Computational and Statistical Aspects of Statistical Machine Learning

John Lafferty

Department of Statistics Retreat
Gleacher Center

Outline

- “Modern” nonparametric inference for high dimensional data
  - Nonparametric reduced rank regression
- Risk-computation tradeoffs
  - Covariance-constrained linear regression
- Other research and teaching activities

Context for High Dimensional Nonparametrics

Great progress in recent years on high dimensional linear models

Many problems have important nonlinear structure.

We've been studying "*purely functional*" methods for high dimensional, nonparametric inference

- no basis expansions
- no Mercer kernels

Additive Models

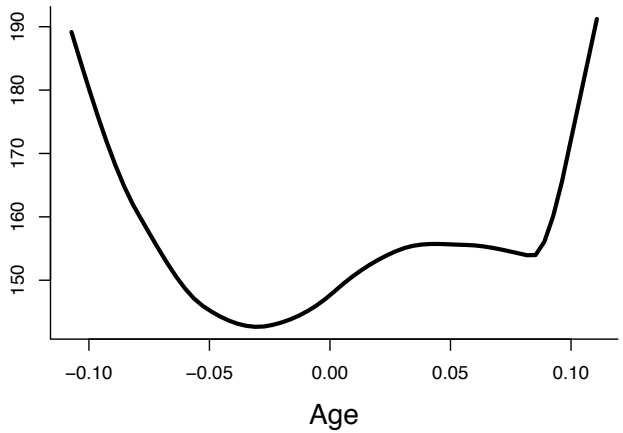
Fully nonparametric models appear hopeless

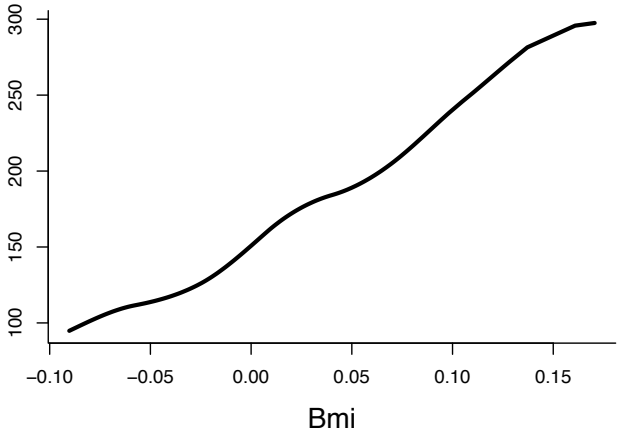
- Logarithmic scaling, $p = \log n$ (e.g., "Rodeo" Lafferty and Wasserman (2008))

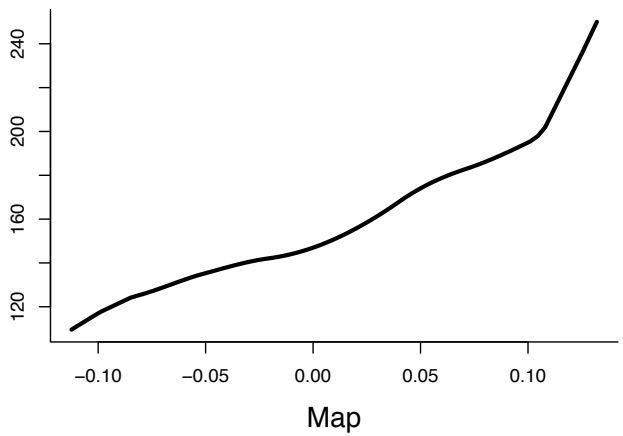
Additive models are useful compromise

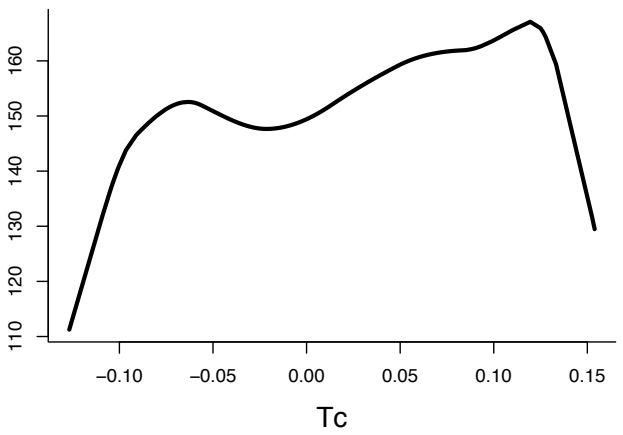
- Exponential scaling, $p = \exp(n^c)$ (e.g., "SpAM" Ravikumar, Lafferty, Liu and Wasserman (2009))

Additive Models









Multivariate Regression

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

Linear model $Y = BX + \epsilon$ where $B \in \mathbb{R}^{q \times p}$.

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

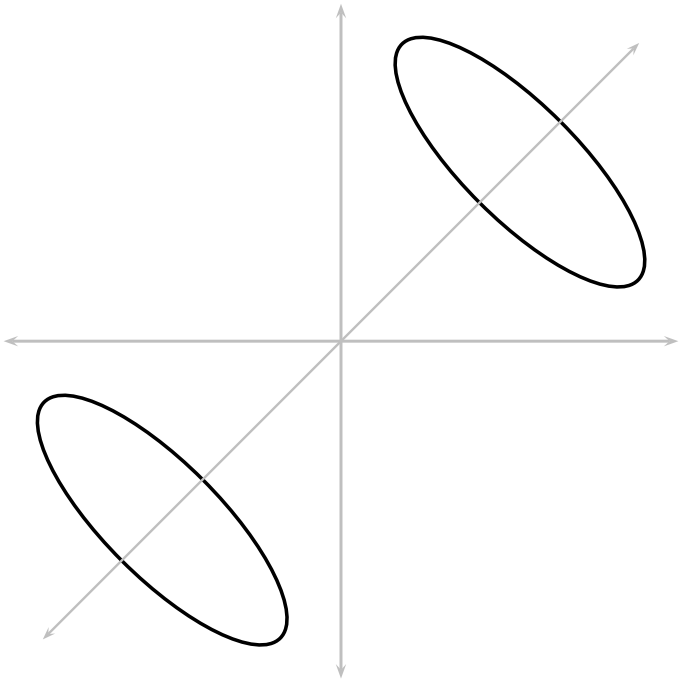
Recent work has studied properties and high dimensional scaling of reduced rank regression where nuclear norm $\|B\|_*$ is used as convex surrogate for rank constraint (Yuan et al., 2007; Negahban and Wainwright, 2011). E.g.,

$$\|\widehat{B}_n - B^*\|_F = O_P\left(\sqrt{\frac{\text{Var}(\epsilon) r(p+q)}{n}}\right)$$

Low-Rank Matrices and Convex Relaxation

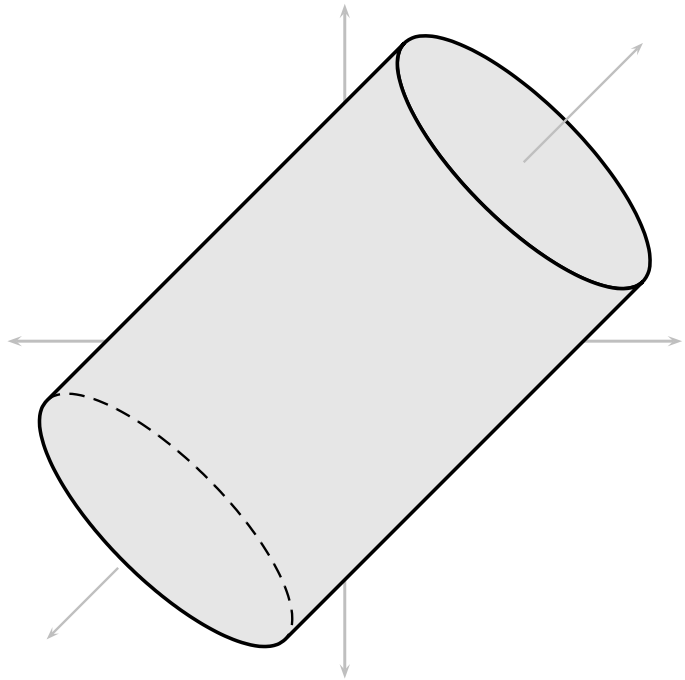
low rank matrices

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



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:

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

Nonparametric Reduced Rank Regression

Foygel, Horrell, Drton and Lafferty (NIPS 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

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

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

Linear case:

$$\sum_{j=1}^{p} \left\|\Sigma_p^{1/2}\right\|_* = \sum_{j=1}^{p} \|B_j\|_2$$

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

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

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

Scaling of Estimation Error

Using a "double covering" technique, ($\frac{1}{2}$-parametric, $\frac{1}{2}$-nonparametric), we bound the deviation between empirical and population functional covariance matrices in spectral norm:

$$\sup_V \left\| \Sigma(V) - \widehat{\Sigma}_n(V) \right\|_{\text{sp}} = O_P\left(\sqrt{\frac{q + \log(pq)}{n}} \right).$$

This allows us to bound the excess risk of the empirical estimator relative to an oracle.

Summary

- *Variations on additive models enjoy most of the good statistical and computational properties of sparse or low-rank linear models.*
- We're building a toolbox for large scale, high dimensional nonparametric inference.

Computation-Risk Tradeoffs

- In “traditional” computational learning theory, dividing line between learnable and non-learnable is polynomial vs. exponential time
- Valiant’s PAC model
- Mostly negative results: It is not possible to efficiently learn in natural settings
- *Claim: Distinctions in polynomial time matter most*

Analogy: Numerical Optimization

In numerical optimization, it is understood how to tradeoff computation for speed of convergence

- First order methods: linear cost, linear convergence
- Quasi-Newton methods: quadratic cost, superlinear convergence
- Newton's method: cubic cost, quadratic convergence

Are similar tradeoffs possible in statistical learning?

Hints of a Computation-Risk Tradeoff

Graph estimation:

- Our method for estimating graph for Ising models: $n = \Omega(d^3 \log p)$, $T = O(p^4)$ for graphs with p nodes and maximum degree d
- Information-theoretic lower bound: $n = \Omega(d \log p)$

Statistical vs. Computational Efficiency

Challenge: Understand how families of estimators with different computational efficiencies can yield different statistical efficiencies

$$\text{Rate}_{\mathcal{H},\mathcal{F}}(n) = \inf_{\widehat{m}_n \in \mathcal{H}} \sup_{m \in \mathcal{F}} \text{Risk}(\widehat{m}_n, m)$$

- $\mathcal{H}$: computationally constrained hypothesis class
- $\mathcal{F}$: smoothness constraints on "true" model

Computation-Risk Tradeoffs for Linear Regression

Dinah Shender has been studying such a tradeoff in the setting of high dimensional linear regression

Computation-Risk Tradeoffs for Linear Regression

Standard ridge estimator solves

$$\left(\frac{1}{n}X^T X + \lambda_n I\right)\widehat{\beta}_\lambda = \frac{1}{n}X^T Y$$

Sparsify sample covariance to get estimator

$$\left(T_t[\widehat{\Sigma}] + \lambda_n I\right)\widetilde{\beta}_{t,\lambda} = \frac{1}{n}X^T Y$$

where $T_t[\widehat{\Sigma}]$ is hard-thresholded sample covariance:

$$T_t([m_{ij}]) = \left[m_{ij}\,\mathbf{1}(|m_{ij}| > t)\right]$$

Recent advance in theoretical CS (Spielman et al.): Solving a symmetric diagonally-dominant linear system with m nonzero matrix entries can be done in time

$$\widetilde{O}(m \log^2 p)$$

Computation-Risk Tradeoffs for Linear Regression

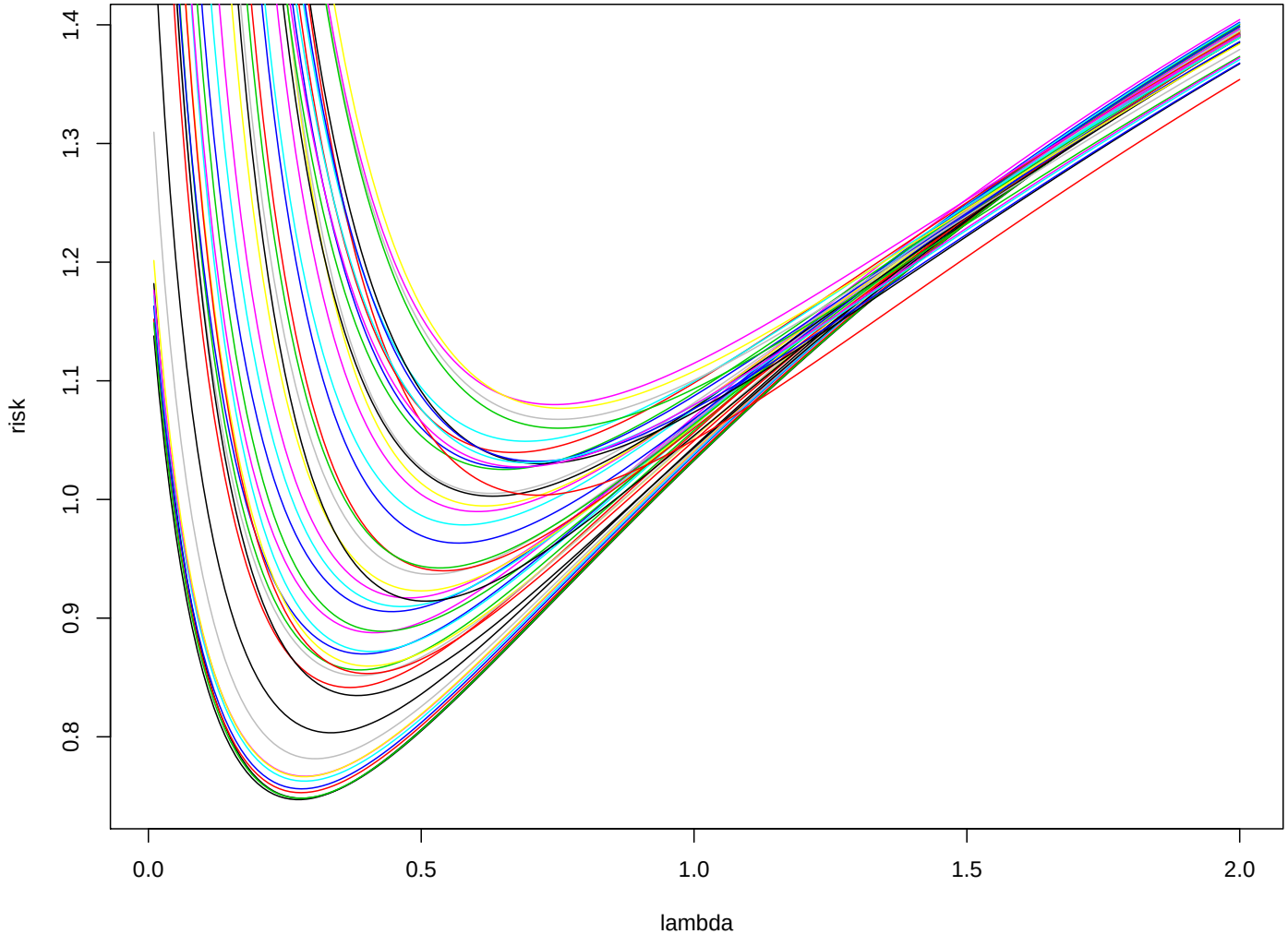
Dinah has recently proved the statistical error scales as

$$\frac{\|\widetilde{\beta}_{t,\lambda} - \beta^*\|}{\|\beta^*\|} = O_P\left(\|T_t(\Sigma) - \Sigma\|_2\right) = O(t^{1-q})$$

for class of covariance matrices with rows in sparse ℓ_q balls (as studied by Bickel and Levina).

- Combined with the computational advance, this gives us an explicit, fine-grained risk/computation tradeoff

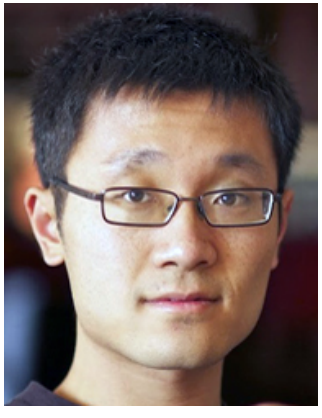
Simulation



Some Other Projects

Minhua Chen: Convex optimization for dictionary learning

Eric Janofsky: Nonparanormal component analysis



Min Xu: High dimensional conditional density and graph estimation

Courses in the Works

- Winter 2013: **Nonparametric Inference** (Undergraduate and Masters)

- Spring 2013: **Machine Learning for Big Data** (Undergraduate Statistics and Computer Science)

Charles Cary: Developing Cloud-based infrastructure for the course. Candidate data: 80 million images, Yahoo! clickthrough data, *Science* journal articles, City of Chicago datasets.