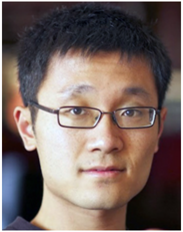
Sensing and Analysis of High-Dimensional Data
UCL-Duke Workshop
September 4 & 5, 2014

Variable Selection in High Dimensional Convex Regression

John Lafferty

Department of Statistics &
Department of Computer Science
University of Chicago

Collaborators



Min Xu
CMU / UChicago

Minhua Chen
Duke / UChicago / Amazon

Collaborators

Sabyasachi Chatterjee
Yale / UChicago

YJ Choe
UChicago

Outline

- Nonparametric additive models
- Faithful variable selection in convex regression
- Algorithm using convex and concave additive models
- Finite sample analysis
- Convexity pattern decoding

Context

- Great progress in recent years on high dimensional models
- We have been trying to push nonparametric methods further
- Shape-constrained approaches are attractive for many reasons

High Dimensional Variable Selection

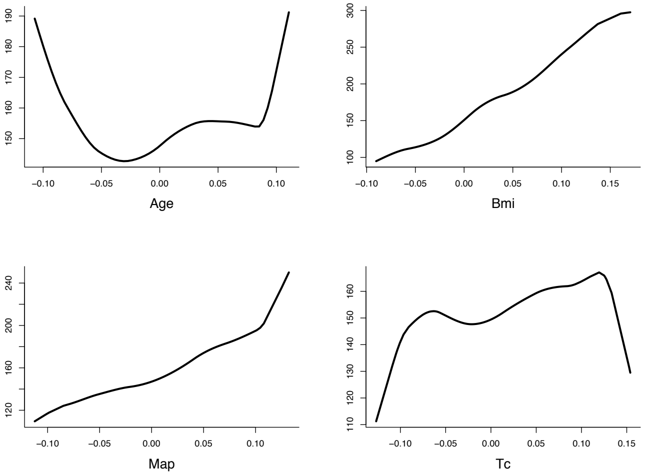
Fully nonparametric models appear hopeless

- Logarithmic scaling, $p = \log n$ (e.g., “Rodeo” L. and Wasserman, 2008)

Additive models are useful compromise

- Exponential scaling, $p = \exp(n^c)$ (e.g., “SpAM” Ravikumar et al., 2009)
- But do not give faithful variable selection

Additive Models



Difficulty: Choosing smoothing parameters

Convex Regression: High Level

- Convex regression is fully nonparametric, with no tuning parameters.
- Shape constraints often natural in economics, marketing, reinforcement learning, etc.
- Estimation is a convex optimization problem. Efficient, scalable QP algorithms.
- We can recover sparsity pattern for convex regression assuming an (incorrect) additive model
- “21st century version of the 4 B’s” (Efron)

Barlow, Bartholomew, Bremner and Brunk (1972), “Statistical inference under order restrictions”

Convex Regression

The infinite-dimensional nonparametric convex regression

$$\min_{f \text{ convex}} \sum_i (y_i - f(x_i))^2$$

is equivalent to the finite dimensional QP

$$\begin{aligned} \min_{f,\beta} \quad & \sum_i (y_i - f_i)^2 \\ \text{such that} \quad & f_j \geq f_i + \beta_i^T (x_j - x_i) \end{aligned}$$

Guntuboyina (2012): minimax analysis for support functions. Rate $n^{-4/(3+p)}$ equivalent to requiring two derivatives

Minimax analysis for convex regression not yet complete; new results by Guntuboyina and Sen (2013) in 1-d setting

Other Previous Work

N. Pya and S. Wood, "Shape constrained additive models," *Statistics and Computing*, February 2014.

- Uses P-splines and requires smoothing parameters

L. Hannah and D. Dunson, (1) "Multivariate convex regression with adaptive partitioning," *JMLR*, 2013; (2) "Bayesian nonparametric convex regression," 2011.

(1) partitions data and constructs linear estimates

(2) places prior over piecewise planar functions

Variable Selection in Convex Regression: Results

- Variable selection using a (potentially mis-specified) convex additive model is “faithful” — no false negatives
- “Sparsistent” variable selection achievable with sample complexity

$$n^{4/5} \geq Cs^5\sigma^2 \log^2 p$$

where s is the number of relevant variables.

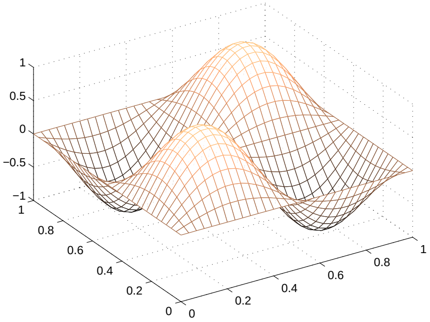
Faithfulness

Additive approximation is

$$\{f_k^*\}, \mu^* := \operatorname*{arg\,min}_{f_1,\ldots,f_p,\mu} \left\{ \mathbb{E}\left[\left(f(X) - \sum_{k=1}^{p} f_k(X_k) - \mu\right)^2\right] \,:\, \mathbb{E}f_k(X_k) = 0 \right\}.$$

We say f is *additively faithful* in case $f_k^* = 0$ implies that f does not depend on coordinate k.

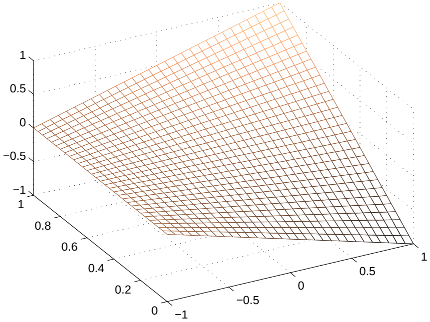
Nonconvex and Unfaithful



Egg carton: $f(x_1, x_2) = \sin(2\pi x_1)\sin(2\pi x_2)$.

An additive approximation would set $f_1 = 0$ and $f_2 = 0$

Nonconvex and Unfaithful



Tilting slope: $f(x_1, x_2) = x_1 x_2$ for $x_1 \in [-1, 1]$ and $x_2 \in [0, 1]$.

An additive approximation would set $f_2 = 0$

Faithfulness under Convexity

Theorem. Suppose the data density is supported on $[0,1]^p$ and satisfies the *boundary points condition*

$$\frac{\partial p(\boldsymbol{x}_{-j} \mid x_j)}{\partial x_j} = \frac{\partial^2 p(\boldsymbol{x}_{-j} \mid x_j)}{\partial x_j^2} = 0 \quad \text{at } x_j = 0, x_j = 1.$$

If f is convex and twice differentiable, then f is additively faithful with respect to p.

Intuition

Suppose the underlying distribution has a product density.

Then the additive approximation zeroes out k when, fixing x_k, every "slice" of f integrates to zero.

The proof of this result shows that "slices" of convex functions that integrate to zero cannot be "glued together" while still maintaining convexity.

Using Shape Constraints

Difficult to estimate optimal additive functions f_k^*—need not be convex

When can a convex additive model be used?

We need to couple with fitting *concave* functions on the residuals:

$$g_k^* = \arg\min \left\{ \mathbb{E}\Big(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k \Big)^2 \; : \; g_k \in \text{-}\mathcal{C}^1, \mathbb{E} g_k(X_k) = 0 \right\}.$$

Faithful Variable Screening

Theorem. Suppose the density satisfies the boundary points condition, and f is convex and twice differentiable. Then $f_k^* = 0$ and $g_k^* = 0$ implies that f does not depend on x_k.

AC/DC Algorithm

1. *AC Stage*: Estimate an additive convex model

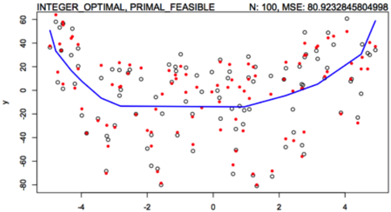
$$\{\widehat{f}_k\}, \widehat{\mu} = \underset{f_1,\dots,f_p \in \mathcal{C}^1,\, \mu \in \mathbb{R}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \mu - \sum_{k=1}^{p} f_k(x_{ik}) \Big)^2 + \lambda \sum_{k=1}^{p} \|f_k\|_\infty$$

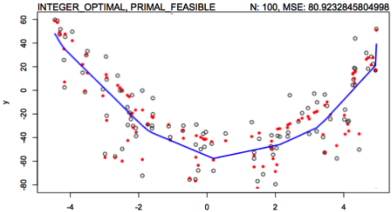
2. *DC Stage*: If $\|\widehat{f}_k\|_\infty = 0$, estimate decoupled concave function:

$$\widehat{g}_k = \underset{g_k \in -\mathcal{C}^1}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \widehat{\mu} - \sum_{k'} \widehat{f}_{k'}(x_{ik'}) - g_k(x_{ik}) \Big)^2 + \lambda \|g_k\|_\infty$$

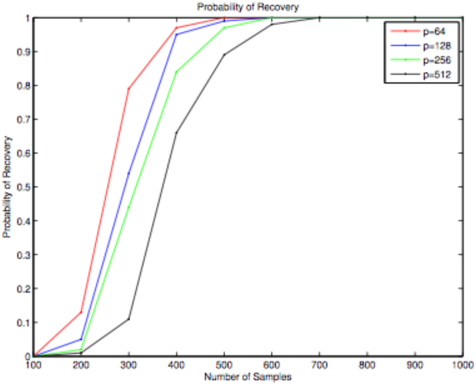
3. Estimated support $\widehat{S}_n = \{k \,:\, \|\widehat{f}_k\|_\infty > 0 \text{ or } \|\widehat{g}_k\|_\infty > 0\}$

Example simulation





AC⚡DC Sparsity Recovery Curves



Finite Sample Analysis: Assumptions

A1: f_0 convex, twice differentiable

A2: $\|f_0\|_\infty \leq sB$

A3: sub-Gaussian noise

A4: X_S and X_{S^c} are independent

A5: boundary points condition

Finite Sample Analysis: Signal-to-Noise

Define

$$\alpha_+ = \inf_{f \in \mathcal{C}_1^p \,:\, \exists k,\, f_k^* \neq 0 \wedge f_k = 0} \left\{ \mathbb{E}\big(f_0(X) - f(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X)\big)^2 \right\}$$

Smallest excess approximation error if a relevant variable is omitted in AC stage.

If α_+ is small, false negatives may occur in the AC stage.

Plays role of smallest coefficient in lasso theory.

Finite Sample Analysis: Signal-to-Noise

Define

$$\alpha_- = \min_{k \in S\,:\, g_k^* \neq 0} \left\{ \mathbb{E}\big(f_0(X) - f^*(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X) - g_k^*(X_k)\big)^2 \right\}.$$

Smallest excess approximation error if a relevant variable is omitted in DC stage.

If α_- is small, false negatives may occur in the DC stage.

Finite Sample Analysis: Sparsistency

Theorem. Suppose that the regularization level is

$$\lambda_n = c_1 sB\sqrt{\frac{\sigma^2 \log^2 np}{n}}$$

and the signal-to-noise ratio satisfies

$$\frac{\alpha_+}{\sigma}, \left(\frac{\alpha_-}{\sigma}\right)^2 \geq c_2 B^3 \sqrt{\frac{s^5 \log^2 np}{n^{4/5}}}.$$

Then for $n^{4/5} \geq c_3 \sigma^2 s^5 \log^2 p$, the AC/DC algorithm outputs a support set $\widehat{S}_n$ satisfying

$$\mathbb{P}(\widehat{S}_n = S) \geq 1 - \frac{1}{n}.$$

Finite Sample Analysis: Sparsistency

- Allows exponential scaling $p = O(\exp(n^c))$ in ambient dimension
- Allows intrinsic dimension to scale as $|S| \equiv s = o(n^{4/25})$
- Gives $n = O(\text{poly}(s))$ sample complexity.
- Comminges and Dalalyan (2012) show that under traditional smoothness constraints, consistent variable selection in high dimensions is only possible if $n \geq \exp(s)$.

Finite Sample Analysis: Proof

The proof exploits recent bracketing number bounds for convex function classes by Kim and Samworth (2014). Specifically, we bound

$$\left\langle W, \widehat{f} - f^* - \overline{f^*} \right\rangle$$

using bracketing entropy, where W is the noise.

This removes some of the limitations of covering number bounds developed by Guntuboyina and Sen (2013).

Current Work: Convexity Pattern Decoding

- Suppose we have an additive model with a sum of convex and concave functions
- Estimation is a QP with no smoothing parameters
- *What if we don't know the convexity pattern—which functions are convex and which are concave? Can it be learned?*

Convexity Pattern Decoding

Model:

$$Y = \sum_{j=1}^{p} z_j \, f_j(x_j) + \varepsilon$$

$$z_j \in \{-1, 1\}, \quad f_j \text{ convex}$$

Problem:

$$\text{Given data } \{(X_i, Y_i)\}_{i=1}^{n},\, X_i \in \mathbb{R}^p,\; Y_i \in \mathbb{R},$$

$$\text{decode } z = (z_1, \ldots, z_p) \in \{-1, 1\}^p$$

Solving this problem will lead to a new, useful approach to high-dimensional nonparametric estimation with no tuning parameters.

Mixed Integer SOCP Formulation

$$\begin{aligned}
\min_{f,g,z,w} \quad & \sum_{i=1}^{n} \Big(Y_i - \sum_{j=1}^{p} (f_{ij} + g_{ij}) \Big)^2 \\
\text{such that} \quad & \text{convexity constraints on } f_j \\
& \text{concavity constraints on } g_j \\
& \sqrt{\sum_{i=1}^{n} f_{ij}^2} \leq z_j B \\
& \sqrt{\sum_{i=1}^{n} g_{ij}^2} \leq w_j B \\
& z_j + w_j \leq 1 \\
& z_j, w_j \in \{0, 1\}.
\end{aligned}$$

A Better, Convex Approach

$$
\begin{aligned}
\min_{f,g,\beta,\gamma,z,w} \quad & \sum_{i=1}^{n}\Big(Y_i - \sum_{j=1}^{p}(f_{ij}+g_{ij})\Big)^2 \\
\text{such that} \quad & \text{convexity constraints on } f_j \\
& \text{concavity constraints on } g_j \\
& \sum_{j=1}^{p}\left\{\beta_{(n)j} - \beta_{(1)j} + \gamma_{(1)j} - \gamma_{(n)j}\right\} \le L
\end{aligned}
$$

$\beta_{(1)j}, \beta_{(n)j}, \gamma_{(1)j}, \gamma_{(n)j}$ are first and last subgradient vectors of f_j and g_j

A nonstandard type of lasso. Works well – requires special analysis.

Summary

- Gave conditions for *additive faithfulness* in convex function estimation
- Proposed *AC/DC algorithm* for variable selection using convex additive models
- Analyzed finite sample behavior, giving *sparsistency rate of convergence*
- Introduced problem of *convexity pattern decoding*