Lecture 3

Nonparametric Methods

Statistical models with weak assumptions

Topics

- *Nonparametric regression*
- Sparse additive models
- Constrained rank additive models
- Nonparametric graphical models

Nonparametric Regression

Given $(X_1, Y_1), \ldots, (X_n, Y_n)$ predict Y from X.

Assume only that $Y_i = m(X_i) + \epsilon_i$ where where $m(x)$ is a smooth function of x.

The most popular methods are *kernel methods*. However, there are two types of kernels:

1. Smoothing kernels
2. Mercer kernels

Smoothing kernels involve local averaging.
Mercer kernels involve regularization.

Smoothing Kernels

- Smoothing kernel estimator:

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^n Y_i \, K_h(X_i, x)}{\sum_{i=1}^n K_h(X_i, x)}$$

where $K_h(x, z)$ is a *kernel* such as

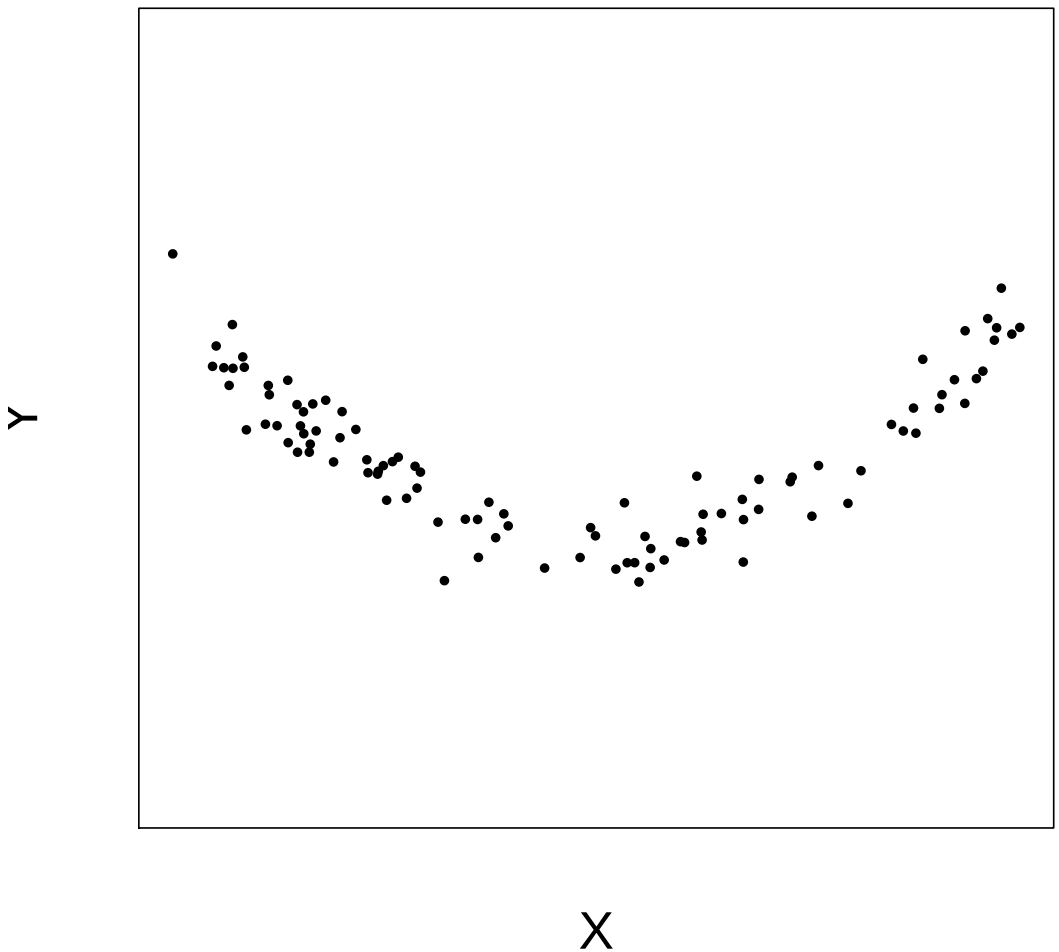
$$K_h(x, z) = \exp\left(-\frac{\|x - z\|^2}{2h^2}\right)$$

and $h > 0$ is called the *bandwidth*.

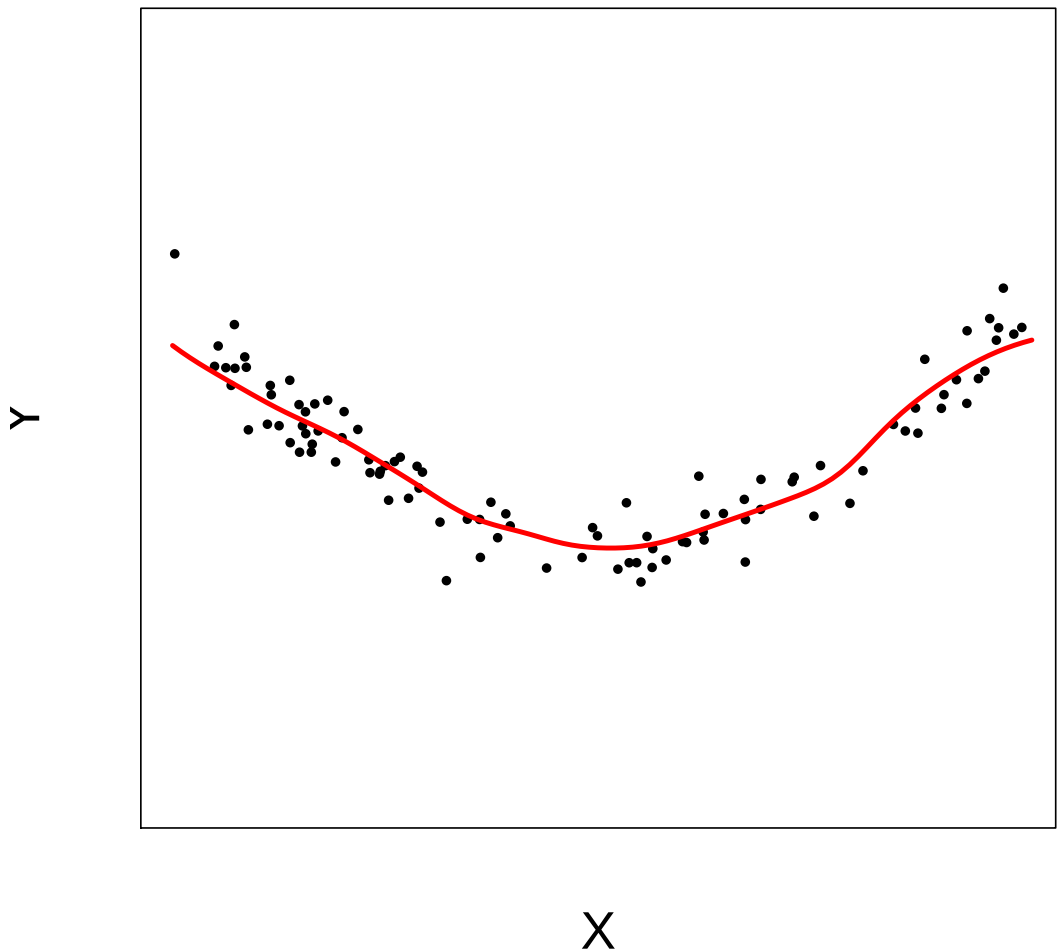
- $\widehat{m}_h(x)$ is just a local average of the Y_i's near x.

- The bandwidth h controls the bias-variance tradeoff:
Small h = large variance while *large h = large bias*.

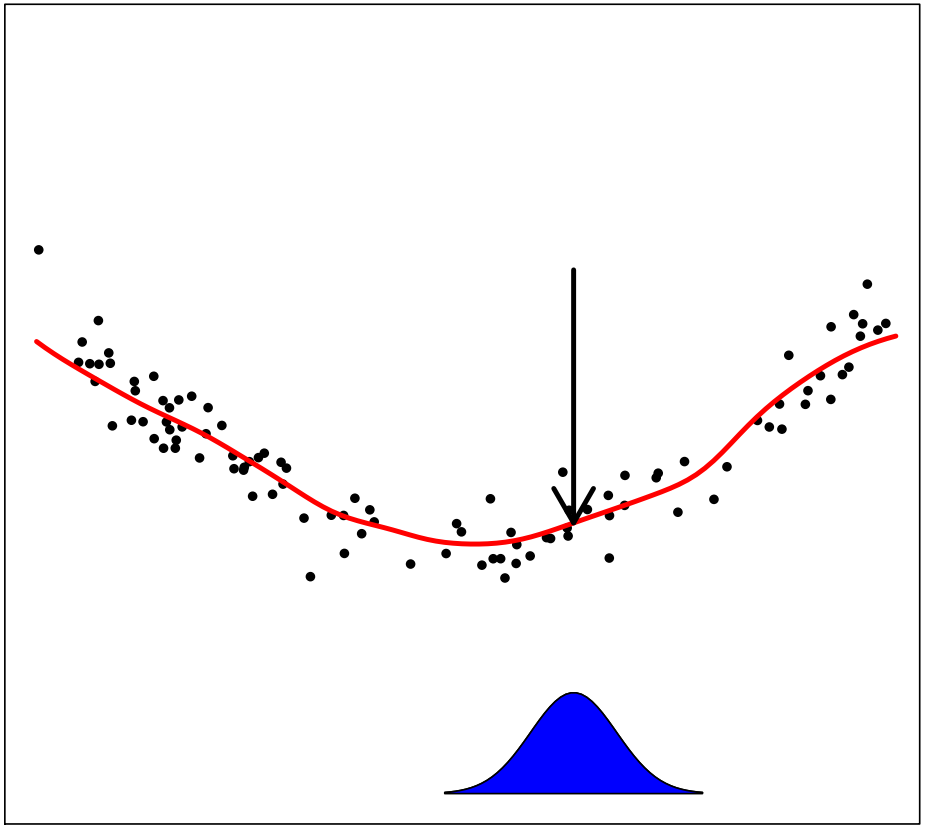
Example: Some Data – Plot of Y_i versus X_i



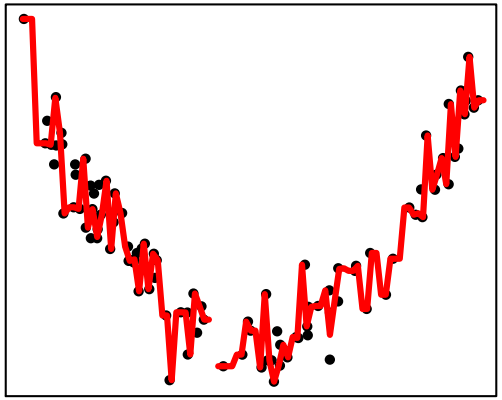
Example: $\widehat{m}(x)$



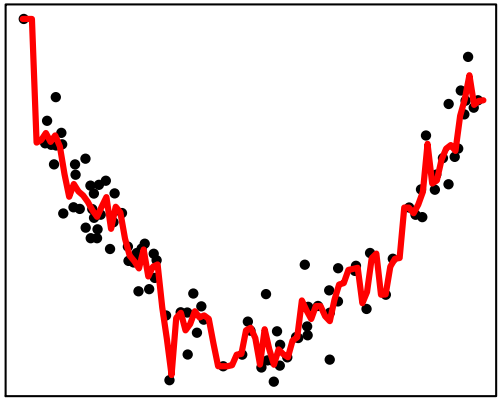
$\widehat{m}(x)$ is a local average



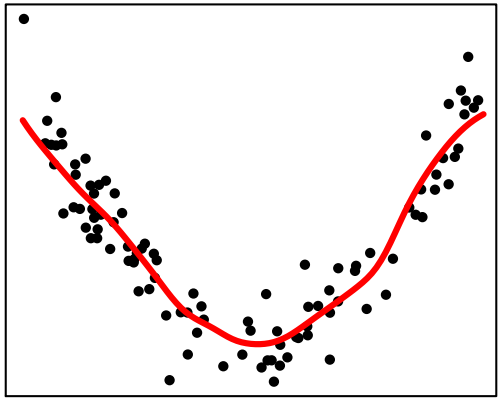
Effect of the bandwidth *h*



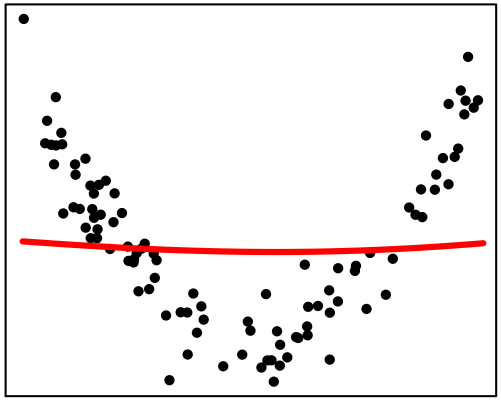
very small bandwidth



small bandwidth



medium bandwidth



large bandwidth

Smoothing Kernels

$$\text{Risk} = \mathbb{E}(Y - \widehat{m}_h(X))^2 = \text{bias}^2 + \text{variance} + \sigma^2.$$

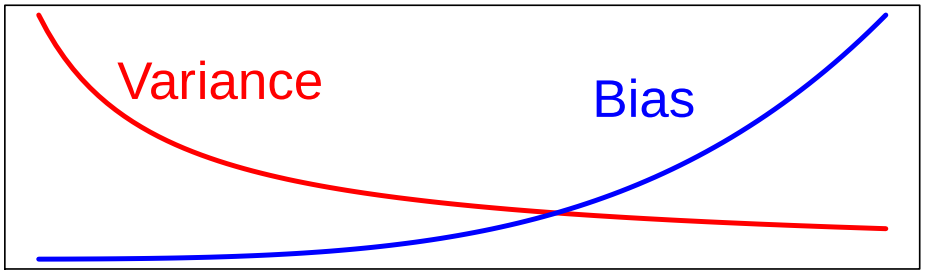
$$\text{bias}^2 \approx h^4,$$

$$\text{variance} \approx \frac{1}{nh^p} \quad \text{where } p = \text{dimension of } X.$$

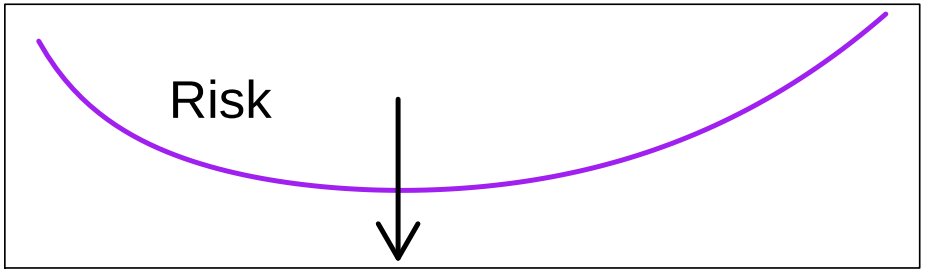
$\sigma^2 = \mathbb{E}(Y - m(X))^2$ is the unavoidable prediction error.

small h: low bias, high variance (undersmoothing)
large h: high bias, low variance (oversmoothing)

Risk Versus Bandwidth



h



optimal h

Estimating the Risk: Cross-Validation

To choose h we need to estimate the risk $R(h)$. We can estimate the risk by using *cross-validation*.

1. Omit (X_i, Y_i) to get $\widehat{m}_{h,(i)}$, then predict: $\widehat{Y}_{(i)} = \widehat{m}_{h,(i)}(X_i)$.
2. Repeat this for all observations.
3. The cross-validation estimate of risk is:

$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_{(i)})^2.$$

Shortcut formula:

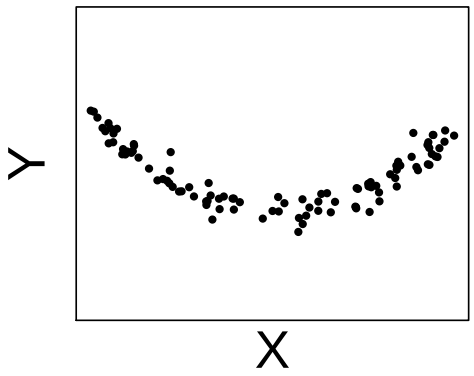
$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{Y}_i}{1 - L_{ii}}\right)^2$$

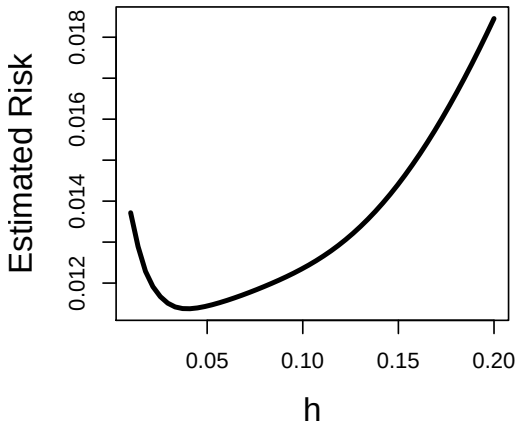
where $L_{ii} = K_h(X_i, X_i)/\sum_t K_h(X_i, X_t)$.

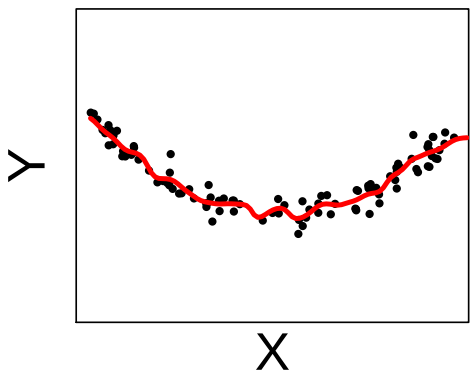
Summary so far

1. Compute $\widehat{m}_h$ for each h.
2. Estimate the risk $\widehat{R}(h)$.
3. Choose bandwidth $\widehat{h}$ to minimize $\widehat{R}(h)$.
4. Let $\widehat{m}(x) = \widehat{m}_{\widehat{h}}(x)$.

Example







Another Approach: Mercer Kernels

Instead of using local smoothing, we can optimize the fit to the data subject to regularization (penalization). Choose $\widehat{m}$ to minimize

$$\sum_i (Y_i - \widehat{m}(X_i))^2 + \lambda \text{ penalty(m)}$$

where penalty(m) is a *roughness penalty*.

λ is a smoothing parameter that controls the amount of smoothing.

How do we construct a penalty that measures roughness? One approach is: *Mercer Kernels* and *RKHS = Reproducing Kernel Hilbert Spaces.*

What is a Mercer Kernel?

A *Mercer Kernel* $K(x, y)$ is symmetric and positive definite:

$$\int \int f(x) f(y) K(x, y)\, dx\, dy \geq 0 \quad \text{for all } f.$$

Example: $K(x, y) = e^{-||x-y||^2/2}$.

Think of $K(x, y)$ as the *similarity* between x and y. We will create a set of *basis functions* based on K.

Fix z and think of $K(z, x)$ as a function of x. That is,

$$K(z, x) = K_z(x)$$

is a function of the second argument, with the first argument fixed.

Mercer Kernels

Let

$$\mathcal{F} = \left\{ f(\cdot) = \sum_{j=1}^{k} \beta_j K(z_j, \cdot) \right\}$$

Define a norm: $\|f\|_K = \sum_j \sum_k \beta_j \beta_k K(z_j, z_k)$. $\|f\|_K$ *small means f smooth.*

If $f = \sum_r \alpha_r K(z_r, \cdot)$, $g = \sum_s \beta_s K(w_s, \cdot)$, the inner product is

$$\langle f, g \rangle_K = \sum_r \sum_s \alpha_r \beta_s K(z_r, w_s).$$

$\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) because

$$\langle f, K(x, \cdot) \rangle = f(x)$$

Nonparametric Regression: Mercer Kernels

Representer Theorem: Let $\widehat{m}$ minimize

$$J = \sum_{i=1}^{n} (Y_i - m(X_i))^2 + \lambda \|m\|_K^2.$$

Then

$$\widehat{m}(x) = \sum_{i=1}^{n} \alpha_i \, K(X_i, x)$$

for some $\alpha_1, \ldots, \alpha_n$.

So, we only need to find the coefficients

$$\alpha = (\alpha_1, \ldots, \alpha_n).$$

Nonparametric Regression: Mercer Kernels

Plug $\widehat{m}(x) = \sum_{i=1}^{n} \alpha_i K(X_i, x)$ into J:

$$J = ||Y - \mathbb{K}\alpha||^2 + \lambda \alpha^T \mathbb{K} \alpha$$

where $\mathbb{K}_{jk} = K(X_j, X_k)$

Now we find α to minimize J. We get: $\widehat{\alpha} = (\mathbb{K} + \lambda I)^{-1} Y$ and $\widehat{m}(x) = \sum_i \widehat{\alpha}_i K(X_i, x)$.

The estimator depends on the amount of regularization λ. Again, there is a bias-variance tradeoff. We choose λ by cross-validation. This is like the bandwidth in smoothing kernel regression.

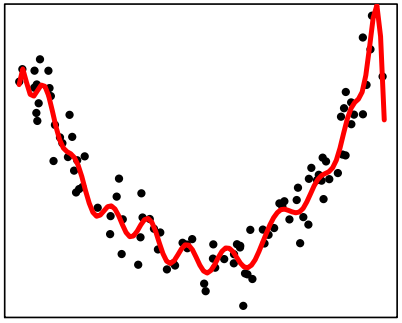
Smoothing Kernels *Versus* Mercer Kernels

Smoothing kernels: the bandwidth h controls the amount of smoothing.

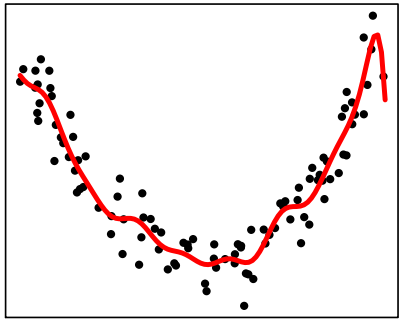
Mercer kernels: norm $\|f\|_K$ controls the amount of smoothing.

In practice these two methods give answers that are very similar.

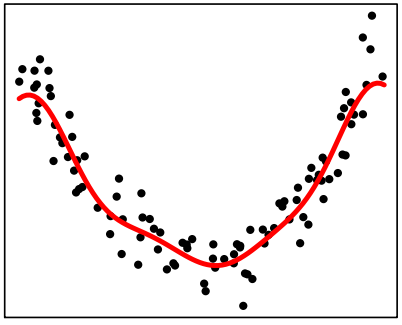
Mercer Kernels: Examples



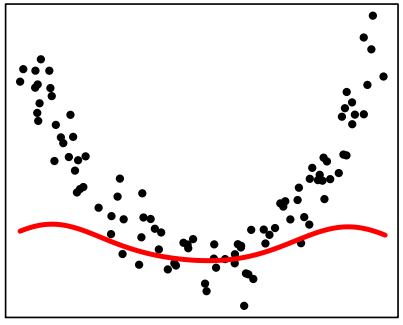
very small lambda



small lambda



medium lambda



large lambda

Multiple Regression

Both methods extend easily to the case where X has dimension $p > 1$. For example, just use

$$K(x, y) = e^{-\|x-y\|^2/2}.$$

However, this is hard to interpret and is subject to the curse of dimensionality. This means that the *statistical performance* and the *computational complexity* degrade as dimension p increases.

An alternative is to use something less nonparametric such as additive model where we restrict $m(x_1, \ldots, x_p)$ to be of the form:

$$m(x_1, \ldots, x_p) = \beta_0 + \sum_j m_j(x_j).$$

Topics

- Nonparametric regression
- *Sparse additive models*
- Nonparametric graphical models

Additive Models

Model: $m(x) = \beta_0 + \sum_{j=1}^{p} m_j(x_j)$.

We can take $\widehat{\beta}_0 = \overline{Y}$ and we will ignore β_0 from now on.

We want to minimize

$$\sum_{i=1}^{n} \left(Y_i - \left(m_1(X_{i1}) + \cdots + m_p(X_{ip}) \right) \right)^2$$

subject to m_j smooth.

Additive Models

The backfitting algorithm:

- Set $\widehat{m}_j = 0$
- Iterate until convergence:
  - Iterate over j:
    - $R_i = Y_i - \sum_{k \neq j} \widehat{m}_k(X_{ik})$
    - $\widehat{m}_j \longleftarrow \text{smooth}(X_j, R)$

Here, $\text{smooth}(X_j, R)$ is any one-dimensional nonparametric regression function.

R: glm

But what if p is large?

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

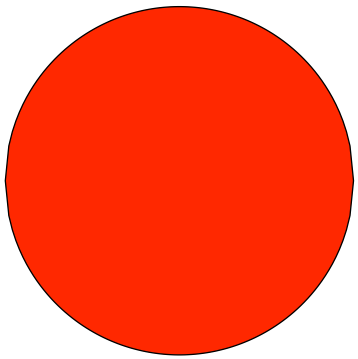
subject to $\sum_{j=1}^{p} \sqrt{\mathbb{E}(m_j^2)} \leq L_n$

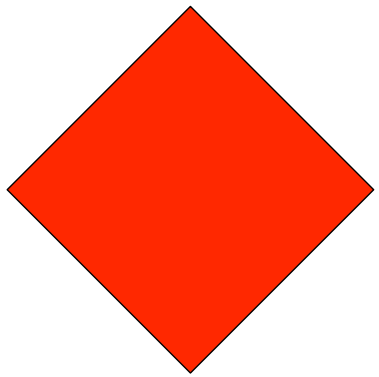
$\mathbb{E}(m_j) = 0$

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

Sparse Additive Models

$$C = \left\{ m \in \mathbb{R}^4 \ : \ \sqrt{m_1(x_1)^2 + m_1(x_2)^2} + \sqrt{m_2(x_1)^2 + m_2(x_2)^2} \leq L \right\}$$

$\pi_{12} C =$ 

$\pi_{13} C =$ 

Stationary Conditions

Lagrangian

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

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; $n = 506$, $p = 30$.

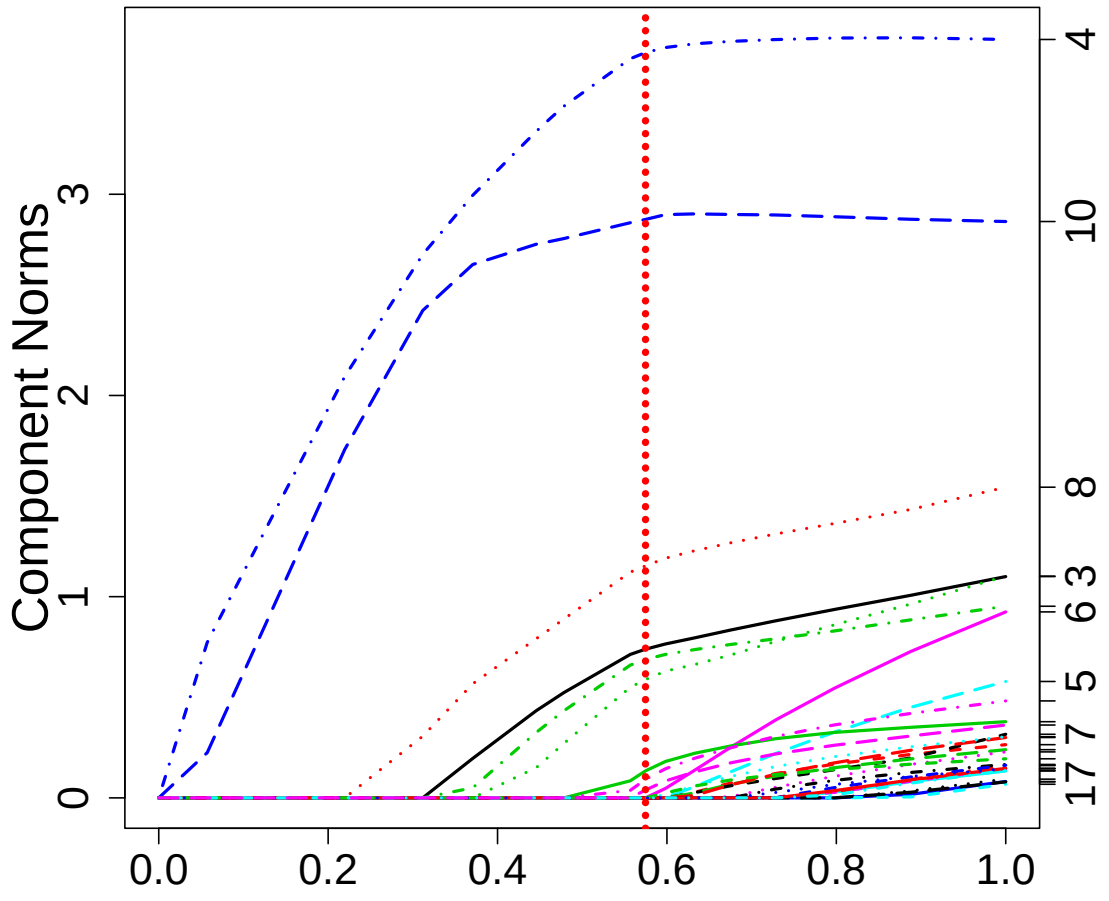
$$Y = \beta_0 + m_1(\text{crime}) + m_2(\text{tax}) + \cdots + \cdots m_{30}(X_{30}) + \epsilon.$$

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

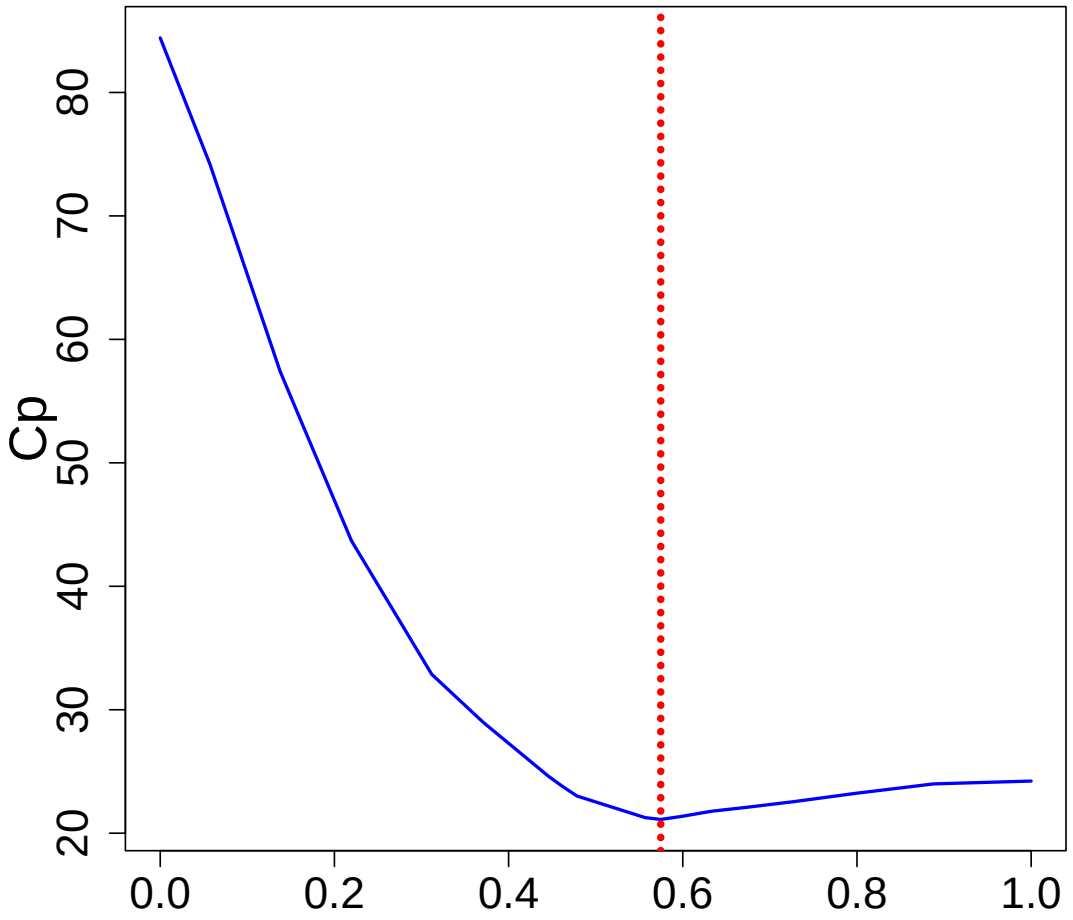
We choose λ by minimizing the estimated risk.

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

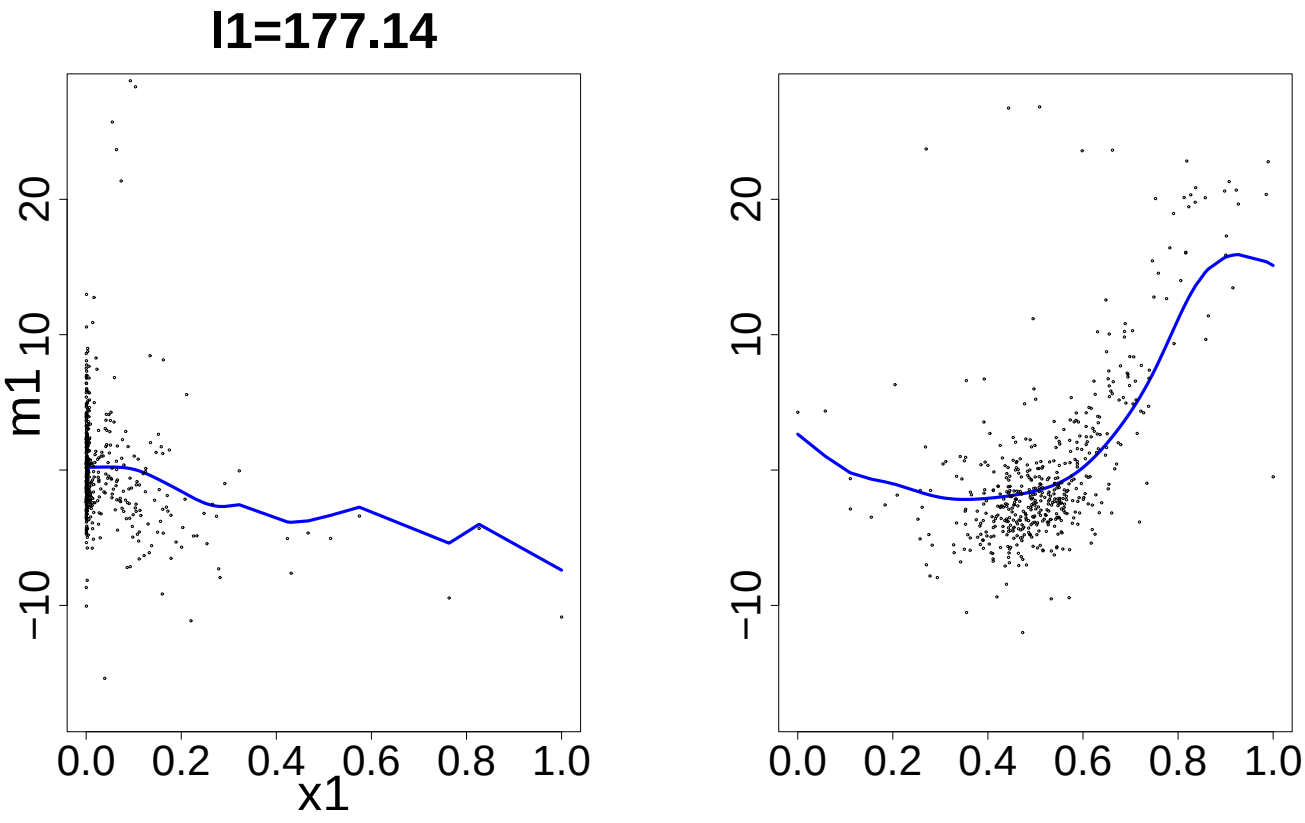
L_2 norms of fitted functions versus $1/\lambda$



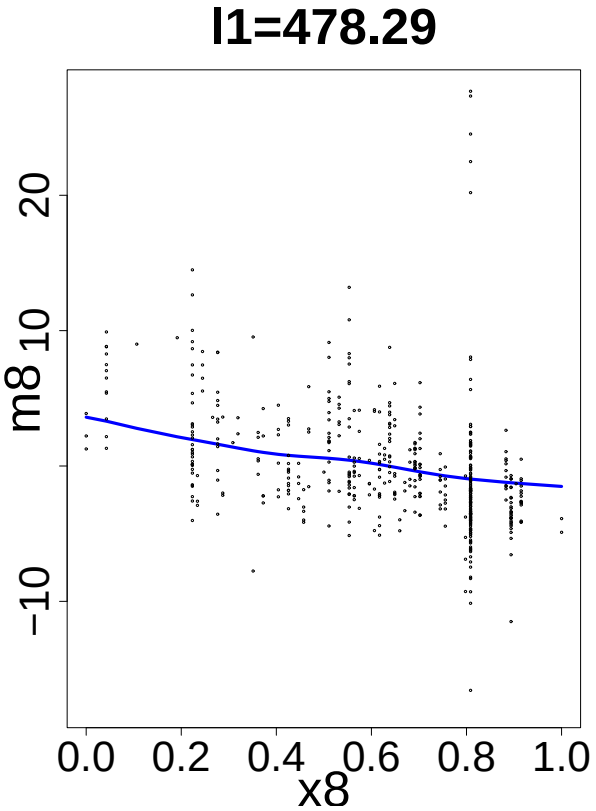
Estimated Risk Versus λ

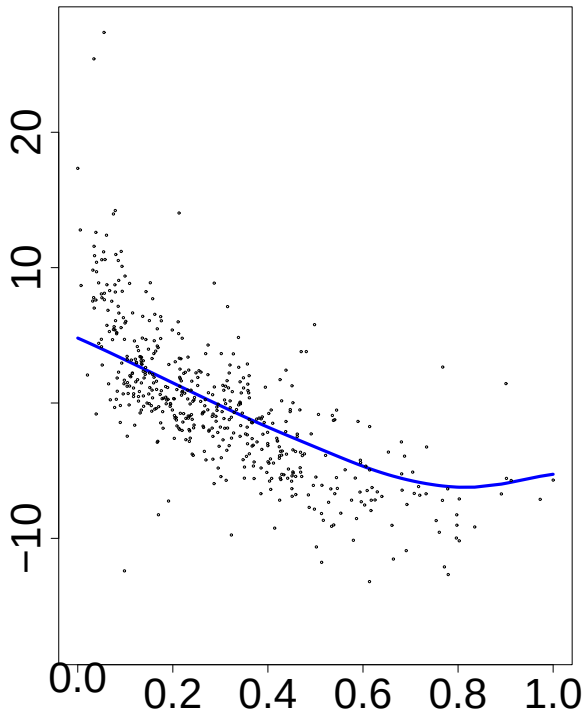


Example Fits



Example Fits





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?

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

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

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

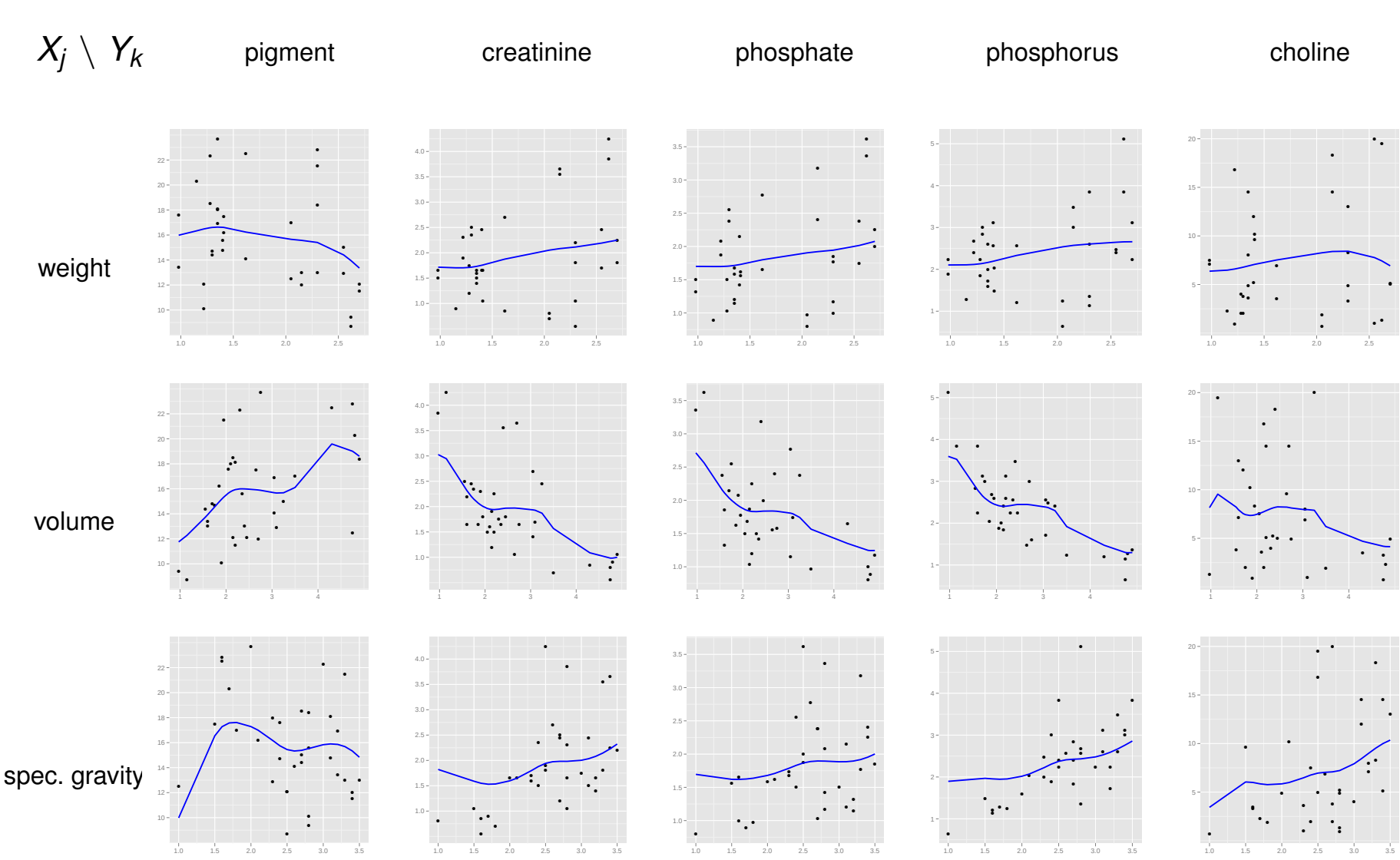
Data of Smith et al. (1962), chemical measurements for 33 individual urine specimens.

$q = 5$ 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.

Topics

- Nonparametric regression
- Sparse additive models
- *Nonparametric graphical models*

Regression vs. Graphical Models

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

The Nonparanormal (Liu, Lafferty, Wasserman, 2009)

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

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

in case

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

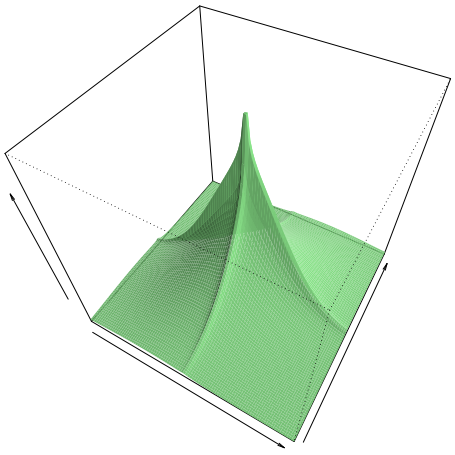
where $f(X) = (f_1(X_1), \ldots, f_p(X_p))$.

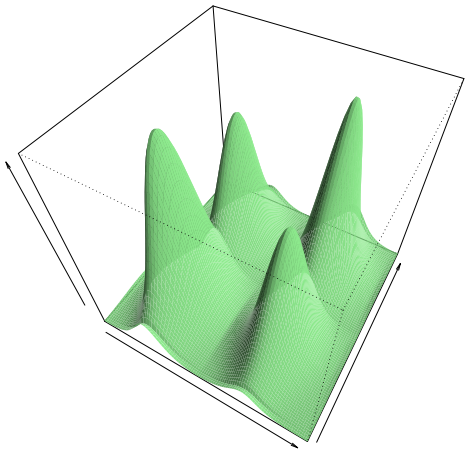
Joint density

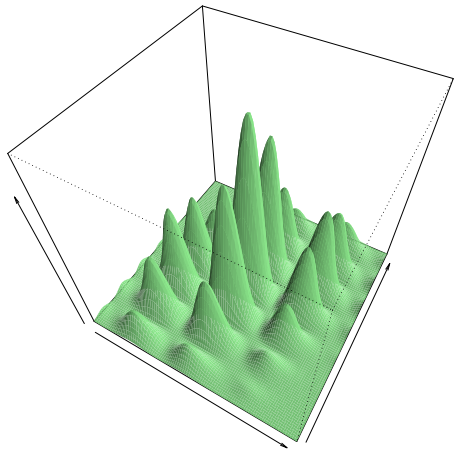
$$p_X(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(f(x)-\mu)^T \Sigma^{-1}(f(x)-\mu)\right\} \prod_{j=1}^{p} |f_j'(x_j)|$$

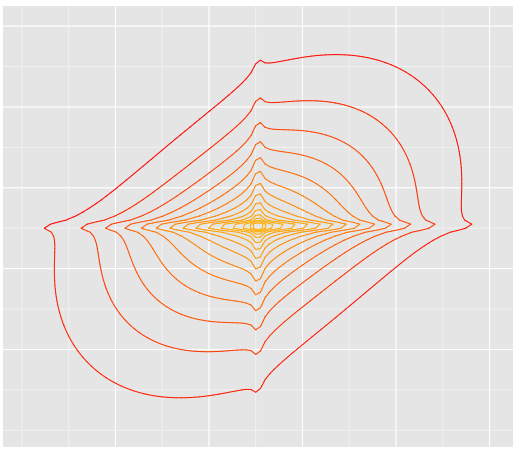
- Semiparametric Gaussian copula

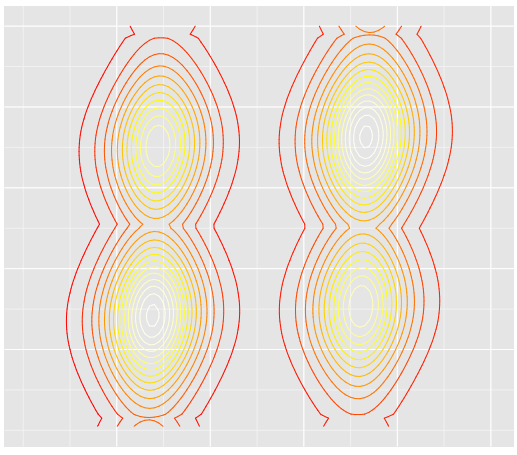
Examples

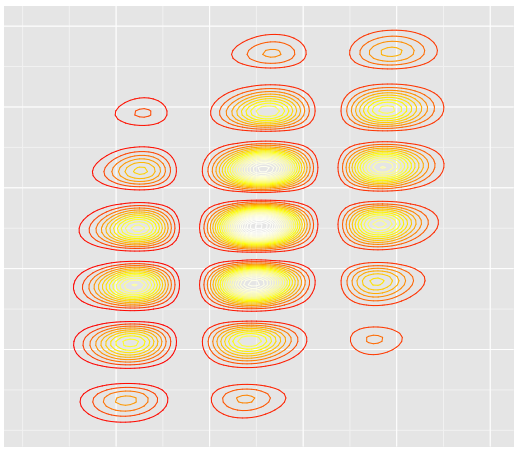












The Nonparanormal

- 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

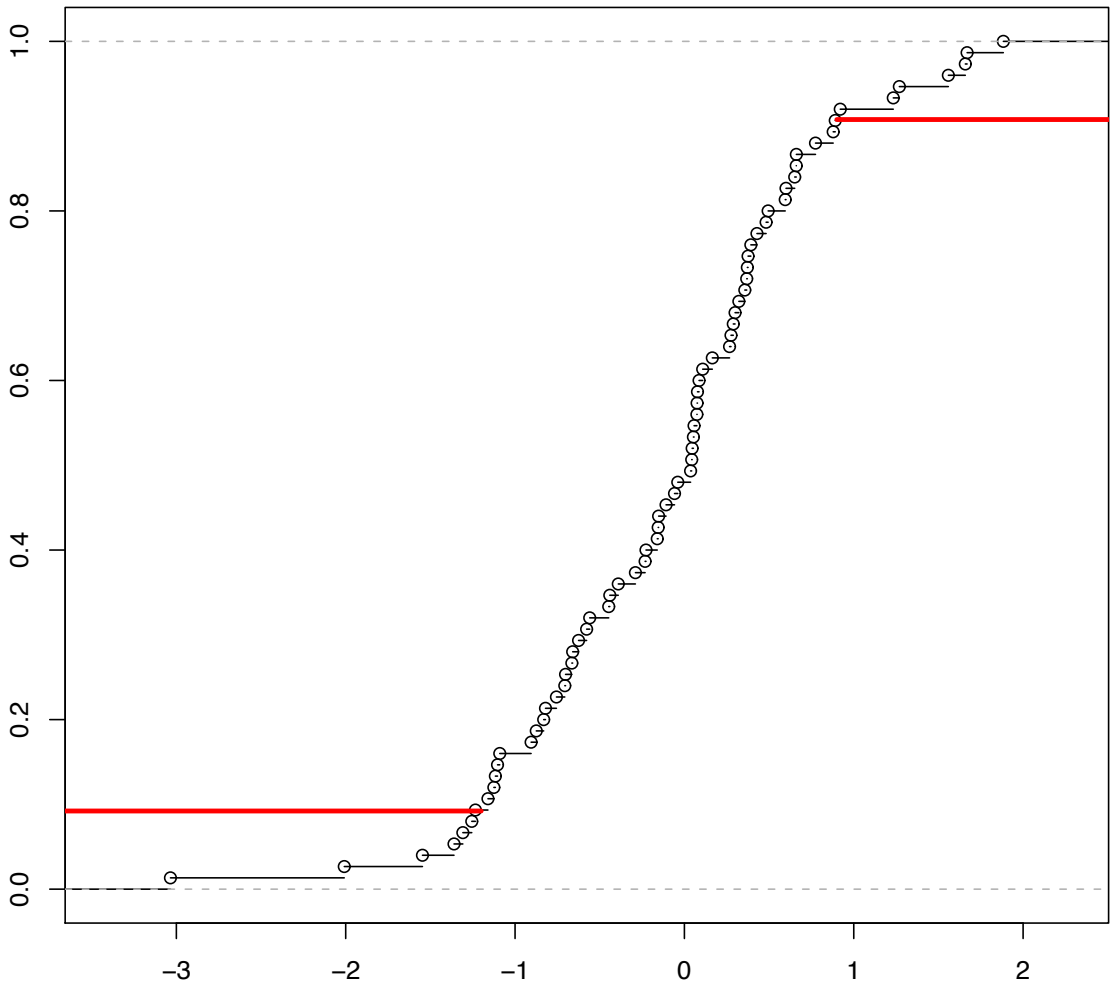
$$X_j \amalg X_k \,\Big|\, X_{\text{rest}}$$

  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.

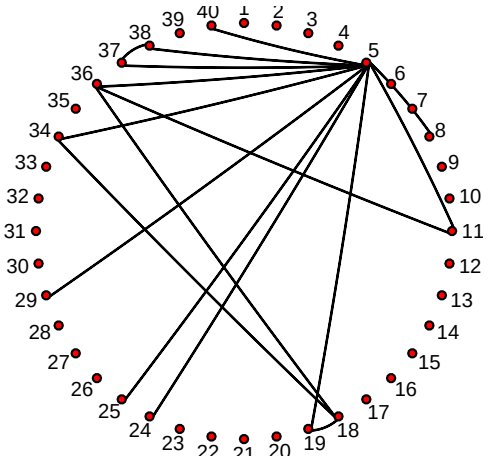
Gene-Gene Interactions for *Arabidopsis thaliana*

source: wikipedia.org

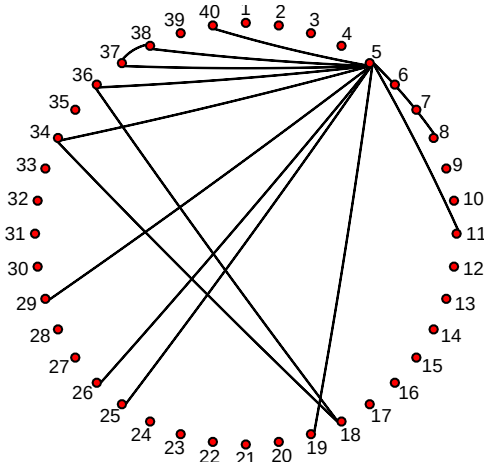
Dataset from Affymetrix microarrays, sample size $n = 118$, $p = 40$ genes (isoprenoid pathway).

Example Results

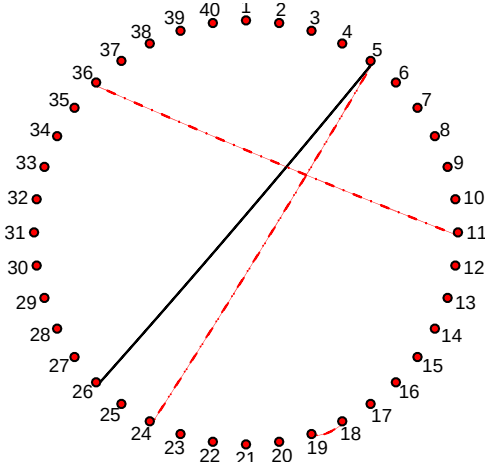
NPN

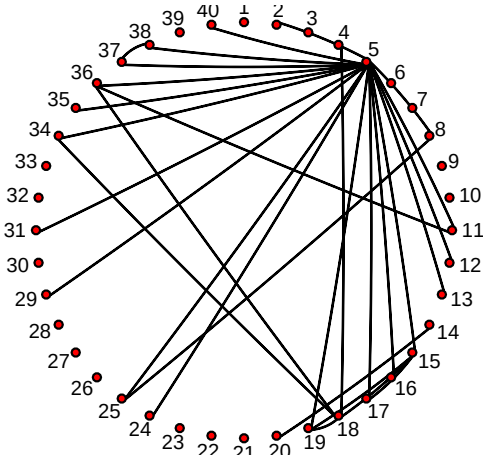


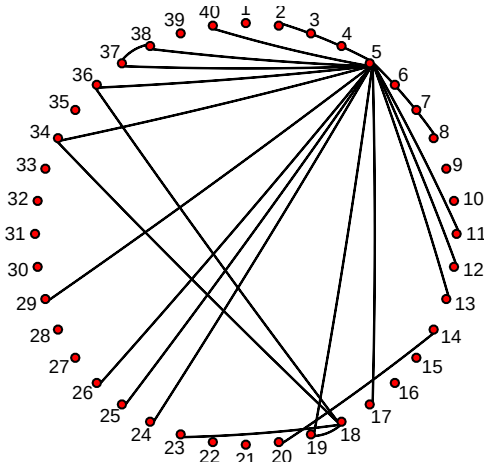
glasso

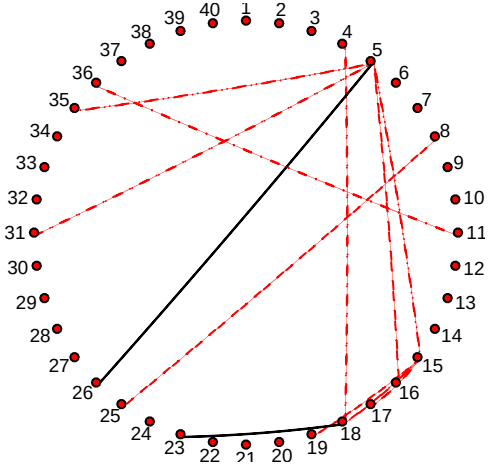


difference

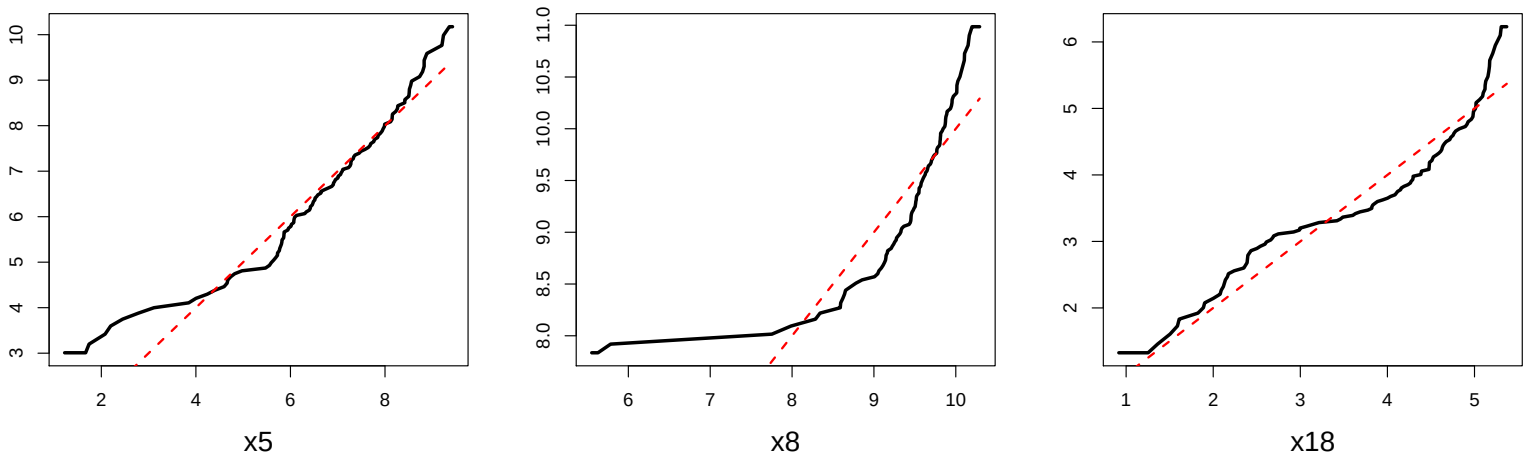






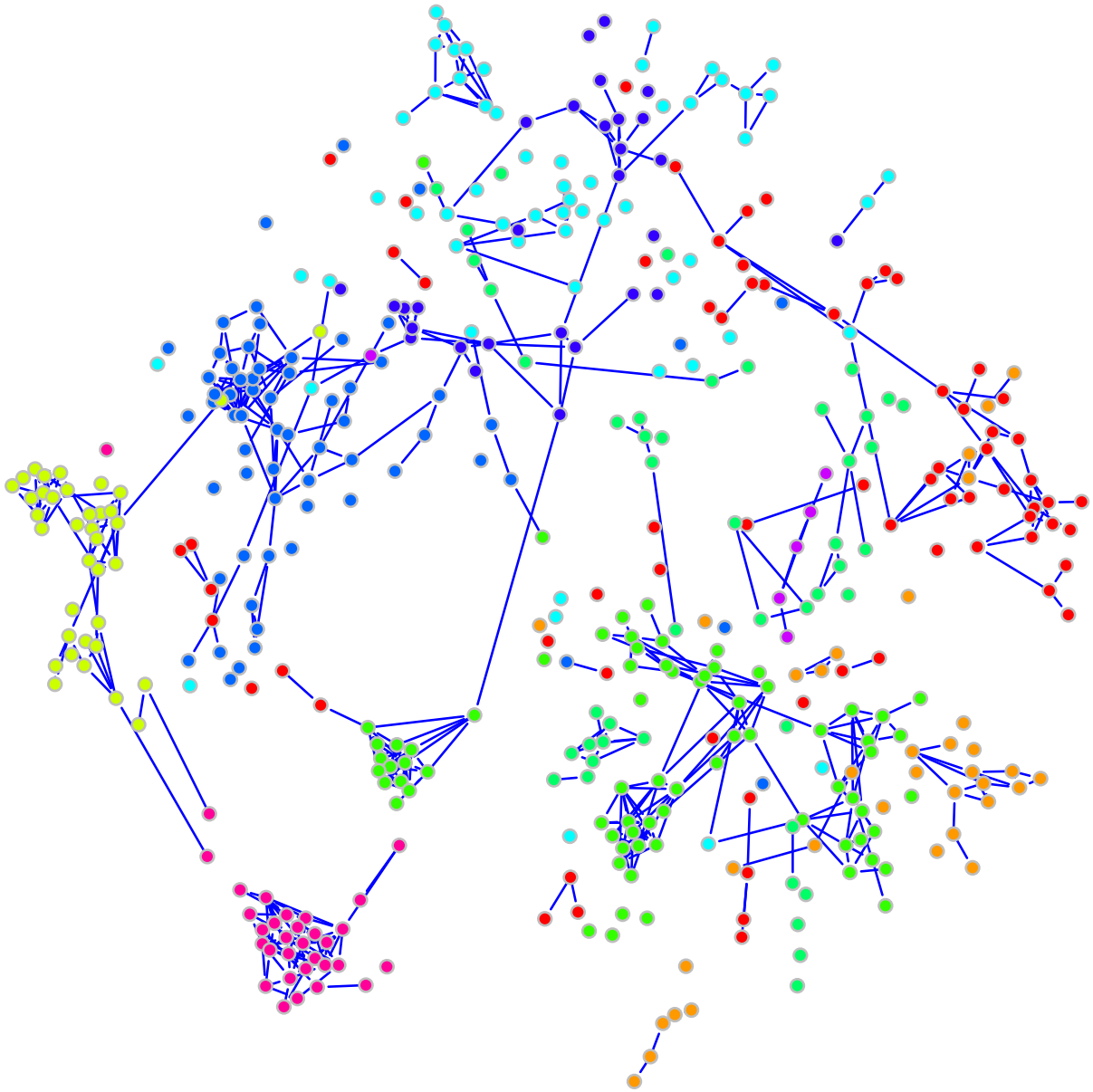


Transformations for 3 Genes

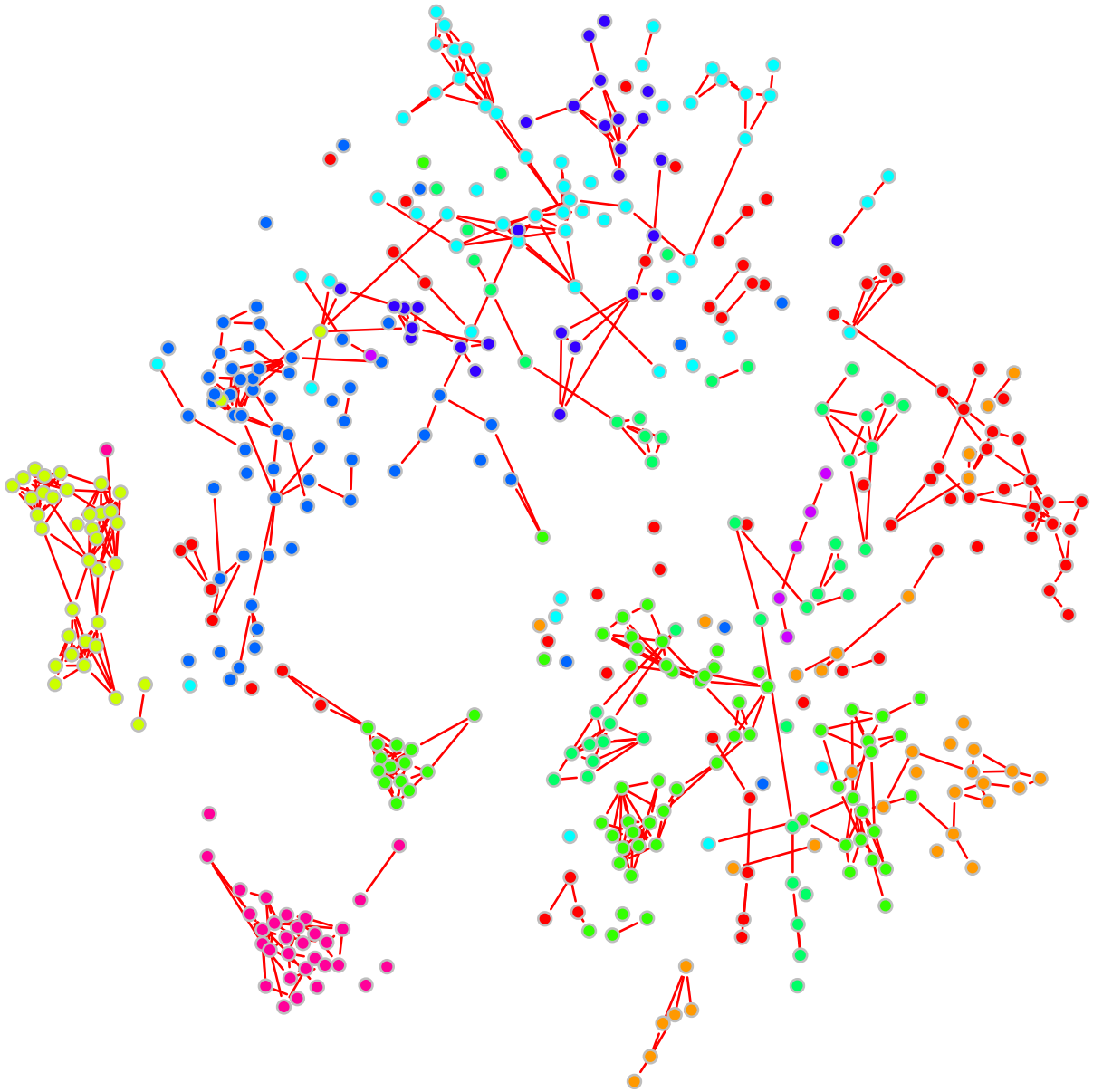


- These genes have highly non-Normal marginal distributions.
- The graphs are different at these genes.

S&P Data (2003–2008): Graphical Lasso

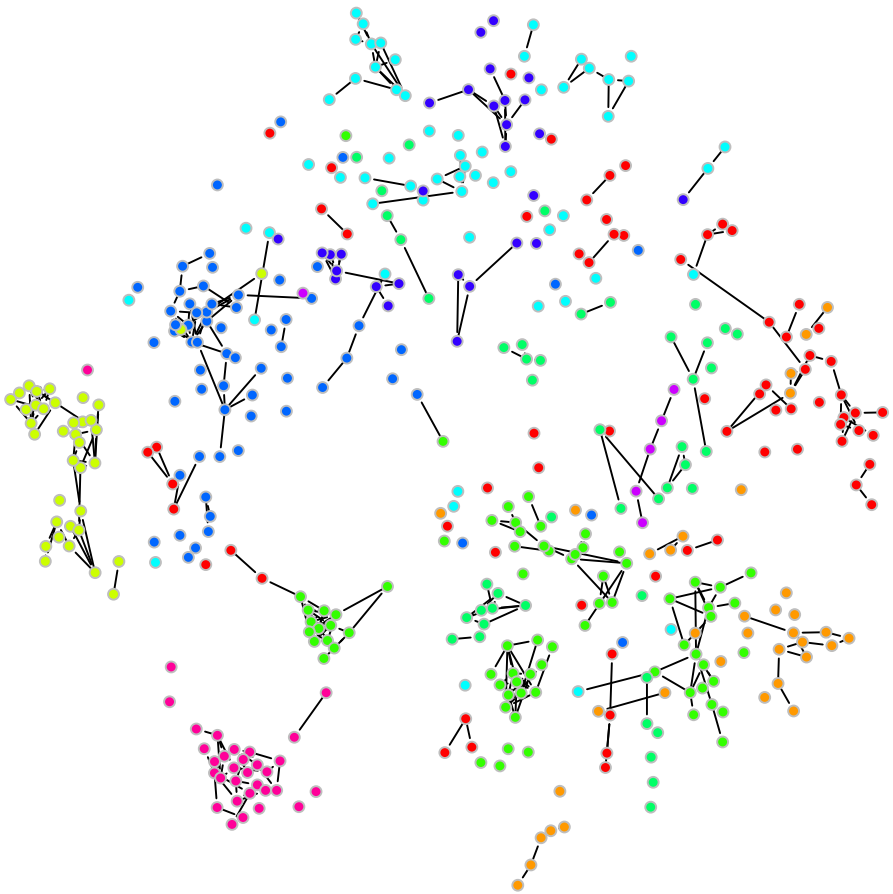


S&P Data: Nonparanormal

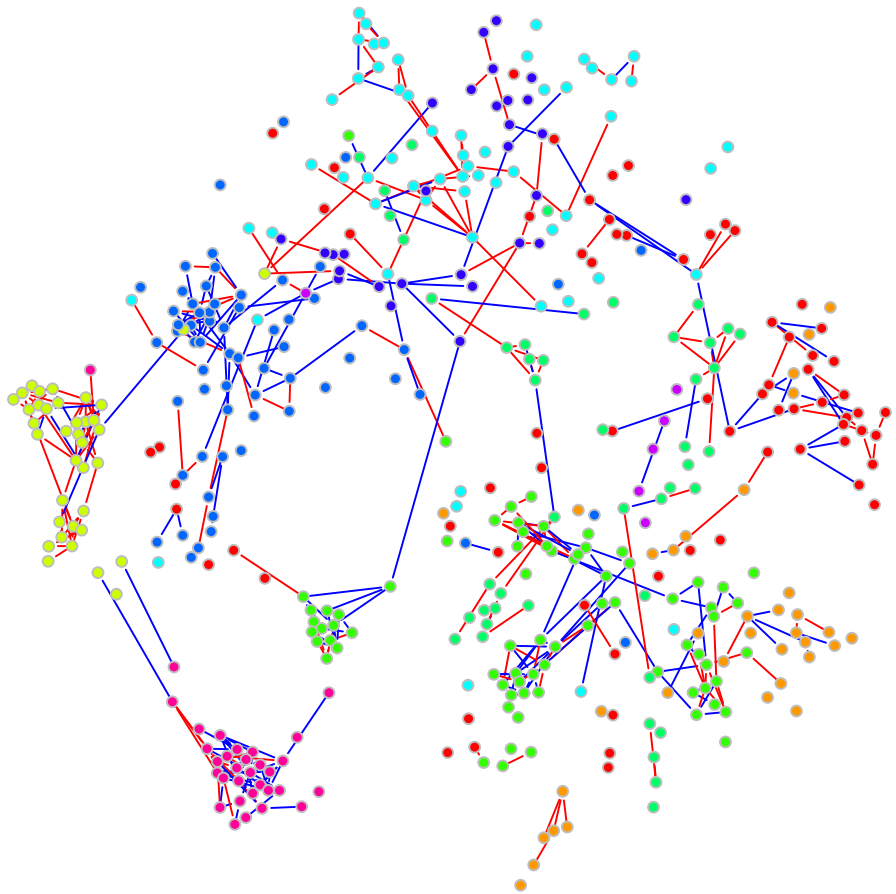


S&P Data: Nonparanormal vs. Glasso

common edges



differences



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.*

Graph-Valued Regression

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ where Y_i is high-dimensional
- We'll discuss one particular version: *graph-valued regression* (Chen, Lafferty, Liu, Wasserman, 2010)
- Let $G(x)$ be the graph for Y based on $p(y|x)$
- This defines a partition $\mathcal{X}_1, \ldots, \mathcal{X}_k$ where $G(x)$ is constant over each partition.
- Three methods to find $G(x)$:
  - Parametric
  - Kernel graph-valued regression
  - GO-CART (Graph-Optimized CART)

Graph-Valued Regression

multivariate regression
(supervised)
$\mu(x) = \mathbb{E}(Y \mid x)$
$Y \in \mathbb{R}^p,\ x \in \mathbb{R}^q$

graphical model
(unsupervised)
$\text{Graph}(Y) = (V, E)$
$(j,k) \notin E \iff Y_j \amalg Y_k \mid Y_{\text{rest}}$

graph-valued regression
$\text{Graph}(Y \mid x)$

- Gene associations from phenotype (or *vice versa*)
- Voting patterns from covariates on bills
- Stock interactions given market conditions, news items

Method I: Parametric

- Assume that $Z = (X, Y)$ is jointly multivariate Gaussian.
- $\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix}$.
- Get $\widehat{\Sigma}_X$, $\widehat{\Sigma}_Y$, and $\widehat{\Sigma}_{XY}$
- Get Ω_X by the glasso.
- $\widehat{\Sigma}_{Y|X} = \widehat{\Sigma}_Y - \widehat{\Sigma}_{YX}\widehat{\Omega}_X\widehat{\Sigma}_{XY}$.
- But, the estimated graph does not vary with different values of X.

Method II: Kernel Smoothing

- $Y|X = x \sim N(\mu(x), \Sigma(x))$.

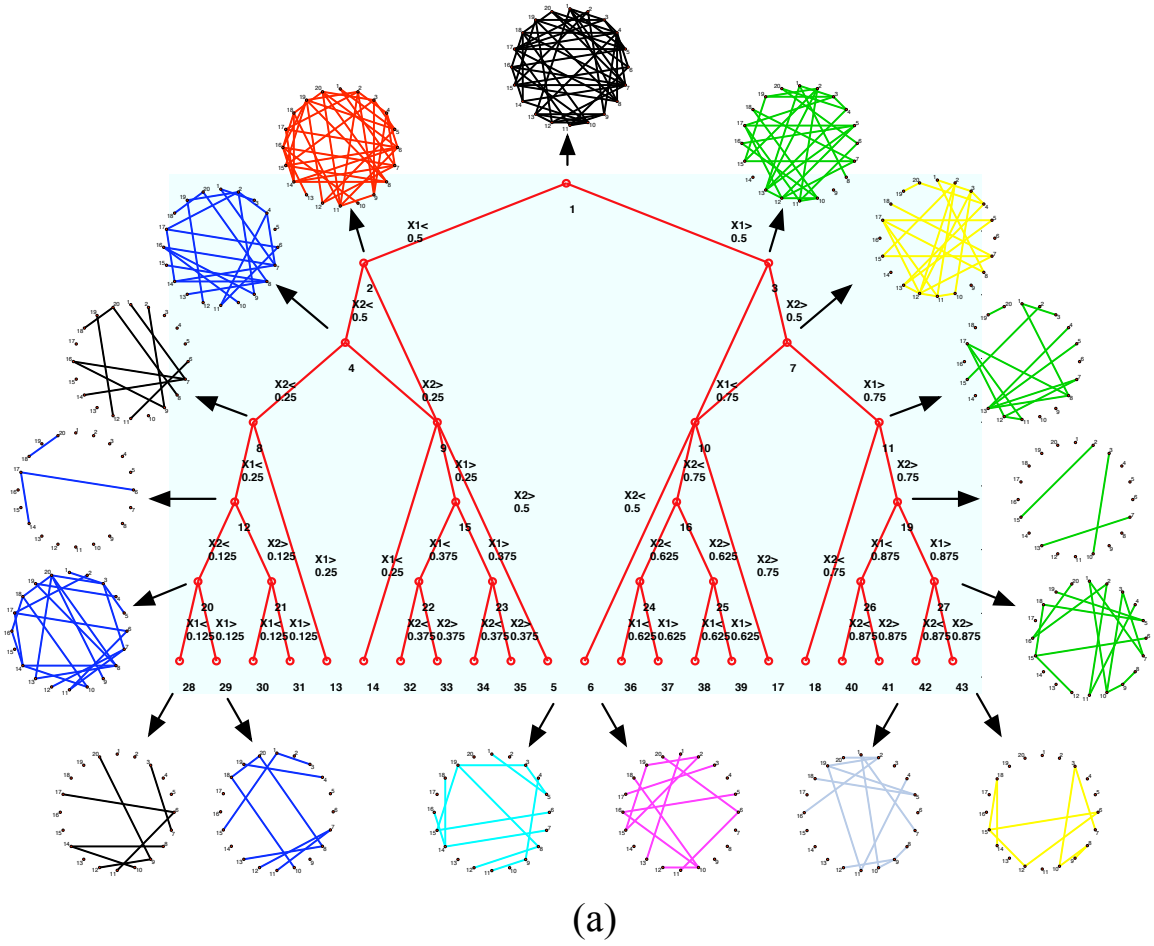
$$\begin{aligned}
\widehat{\Sigma}(x) &= \frac{\sum_{i=1}^{n} K\left(\frac{\|x-x_i\|}{h}\right)(y_i - \widehat{\mu}(x))(y_i - \widehat{\mu}(x))^T}{\sum_{i=1}^{n} K\left(\frac{\|x-x_i\|}{h}\right)} \\
\widehat{\mu}(x) &= \frac{\sum_{i=1}^{n} K\left(\frac{\|x-x_i\|}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{\|x-x_i\|}{h}\right)}.
\end{aligned}$$

- Apply glasso to $\widehat{\Sigma}(x)$
- Easy to do but recovering $\mathcal{X}_1, \ldots, \mathcal{X}_k$ requires difficult post-processing.

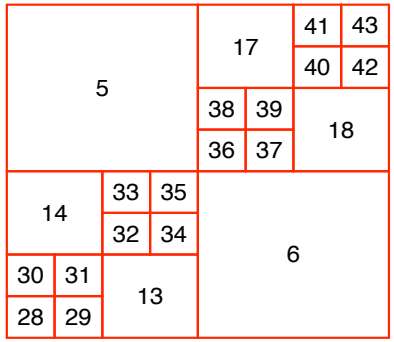
Method III: Partition Estimator

- Run CART but use Gaussian log-likelihood (on held out data) to determine the splits
- This yields a partition $\mathcal{X}_1, \ldots, \mathcal{X}_k$ (and a correspdonding tree)
- Run the glasso within each partition element

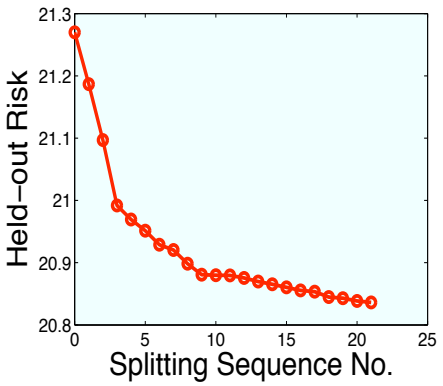
Simulated Data



(a)

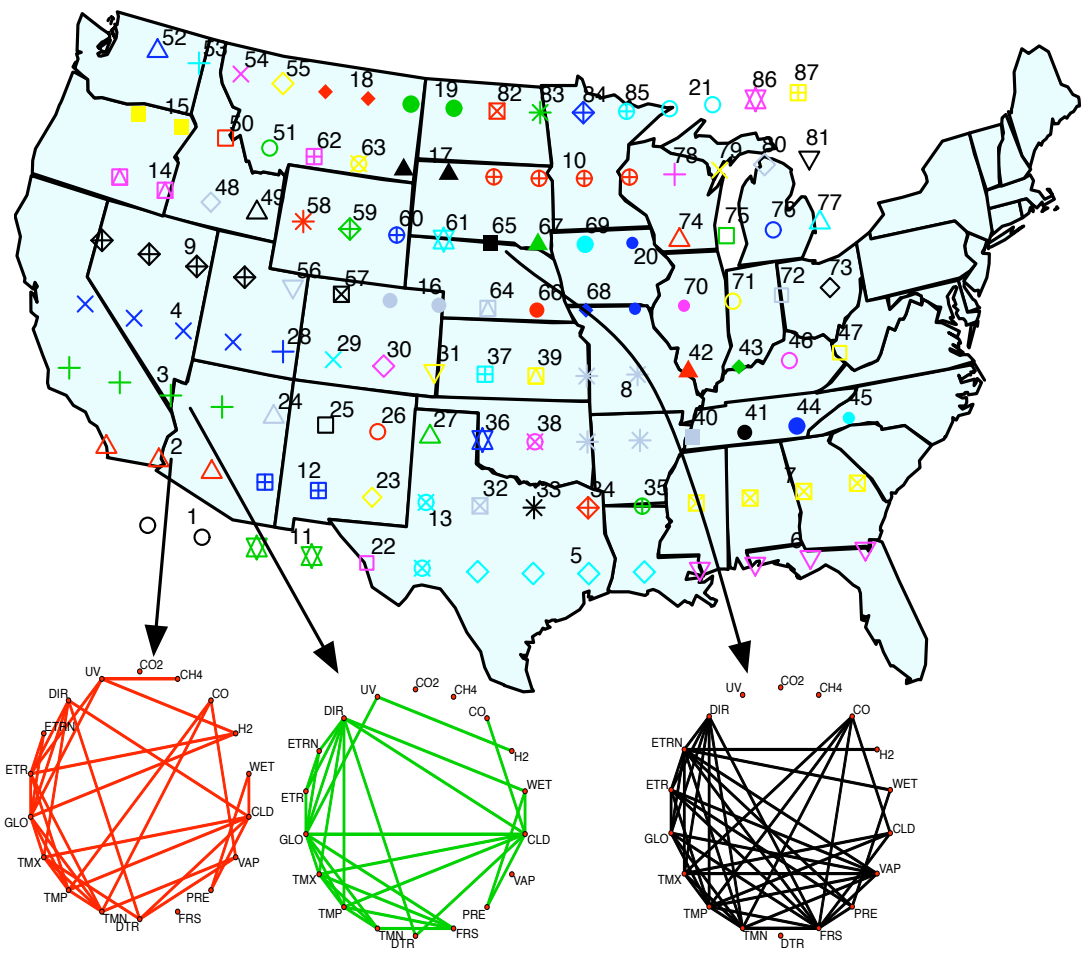


(b)

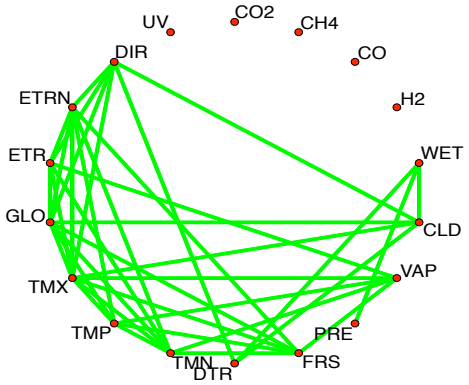


(c)

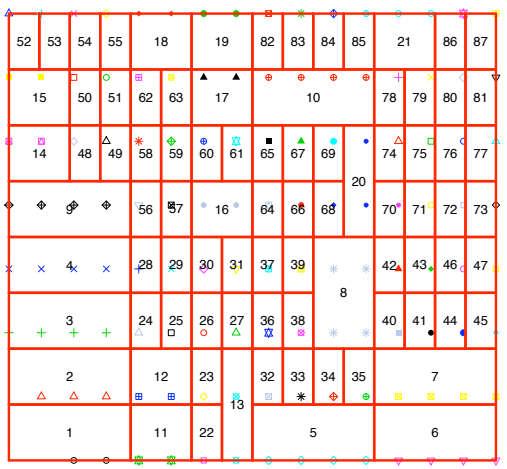
Climate Data



(a)



(b)



(c)

Tradeoff

- Nonparanormal: Unrestricted graphs, semiparametric
- We'll now trade off structural flexibility for greater nonparametricity

Forest Densities (Gupta, Lafferty, Liu, Wasserman, Xu, 2011)

A distribution is *supported by a forest F with edge set $E(F)$* if

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

- For known marginal densities $p(x_i, x_j)$, best tree obtained by minimum weight spanning tree algorithms.
- In high dimensions, a spanning tree will overfit.
- We prune back to a forest.

Step 1: Constructing a Full Tree

- Compute kernel density estimates

$$\widehat{f}_{n_1}(x_i, x_j) = \frac{1}{n_1} \sum_{s \in \mathcal{D}_1} \frac{1}{h_2^2} K\left(\frac{X_i^{(s)} - x_i}{h_2}\right) K\left(\frac{X_j^{(s)} - x_j}{h_2}\right)$$

- Estimate mutual informations

$$\widehat{I}_{n_1}(X_i, X_j) = \frac{1}{m^2} \sum_{k=1}^{m} \sum_{\ell=1}^{m} \widehat{f}_{n_1}(x_{ki}, x_{\ell j}) \log \frac{\widehat{f}_{n_1}(x_{ki}, x_{\ell j})}{\widehat{f}_{n_1}(x_{ki})\, \widehat{f}_{n_1}(x_{\ell j})}$$

- Run Kruskal's algorithm (Chow-Liu) on edge weights

Step 2: Pruning the Tree

- Heldout risk

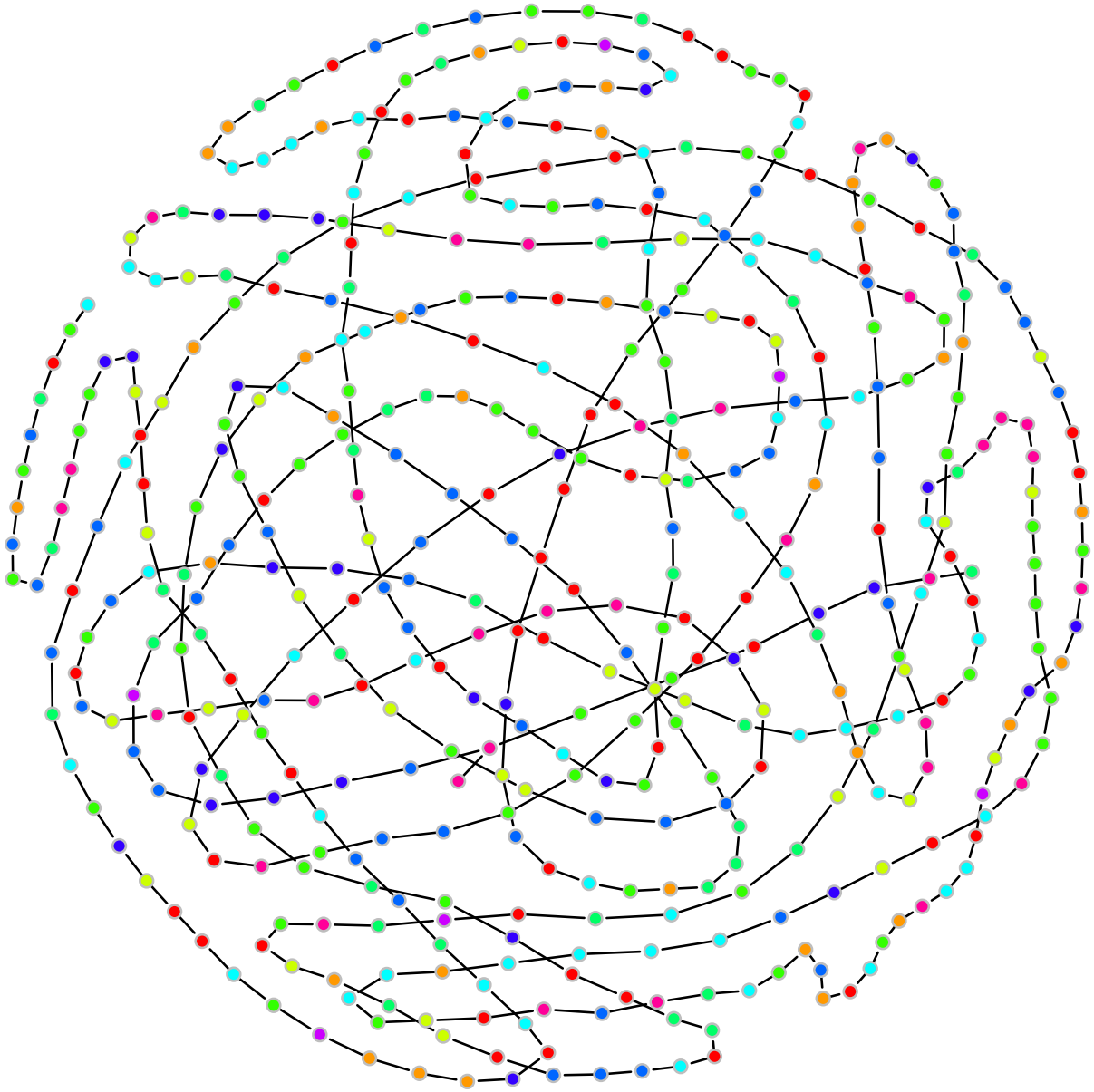
$$\widehat{R}_{n_2}(f_F) = -\sum_{(i,j)\in E} \int \widehat{f}_{n_2}(x_i, x_j) \log \frac{f(x_i, x_j)}{f(x_i)\, f(x_j)}\, dx_i dx_j$$

- Selected forest given by

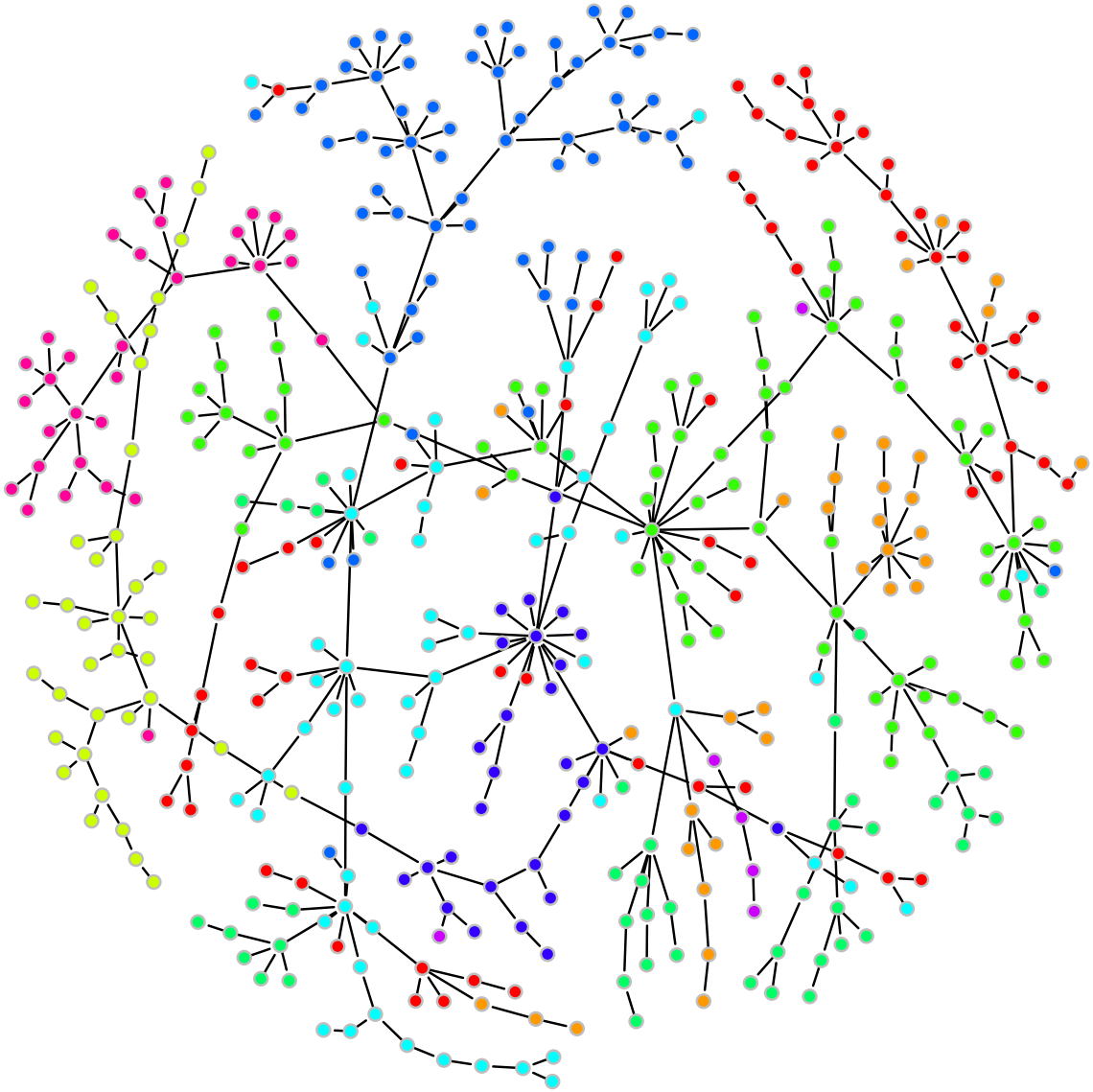
$$\widehat{k} = \underset{k\in\{0,\ldots,p-1\}}{\arg\min}\ \widehat{R}_{n_2}\left(\widehat{f}_{\widehat{T}^{(k)}_{n_1}}\right)$$

where $\widehat{T}^{(k)}_{n_1}$ is forest obtained after k steps of Kruskal

S&P Data: Forest Graph—Oops!

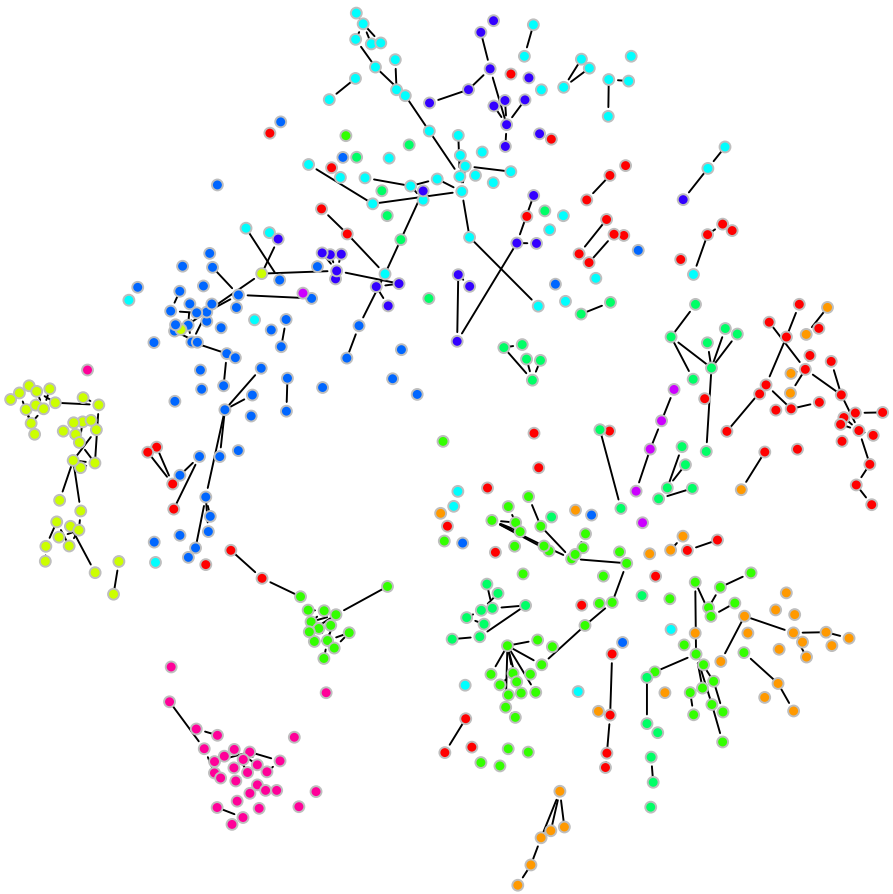


S&P Data: Forest Graph

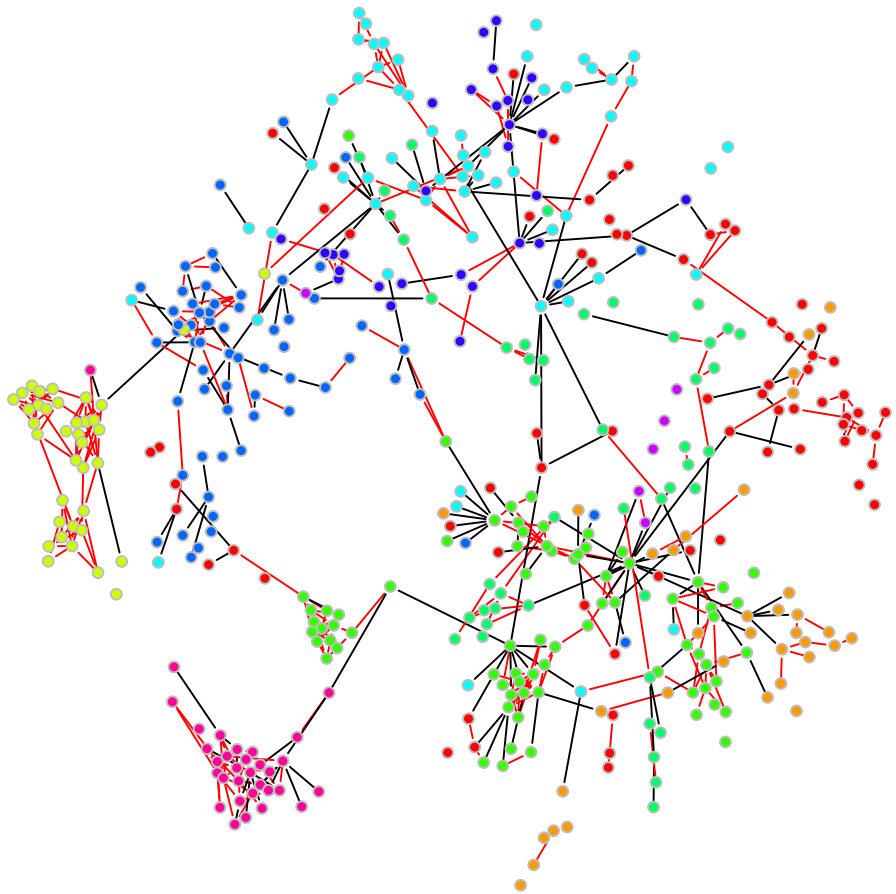


S&P Data: Forest vs. Nonparanormal

common edges



differences



Summary

- Smoothing kernels, Mercer kernels
- Sparse additive models
- Constrained rank additive models
- Nonparametric graphical models: Nonparanormal and forest-structured densities
- A little nonparametricity goes a long way.

Summary

- 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.