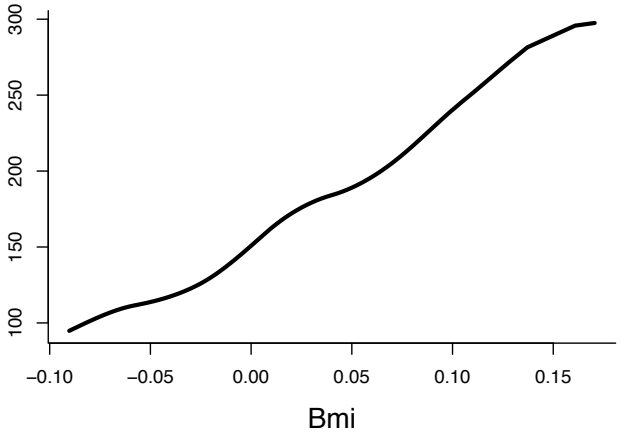
Many problems have clear nonlinear structure

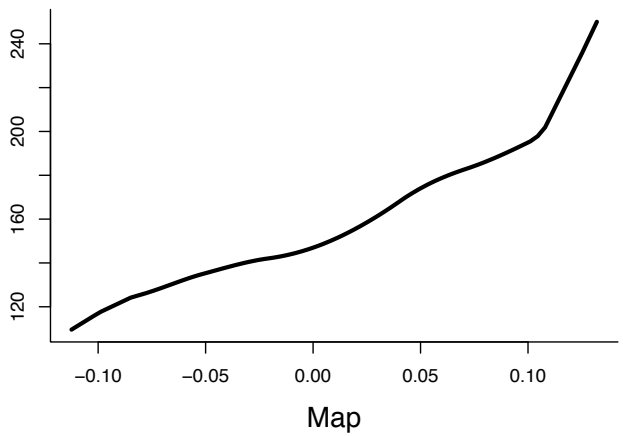
We are interested in *purely functional* methods for high dimensional, nonparametric inference

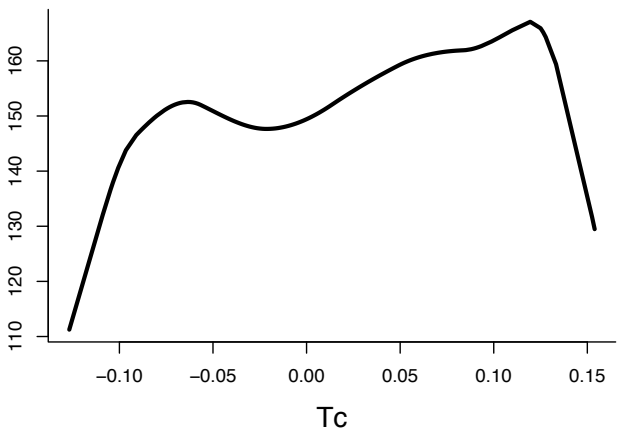
- no basis expansions

Additive Models









Additive Models

Fully nonparametric methods appear hopeless

- Logarithmic scaling, $p = \log n$ (e.g., "Rodeo" Lafferty and Wasserman (2008))

Additive models are useful compromise

- Exponential scaling, $p = \exp(n^c)$ (e.g., "SpAM" Ravikumar et al. (2009))

Themes of this talk

- Variations on additive models enjoy most of the good statistical and computational properties of sparse linear models
- Thresholded backfitting algorithms, via subdifferential calculus
- RKHS formulations are problematic
- A little nonparametricity goes a long way

Outline

- Sparse additive models
- Nonparametric reduced rank regression
- Functional sparse CCA
- The nonparanormal
- Conclusions

Sparse Additive Models

Ravikumar, Lafferty, Liu and Wasserman, JRSS B (2009)

Additive Model: $Y_i = \sum_{j=1}^{p} m_j(X_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n$

High dimensional: $n \ll p$, with most $m_j = 0$.

Optimization: minimize $\mathbb{E}\left(Y - \sum_j m_j(X_j)\right)^2$

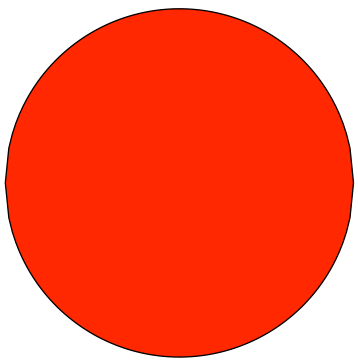
subject to $\sum_{j=1}^{p} \sqrt{\mathbb{E}(m_j^2)} \leq L_n$

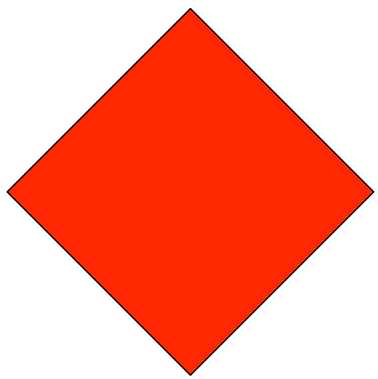
$\mathbb{E}(m_j) = 0$

Related work by Bühlmann and van de Geer (2009), Koltchinskii and Yuan (2010), Raskutti, Wainwright and Yu (2011)

Sparse Additive Models

$$C = \left\{ m \in \mathbb{R}^4 \ : \ \sqrt{m_{11}^2 + m_{21}^2} + \sqrt{m_{12}^2 + m_{22}^2} \leq L \right\}$$

$\pi_{12} C =$ 

$\pi_{13} C =$ 

Stationary Conditions

Lagrangian

$$\mathcal{L}(f, \lambda, \mu) = \frac{1}{2}\mathbb{E}\left(Y - \sum_{j=1}^{p} m_j(X_j)\right)^2 + \lambda \sum_{j=1}^{p} \sqrt{\mathbb{E}(m_j^2(X_j))}$$

Let $R_j = Y - \sum_{k \neq j} m_k(X_k)$ be jth residual. Stationary condition

$$m_j - \mathbb{E}(R_j \mid X_j) + \lambda v_j = 0 \quad a.e.$$

where $v_j \in \partial\sqrt{\mathbb{E}(m_j^2)}$ satisfies

$$\begin{aligned} v_j &= \frac{m_j}{\sqrt{\mathbb{E}(m_j^2)}} \quad \text{if } \mathbb{E}(m_j^2) \neq 0 \\ \sqrt{\mathbb{E}v_j^2} &\leq 1 \quad \text{otherwise} \end{aligned}$$

Stationary Conditions

Rewriting,

$$
\begin{aligned}
m_j + \lambda v_j &= \mathbb{E}(R_j \mid X_j) \equiv P_j \\
\left(1 + \frac{\lambda}{\sqrt{\mathbb{E}(m_j^2)}}\right) m_j &= P_j \text{ if } \mathbb{E}(P_j^2) > \lambda \\
m_j &= 0 \text{ otherwise}
\end{aligned}
$$

This implies

$$
m_j = \left[1 - \frac{\lambda}{\sqrt{\mathbb{E}(P_j^2)}}\right]_+ P_j
$$

SpAM Backfitting Algorithm

Input: Data (X_i, Y_i), regularization parameter λ.

Iterate until convergence:

For each $j = 1, \ldots, p$:

Compute residual: $R_j = Y - \sum_{k \neq j} \widehat{m}_k(X_k)$

Estimate projection $P_j = \mathbb{E}(R_j \mid X_j)$, smooth: $\widehat{P}_j = \mathcal{S}_j R_j$

Estimate norm: $s_j = \sqrt{\mathbb{E}[P_j]^2}$

Soft-threshold: $\widehat{m}_j \leftarrow \left[1 - \frac{\lambda}{\widehat{s}_j}\right]_+ \widehat{P}_j$

Output: Estimator $\widehat{m}(X_i) = \sum_j \widehat{m}_j(X_{ij})$.

Example: Boston Housing Data

Predict house value Y from 10 covariates.

We added 20 irrelevant (random) covariates to test the method.

$Y =$ house value; $n = 506$, $p = 30$.

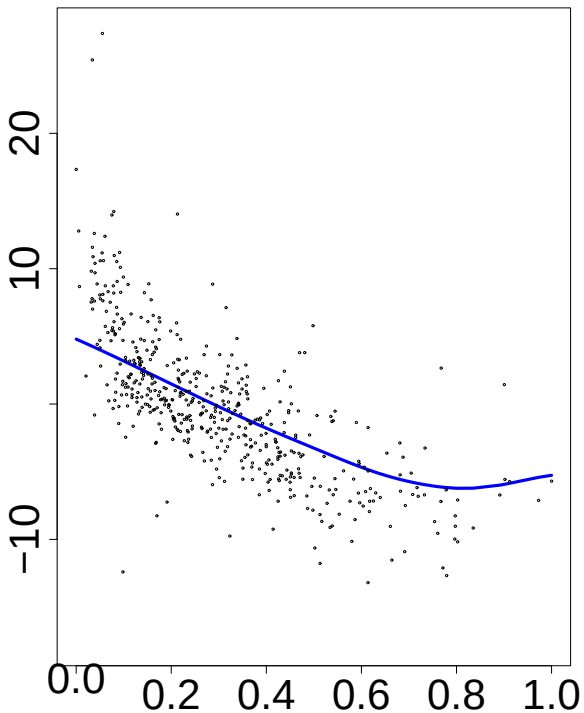
$$Y = \beta_0 + m_1(\text{crime}) + m_2(\text{tax}) + \cdots + \cdots m_{30}(X_{30}) + \epsilon.$$

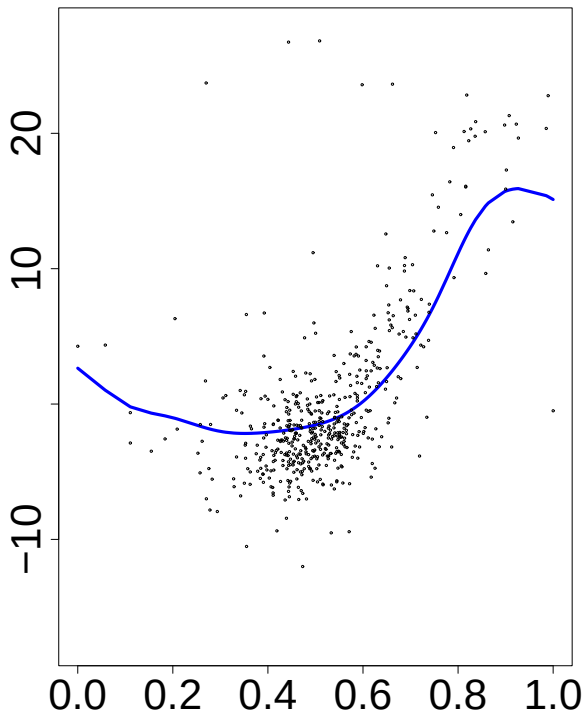
Note that $m_{11} = \cdots = m_{30} = 0$.

We choose λ by minimizing the estimated risk.

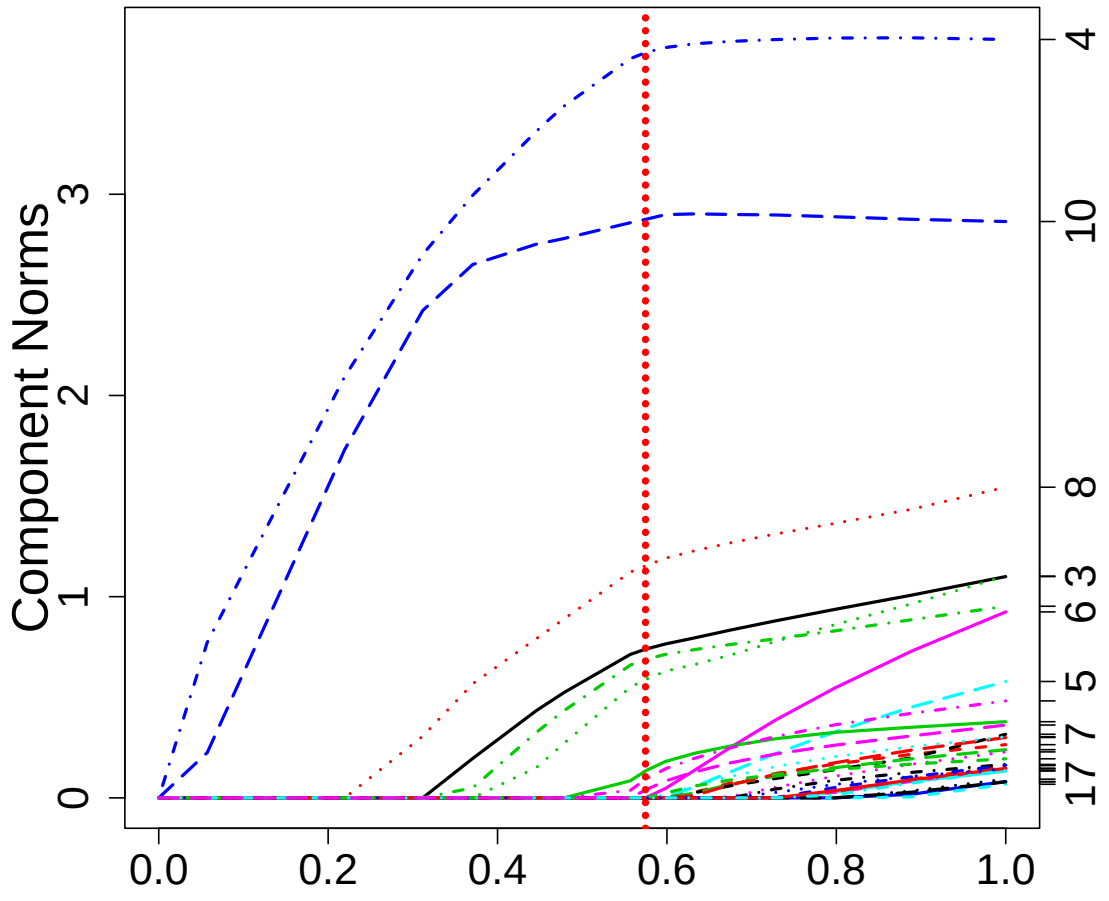
SpAM yields 6 nonzero functions. It correctly reports that $\widehat{m}_{11} = \cdots = \widehat{m}_{30} = 0$.

Example Fits





L_2 norms of fitted functions versus $1/\lambda$



RKHS Version

Raskutti, Wainwright and Yu (2011)

Sample optimization

$$\min_f \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p} m_j(x_{ij})\Big)^2 + \lambda \sum_j \|m_j\|_{\mathcal{H}_j} + \mu \sum_j \|m_j\|_{L_2(\mathbb{P}_n)}$$

where $\|m_j\|_{L_2(\mathbb{P}_n)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} m_j^2(x_{ij})}$.

By Representer Theorem, with $m_j(\cdot) = K_j\alpha_j$,

$$\min_f \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p} K_j\alpha_j\Big)^2 + \lambda \sum_j \sqrt{\alpha_j^T K_j \alpha_j} + \mu \sum_j \sqrt{\alpha_j^T K_j^2 \alpha_j}$$

Finite dimensional SOCP, but no scalable algorithms known.

Open Problems

- Under what conditions do the backfitting algorithms converge?
- What guarantees can be given on the solution to the infinite dimensional optimization?
- Is it possible to simultaneously adapt to unknown smoothness and sparsity?

Multivariate Regression

$Y \in \mathbb{R}^q$ and $X \in \mathbb{R}^p$. Regression function $M(X) = \mathbb{E}(Y \mid X)$.

Linear model $M(X) = BX$ where $B \in \mathbb{R}^{q \times p}$.

Reduced rank regression: $r = \text{rank}(B) \leq C$.

Recent work has studied properties and high dimensional scaling of reduced rank regression where nuclear norm

$$\|B\|_* := \sum_{j=1}^{\min(p,q)} \sigma_j(B)$$

as convex surrogate for rank constraint (Yuan et al., 2007; Negahban and Wainwright, 2011)

Nonparametric Reduced Rank Regression

Foygel, Horrell, Drton and Lafferty (2012)

Nonparametric multivariate regression $M(X) = (m^1(X), \ldots, m^q(X))^T$

Each component an additive model

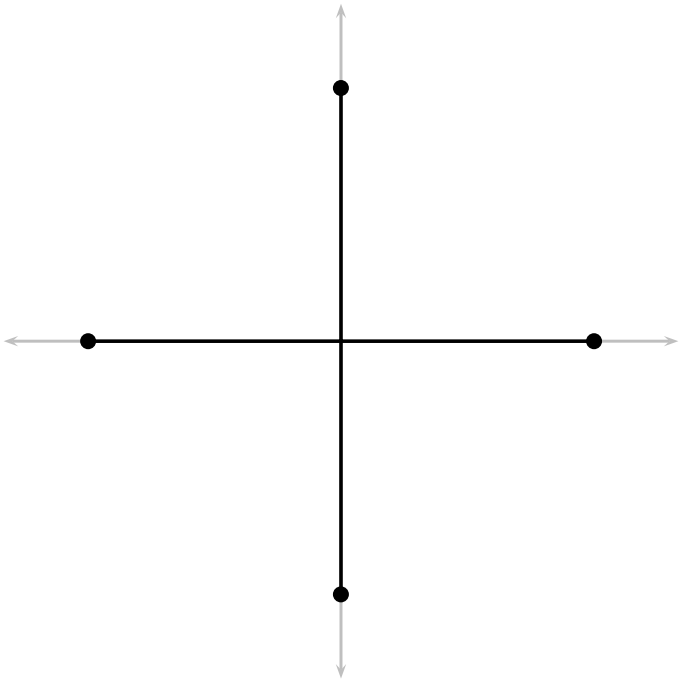
$$m^k(X) = \sum_{j=1}^{p} m_j^k(X_j)$$

What is the nonparametric analogue of $\|B\|_*$ penalty?

Recall: Sparse Vectors and ℓ_1 Relaxation

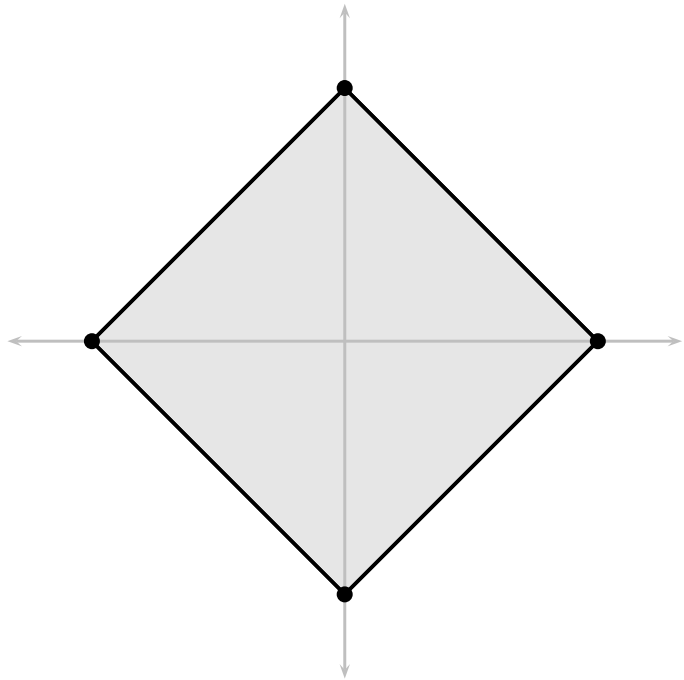
sparse vectors

$\|X\|_0 \leq t$



convex hull

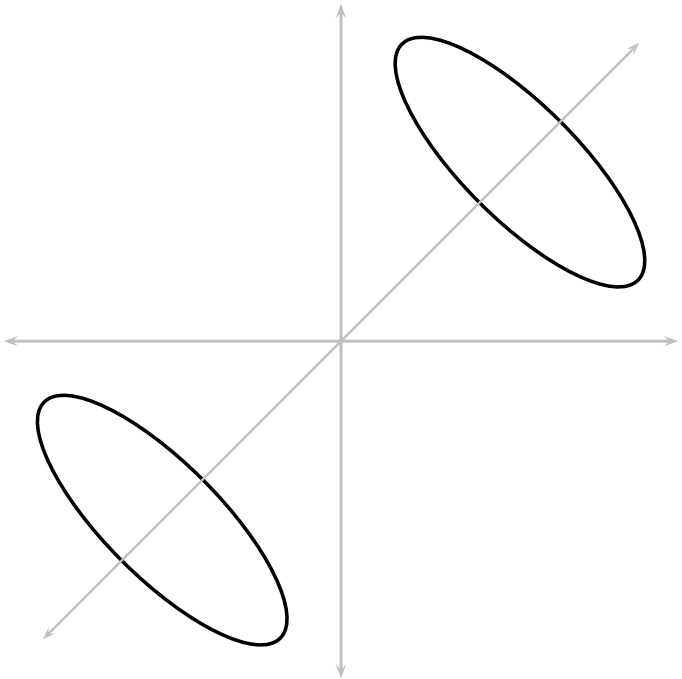
$\|X\|_1 \leq t$



Low-Rank Matrices and Convex Relaxation

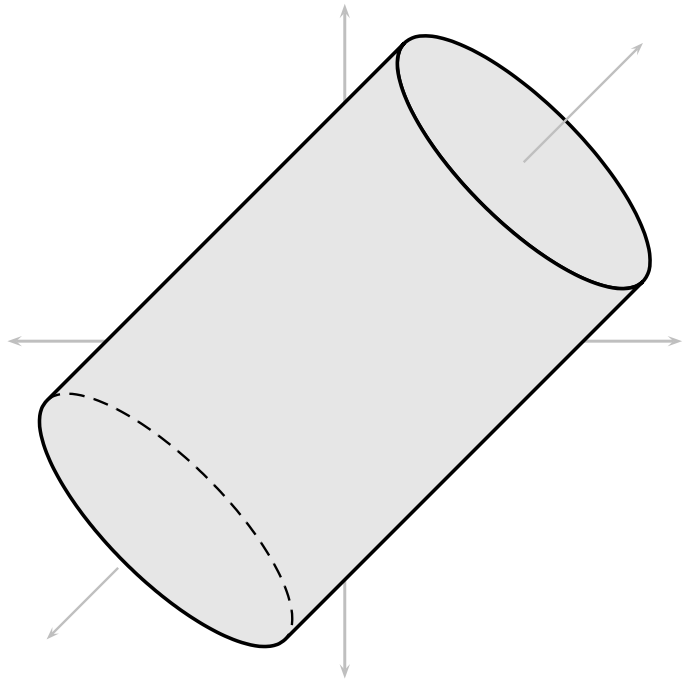
low rank matrices

$\text{rank}(X) \leq t$



convex hull

$\|X\|_* \leq t$



Nuclear Norm Regularization

Algorithms for nuclear norm minimization are a lot like iterative soft thresholding for lasso problems.

To project a matrix B onto the nuclear norm ball $\|X\|_* \leq t$:

- Compute the SVD:

$$B = U \operatorname{diag}(\sigma)\, V^T$$

- Soft threshold the singular values:

$$B \leftarrow U \operatorname{diag}(\operatorname{Soft}_\lambda(\sigma))\, V^T$$

Low Rank Functions

What does it mean for a set of functions $m^1(x), \ldots, m^q(x)$ to be low rank?

Let $x_1, \ldots, x_n$ be a collection of points.

We require the $n \times q$ matrix $\mathbb{M}(x_{1:n}) = [m^k(x_i)]$ is low rank.

Stochastic setting: $\mathbb{M} = [m^k(X_i)]$. Natural penalty is

$$\|\mathbb{M}\|_* = \sum_{s=1}^{q} \sigma_s(\mathbb{M}) = \sum_{s=1}^{q} \sqrt{\lambda_s(\mathbb{M}^T\mathbb{M})}$$

Population version:

$$|||M|||_* := \left\| \sqrt{\mathrm{Cov}(M(X))} \right\|_* = \left\| \Sigma(M)^{1/2} \right\|_*$$

Constrained Rank Additive Models (CRAM)

Let $\Sigma_j = \mathrm{Cov}(M_j)$. Two natural penalties:

$$\left\|\Sigma_1^{1/2}\right\|_* + \left\|\Sigma_2^{1/2}\right\|_* + \cdots + \left\|\Sigma_p^{1/2}\right\|_*$$

$$\left\|(\Sigma_1^{1/2}\Sigma_2^{1/2}\cdots\Sigma_p^{1/2})\right\|_*$$

Population risk functional (first penalty)

$$\frac{1}{2}\mathbb{E}\left\|Y - \sum_j M_j(X_j)\right\|_2^2 + \lambda \sum_j |||M_j|||_*$$

Stationary Conditions

Subdifferential is $\partial \|\!|\!| F |\!|\!|_* = \left\{ \left(\sqrt{\mathbb{E}(FF^\top)} \right)^{-1} F + H \right\}$ where

$\|\!|\!| H |\!|\!|_{\text{sp}} \leq 1,\ \mathbb{E}(FH^\top) = 0,\ \mathbb{E}(FF^\top)H = 0$

Let $P(X) := \mathbb{E}(Y \mid X)$ and consider optimization

$$\frac{1}{2}\mathbb{E}\left\| Y - M(X) \right\|_2^2 + \lambda \|\!|\!| M |\!|\!|_*$$

Let $\mathbb{E}(PP^T) = U \operatorname{diag}(\tau)\, U^T$ be the SVD. Define

$$M = U \operatorname{diag}([1 - \lambda/\sqrt{\tau}]_+)\, U^T P$$

Then M is a stationary point of the optimization, satisfying

$$E(Y \mid X) = M(X) + \lambda V(X) \ \ a.e.,\ \text{for some } V \in \partial \|\!|\!| M |\!|\!|_*$$

CRAM Backfitting Algorithm (Penalty 1)

Input: Data (X_i, Y_i), regularization parameter λ.

Iterate until convergence:

For each $j = 1, \ldots, p$:

Compute residual: $R_j = Y - \sum_{k \neq j} \widehat{f}_k(X_k)$

Estimate projection $P_j = \mathbb{E}(R_j \mid X_j)$, smooth: $\widehat{P}_j = \mathcal{S}_j R_j$

Compute SVD: $\frac{1}{n} \widehat{P}_j \widehat{P}_j^T = U \, \mathrm{diag}(\tau) \, U^T$

Soft-threshold: $\widehat{M}_j = U \, \mathrm{diag}([1 - \lambda/\sqrt{\tau}]_+) U^T \widehat{P}_j$

Output: Estimator $\widehat{M}(X_i) = \sum_j \widehat{M}_j(X_{ij})$.

Example

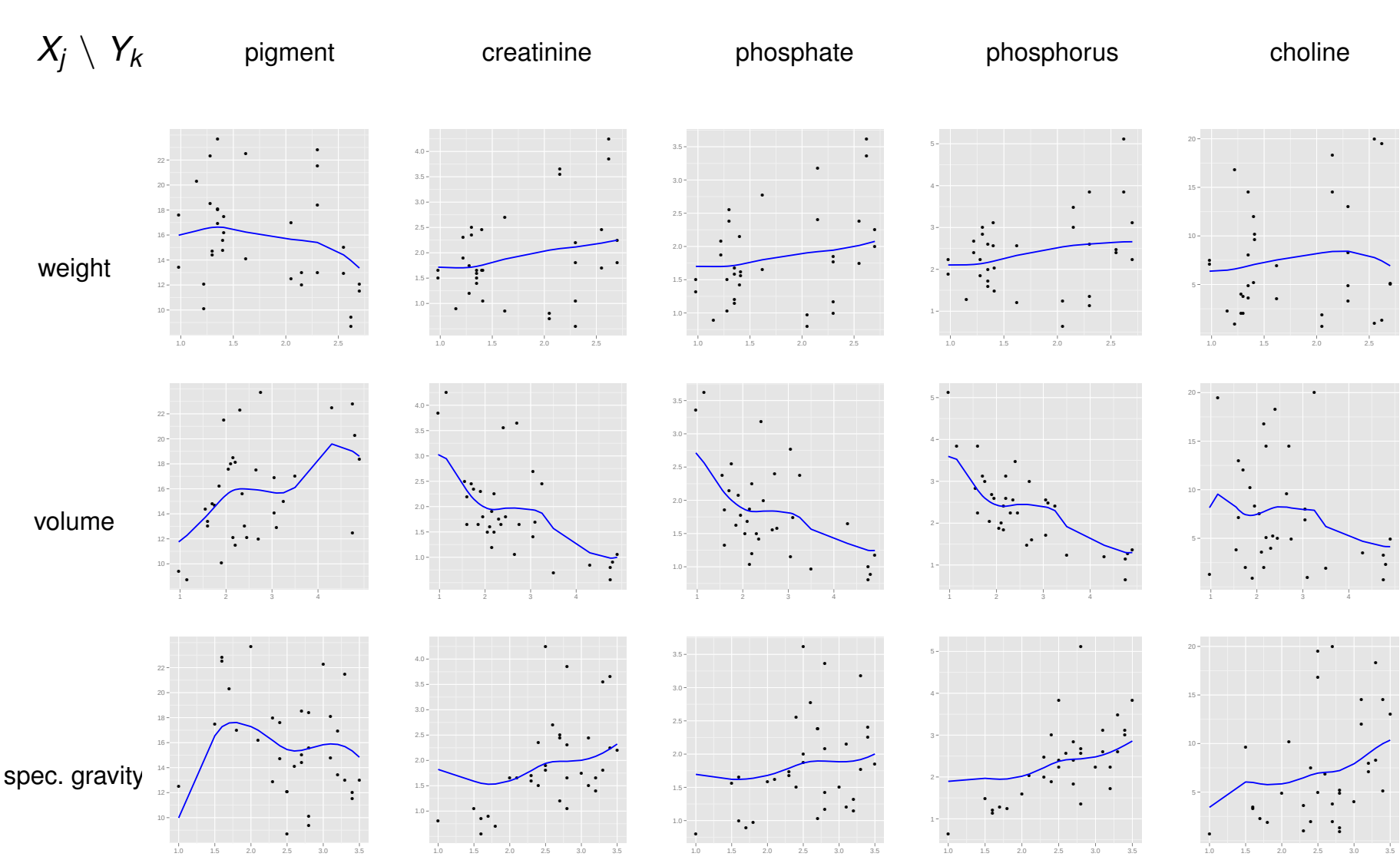
Data of Smith et al. (1962), chemical measurements for 33 individual urine specimens.

$q = 5$ response variables: pigment creatinine, and the concentrations (in mg/ml) of phosphate, phosphorus, creatinine and choline.

$p = 3$ covariates: weight of subject, volume and specific gravity of specimen.

We use Penalty 2 with local linear smoothing.

We take $\lambda = 1$ and bandwidth $h = .3$.


$X_j \setminus Y_k$
pigment
creatinine
phosphate
phosphorus
choline
weight
volume
spec. gravity

Statistical Scaling for Prediction

Let $\mathcal{F}$ be class of matrices of functions that have a functional SVD

$$M(X) = UDV(X)^\top$$

where $\mathbb{E}(V^\top V) = I$, and $V(X) = [v_{sj}(X_j)]$ with each v_{sj} in a second-order Sobolev space. Define

$$\mathcal{M}_n = \left\{ M : M \in \mathcal{F},\ \|D\|_* = o\left(\frac{n}{q + \log(pq)} \right)^{1/4} \right\}.$$

Let $\widehat{M}$ minimize the empirical risk $\frac{1}{n} \sum_i \| Y_i - \sum_j M_j(X_{ij}) \|_2^2$ over the class $\mathcal{M}_n$. Then

$$R(\widehat{M}) - \inf_{M \in \mathcal{M}_n} R(M) \xrightarrow{P} 0 .$$

Nonparametric CCA

Canonical correlation analysis (CCA, Hotelling, 1936) is classical method for finding correlations between components of two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$.

Sparse versions have been proposed for high dimensional data (Witten & Tibshirani, 2009)

Sparse additive models can be extended to this setting.

Sparse Additive Functional CCA

Balasubramanian, Puniyani and Lafferty (2012)

Population version of optimization:

$$\max_{f \in \mathcal{F},\, g \in \mathcal{G}} \mathbb{E}\left(f(X)g(Y)\right) \qquad \text{subject to}$$

$$\max_j \mathbb{E}(f_j^2) \leq 1, \quad \sum_{j=1}^{p} \sqrt{\mathbb{E}(f_j^2)} \leq C_f$$

$$\max_k \mathbb{E}(g_k^2) \leq 1, \quad \sum_{k=1}^{q} \sqrt{\mathbb{E}(g_k^2)} \leq C_g$$

Estimated with analogues of SpAM backfitting, together with screening procedures. See ICML paper.

Regression vs. Graphical Models

assumptions	*regression*	*graphical models*
parametric	lasso	graphical lasso
nonparametric	sparse additive model	*nonparanormal*

The Nonparanormal

Liu, Lafferty and Wasserman, JMLR 2009

A random vector $X = (X_1, \ldots, X_p)^T$ has a *nonparanormal* distribution

$$X \sim NPN(\mu, \Sigma, f)$$

in case

$$Z \equiv f(X) \sim N(\mu, \Sigma)$$

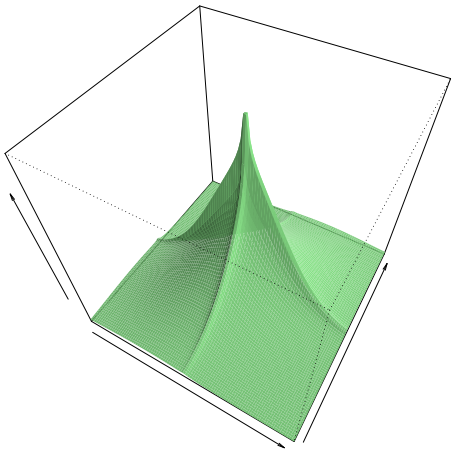
where $f(X) = (f_1(X_1), \ldots, f_p(X_p))$.

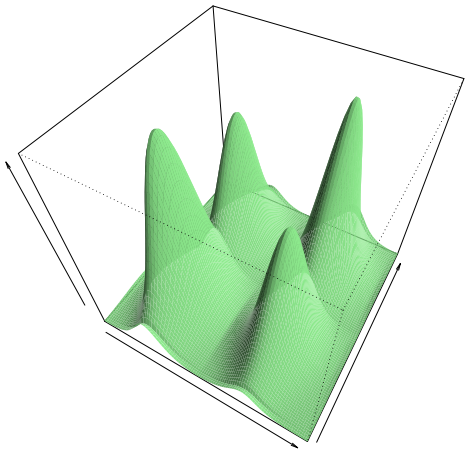
Joint density

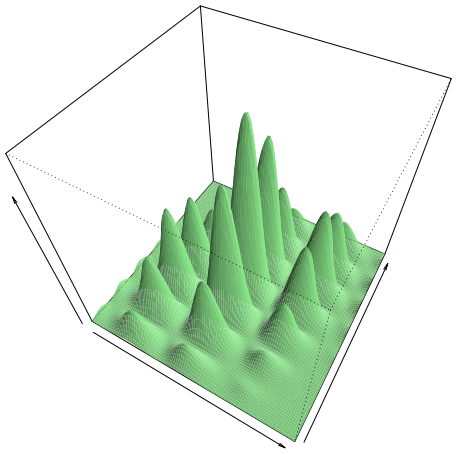
$$p_X(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(f(x)-\mu)^T \Sigma^{-1}(f(x)-\mu)\right\} \prod_{j=1}^{p} |f_j'(x_j)|$$

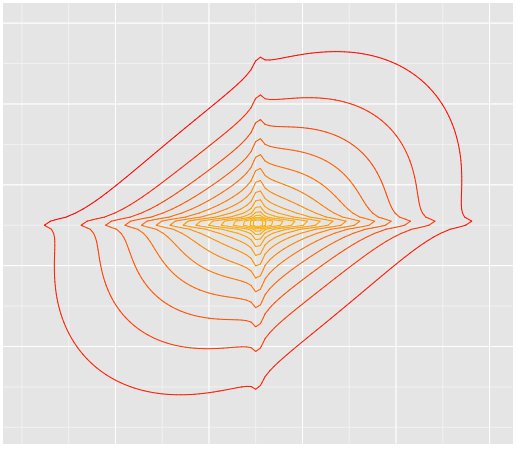
- Semiparametric Gaussian copula

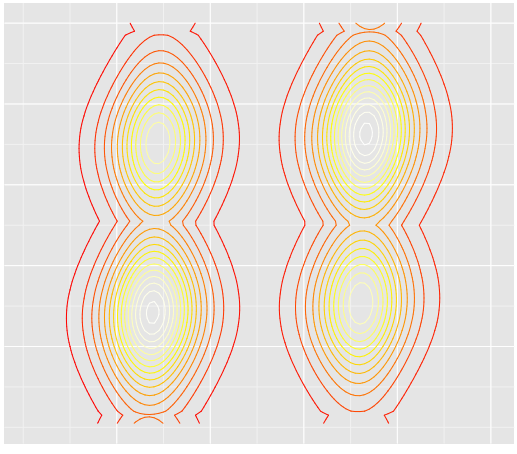
Examples

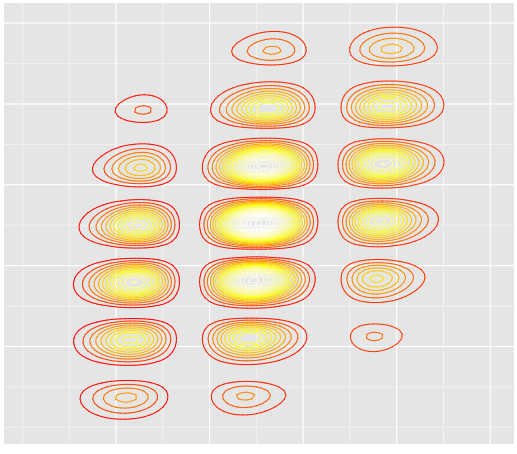












The Nonparanormal

- Define $h_j(x) = \Phi^{-1}(F_j(x))$ where $F_j(x) = \mathbb{P}(X_j \leq x)$.
- Let Λ be the covariance matrix of $Z = h(X)$. Then

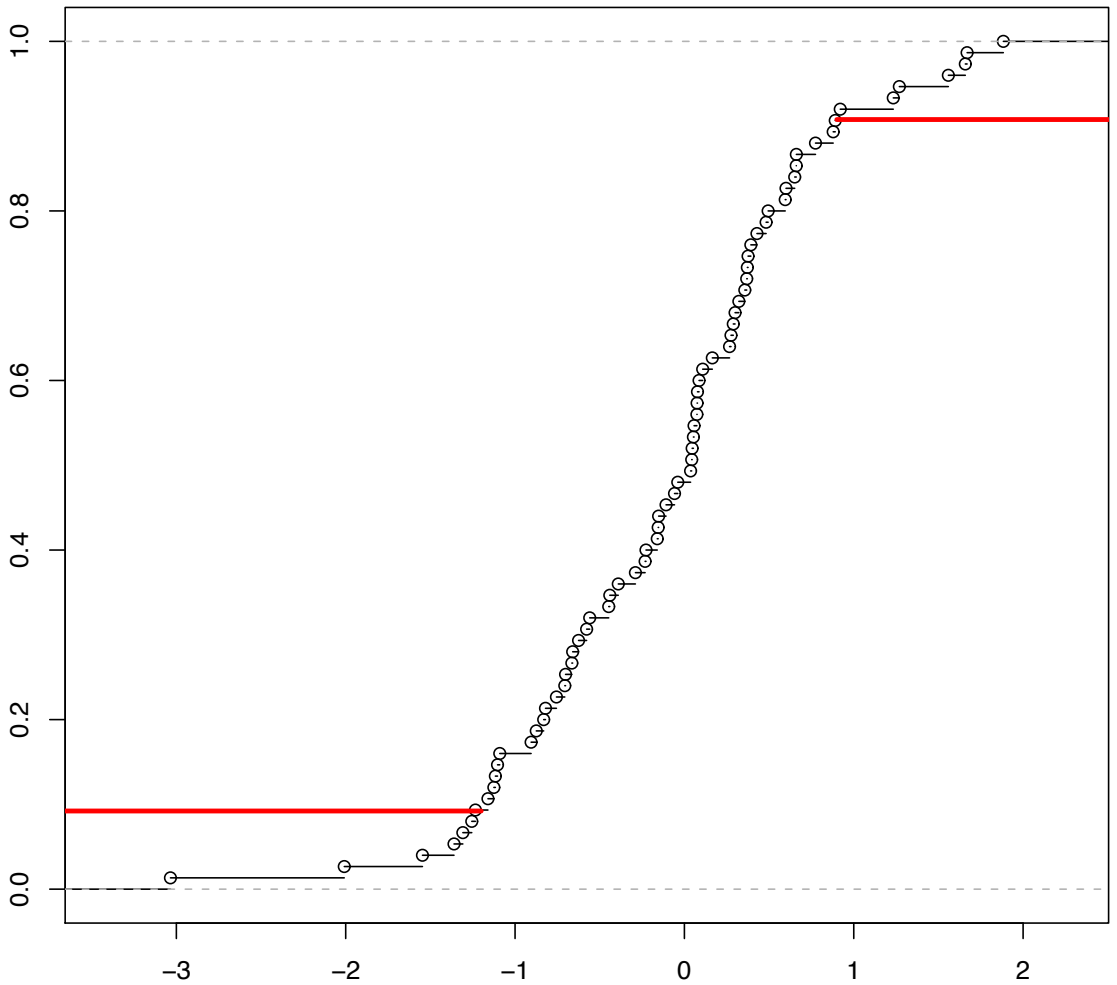
$$X_j \amalg X_k \,\Big|\, X_{\text{rest}}$$

  if and only if $\Lambda^{-1}_{jk} = 0$.
- Hence we need to:
  1. Estimate $\widehat{h}_j(x) = \Phi^{-1}(\widehat{F}_j(x))$.
  2. Estimate covariance matrix of $Z = \widehat{h}(X)$ using the glasso.

Winsorizing the CDF

Truncation to estimate $\widehat{F}_j$ for $n > p$:

$$\delta_n \equiv \frac{1}{4n^{1/4}\sqrt{\pi \log n}}$$



$$\delta_n \equiv \frac{1}{4n^{1/4}\sqrt{\pi \log n}}$$

Properties

- LLW (2009) show that the resulting procedure has the same theoretical properties as the glasso, even with dimension p increasing with n.
- The truncation of the empirical distribution is crucial for the theoretical results when p is large, although in practice it does not seem to matter too much.
- If the nonparanormal is used when the data are actually Normal, little efficiency is lost.

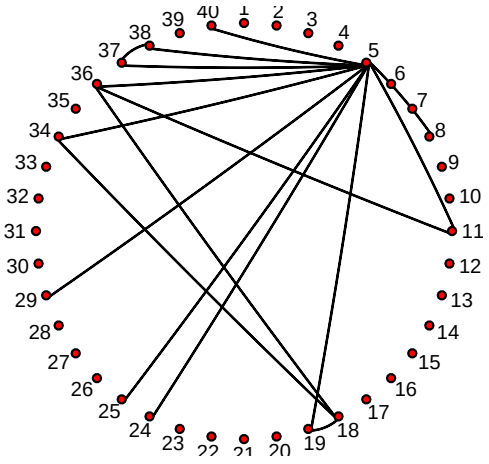
Gene-Gene Interactions for *Arabidopsis thaliana*

source: wikipedia.org

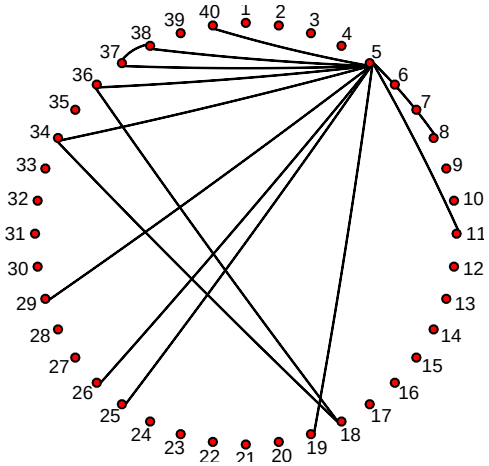
Dataset from Affymetrix microarrays, sample size $n = 118$, $p = 40$ genes (isoprenoid pathway).

Example Results

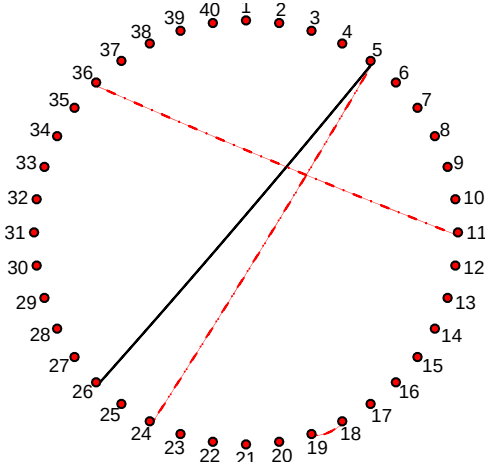
NPN

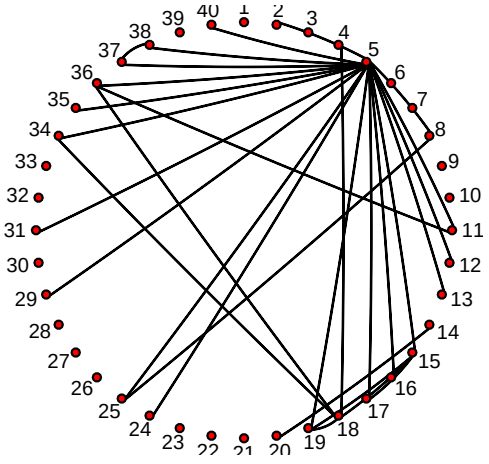


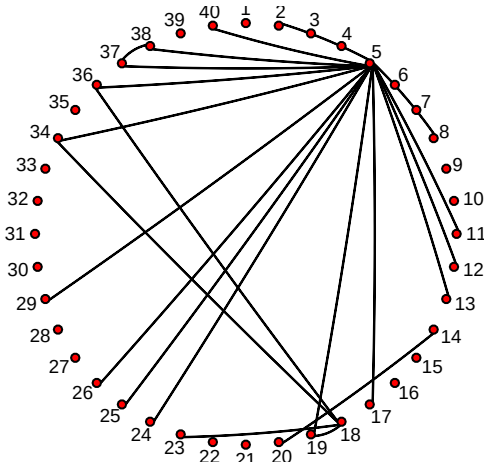
glasso

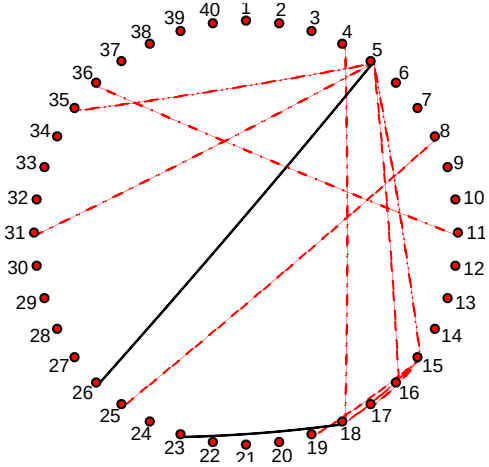


difference

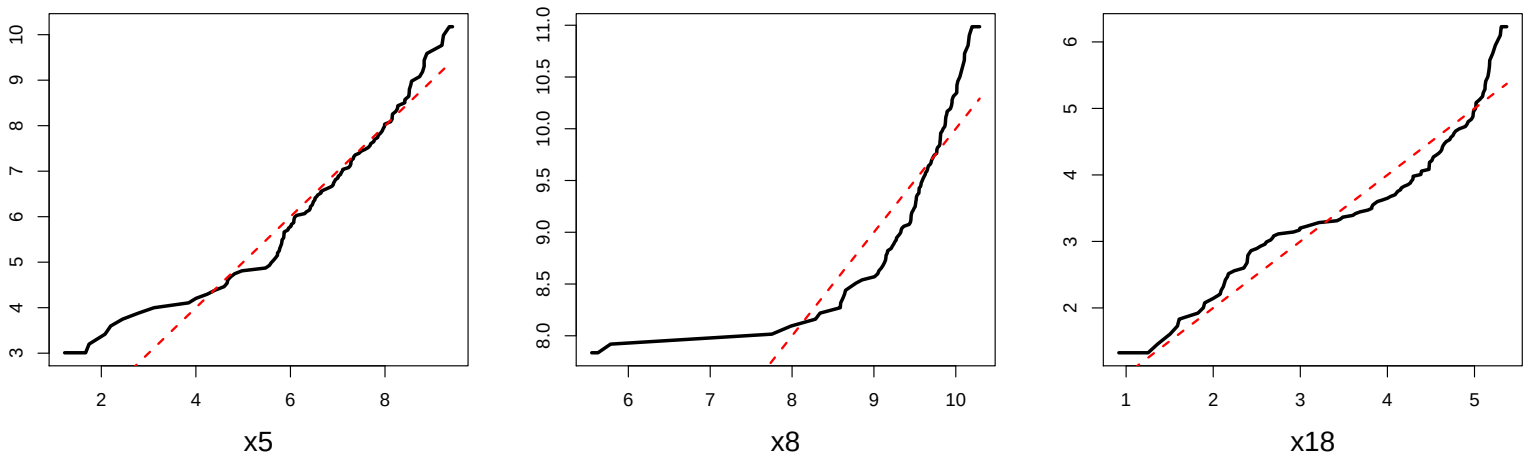








Transformations for 3 Genes



- These genes have highly non-Normal marginal distributions.
- The graphs are different at these genes.

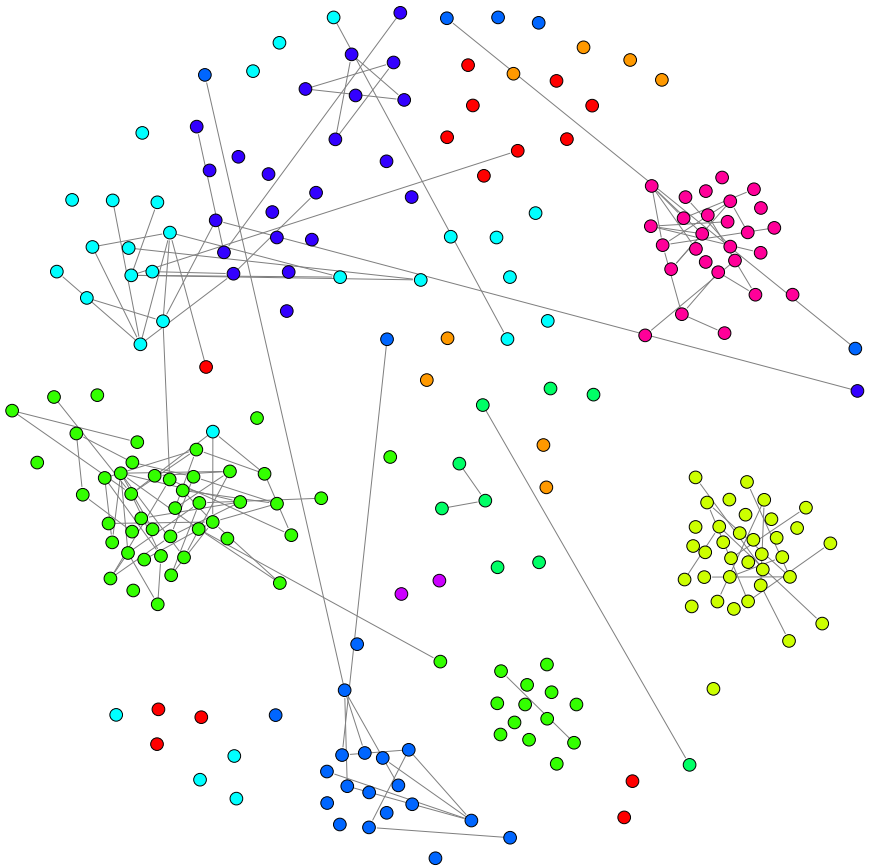
Graphs on the S&P 500

- Data from Yahoo! Finance (`finance.yahoo.com`).
- Daily closing prices for 452 stocks in the S&P 500 between 2003 and 2008 (before onset of the "financial crisis").
- Log returns $X_{tj} = \log\left(S_{t,j}/S_{t-1,j}\right)$.
- Winsorized to trim outliers.
- In following graphs, each node is a stock, and color indicates GICS industry.

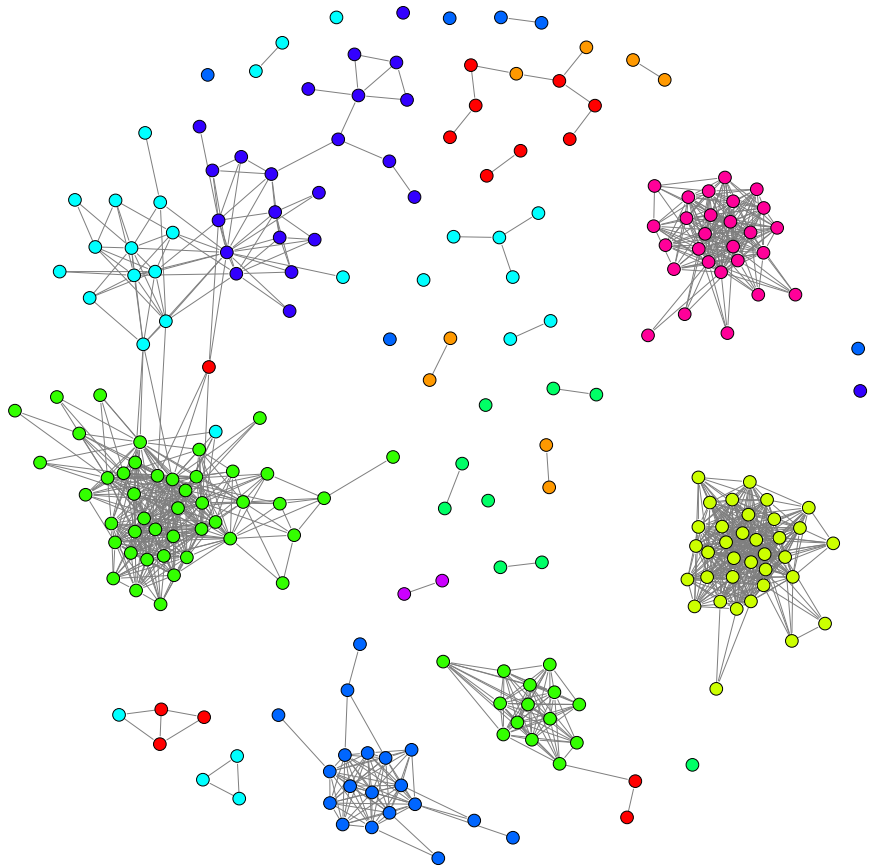
```
Consumer Discretionary          Consumer Staples
Energy                          Financials
Health Care                     Industrials
Information Technology          Materials
Telecommunications Services     Utilities
```

S&P Data: Glasso vs. Nonparanormal

difference



common



The Nonparanormal SKEPTIC

Liu, Han, Yuan, Lafferty & Wasserman, 2012

Assuming $X \sim NPN(f, \Sigma^0)$, we have

$$\Sigma^0_{jk} = 2 \sin\left(\frac{\pi}{6}\rho_{jk}\right)$$

where ρ is Spearman's rho:

$$\rho_{jk} := \text{Corr}\left(F_j(X_j), F_k(X_k)\right).$$

Empirical estimate:

$$\widehat{\rho}_{jk} = \frac{\sum_{i=1}^n (r_j^i - \bar{r}_j)(r_k^i - \bar{r}_k)}{\sqrt{\sum_{i=1}^n (r_j^i - \bar{r}_j)^2 \cdot \sum_{i=1}^n (r_k^i - \bar{r}_k)^2}}.$$

Similar relation holds for Kendall's tau.

The Nonparanormal SKEPTIC

Using a Hoeffding inequality for U-statistics, we get

$$\max_{jk} \left| \widehat{\Sigma}^{\rho}_{jk} - \Sigma^{0}_{jk} \right| \leq \frac{3\sqrt{2}\pi}{2} \sqrt{\frac{\log d + \log n}{n}},$$

with probability at least $1 - 1/n^2$.

Can thus estimate the covariance at the parametric rate

Punch line: *For graph and covariance estimation, no loss in statistical or computational efficiency comes from using Nonparanormal rather than Normal graphical model.*

Conclusions

- Thresholded backfitting algorithms derived from subdifferential calculus
- RKHS formulations are problematic
- Theory for infinite dimensional optimizations still incomplete
- Many extensions possible: Nonparanormal component analysis, etc.
- *Variations on additive models enjoy most of the good statistical and computational properties of sparse linear models, with relaxed assumptions*
- We're building a toolbox for large scale, high dimensional nonparametric inference.