Nonparametric Convex Regression[1]

Bodhisattva Sen
Department of Statistics
Columbia University, New York
bodhi@stat.columbia.edu

University of North Carolina
Chapel Hill
4 April, 2014

[1]Supported by NSF grant DMS-1150435 (CAREER)

Outline

What is convex regression?

- Regression model of the form

$$Y_i = \phi_0(X_i) + \epsilon_i, \qquad \text{for } i = 1, \ldots, n,$$

  where

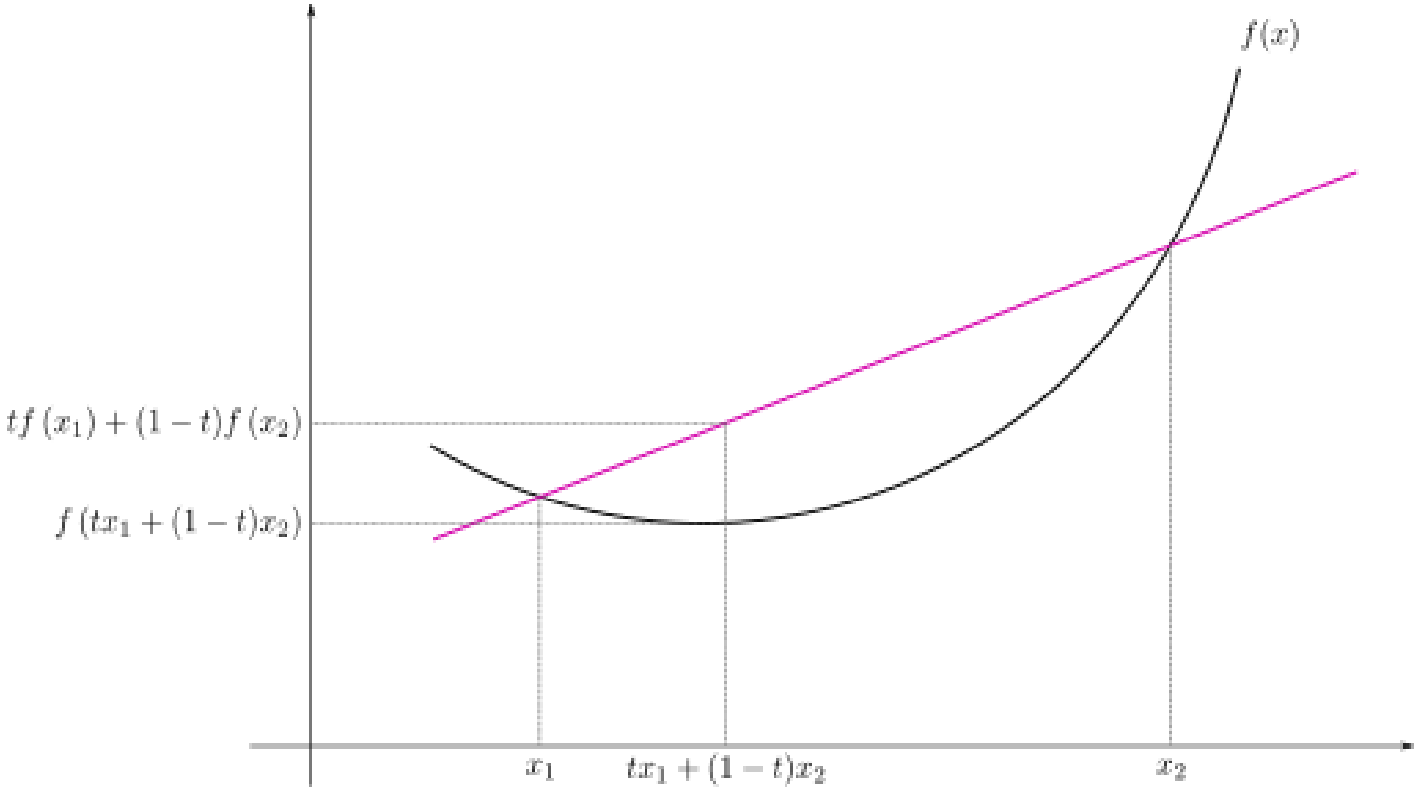
  - $\phi_0 : \mathbb{R}^d \to \mathbb{R}$, $d \geq 1$, is an unknown *convex* function
  - X_i is an $\mathfrak{X}$-valued vector ($\mathfrak{X} \subset \mathbb{R}^d$ is a closed convex set)
  - ϵ_i is a random variable with $\mathbb{E}(\epsilon_i | X_i) = 0$ and $\mathrm{Var}\,(\epsilon_i) < \infty$

- Goal: Estimate ϕ_0 *nonparametrically*, assuming the *convexity*

What is a convex function?

- $f : \mathbb{R}^d \to \mathbb{R}$ is *convex* iff for all $X_1, X_2 \in \mathbb{R}^d$ and $t \in [0, 1]$,

$$f(tX_1 + (1-t)X_2) \leq t\, f(X_1) + (1-t)\, f(X_2).$$



- Arises frequently in *economics*, operations research, financial engineering, e.g., production functions are known to be *concave* & $\uparrow$
- Approximating a function (objective function, constraint) for a convex optimization problem
- Convexity is probably the simplest *shape* constraint that has a natural generalization to multi-dimension.

Two further applications of the method

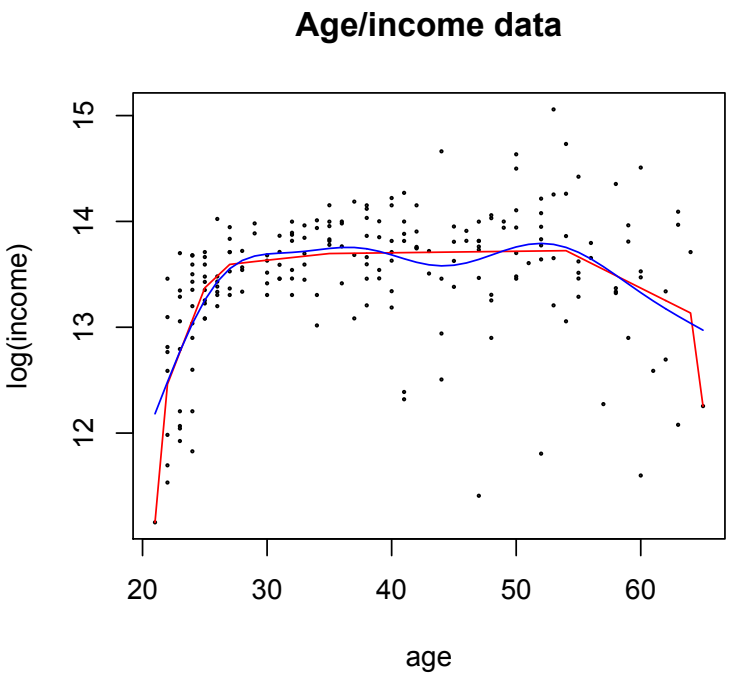
- ϕ_0 is *affine* (i.e. $\phi_0(X) = \alpha + \beta^T X$) iff ϕ_0 both convex and concave
  - Test the goodness-of-fit of a linear model (Sen and Meyer (2013))
- Any *smooth* function on a compact domain can be represented as the difference of two convex functions
  - Estimate any smooth function (Work in progress)

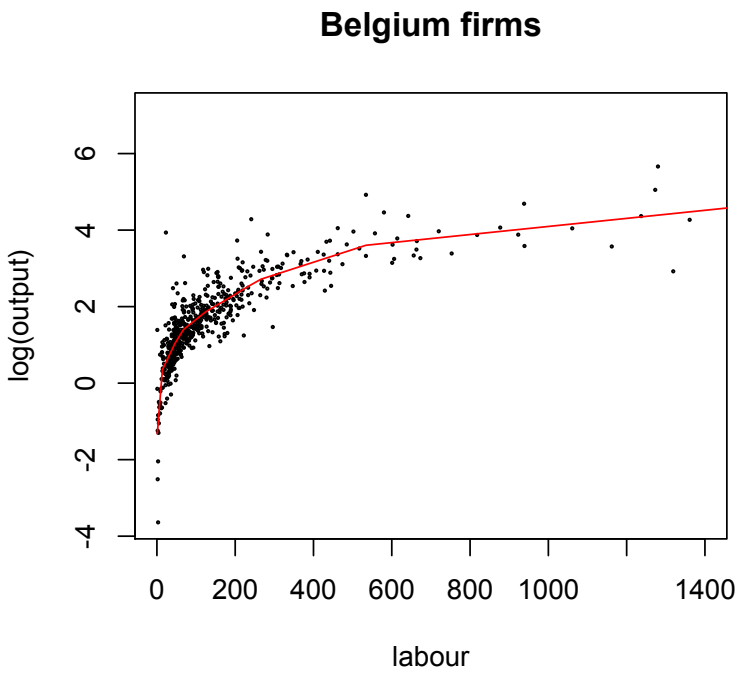
Convex regression when $d = 1$

- Model
$$Y_i = \phi_0(x_i) + \epsilon_i, \qquad \text{for } i = 1, 2, \ldots, n$$

- $\phi_0 : \mathbb{R} \to \mathbb{R}$ is an unknown *convex* (or concave) function
- $x_1 < x_2 < \cdots < x_n$ are known constants; $\epsilon_1, \ldots, \epsilon_n$ mean 0 errors.

- Want to estimate ϕ_0 *nonparametrically*, given ϕ_0 is *convex/concave*





The estimation procedure: least squares

- Model

$$Y_i = \phi_0(X_i) + \epsilon_i \qquad \text{for } i = 1, \ldots, n; \quad X_i \in \mathbb{R}^d; d \geq 1$$

- We estimate ϕ_0 via the method of *least squares*

$$\hat{\phi} = \underset{\psi:\mathbb{R}^d \to \mathbb{R} \text{ convex}}{\operatorname{argmin}} \sum_{i=1}^{n} \{Y_i - \psi(X_i)\}^2 .$$

- $\hat{\boldsymbol{\theta}} = (\hat{\phi}(x_1), \ldots, \hat{\phi}(x_n))$ is *unique* (Why? next slide ...)

- Long history: Hildreth (1954) ...

- Fully *automated*; does not require the choice of tuning parameter(s)

- *Sequence model*: $Y_i = \phi_0(X_i) + \epsilon_i,\ i = 1, 2, \ldots, n$, is equivalent to

$$\mathbf{Y} = \boldsymbol{\theta}_0 + \boldsymbol{\epsilon}$$

- $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top, \boldsymbol{\theta}_0 = (\phi_0(X_1), \ldots \phi_0(X_n))^\top, \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$

- $\min_{\psi \text{ convex}} \sum_{i=1}^n \{Y_i - \psi(X_i)\}^2 \equiv \min_{\boldsymbol{\theta} \in \mathcal{K}} \sum_{i=1}^n (Y_i - \theta_i)^2$

- $\mathcal{K} = \{\boldsymbol{\theta} \in \mathbb{R}^n : \exists\ \psi \text{ convex s.t. } \psi(X_i) = \theta_i,\ \forall i = 1, \ldots, n\}$

- $\hat{\boldsymbol{\theta}}$ is the *projection* of $\mathbf{Y}$ onto $\mathcal{K}$, i.e.,

$$\hat{\boldsymbol{\theta}} := \Pi(\mathbf{Y}|\mathcal{K}) = \underset{\boldsymbol{\theta} \in \mathcal{K}}{\operatorname{argmin}} \|\mathbf{Y} - \boldsymbol{\theta}\|^2$$

- $\mathcal{K}$ is a *closed convex cone* in $\mathbb{R}^n$ ($\mathcal{T}$ is a convex cone iff $\alpha_1\boldsymbol{\theta}_1 + \alpha_2\boldsymbol{\theta}_2 \in \mathcal{T}, \forall \alpha_1, \alpha_2 > 0 \ \ \& \ \ \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{T}$)

- The *projection theorem* gives the existence and uniqueness of $\hat{\boldsymbol{\theta}}$.

Computation when $d = 1$

- $\mathcal{K}$ has a *straightforward characterization* when $d = 1$, i.e.,

$$\mathcal{K} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^n : \frac{\theta_2 - \theta_1}{x_2 - x_1} \leq \frac{\theta_3 - \theta_2}{x_3 - x_2} \leq \cdots \leq \frac{\theta_n - \theta_{n-1}}{x_n - x_{n-1}} \right\}$$

- Quadratic program with $(n - 2)$ linear constraints

- $\hat{\phi}$ ends up being a *piece-wise linear function* constructed by linearly interpolating the pairs $(x_1, \hat{\theta}_1), \ldots, (x_n, \hat{\theta}_n)$

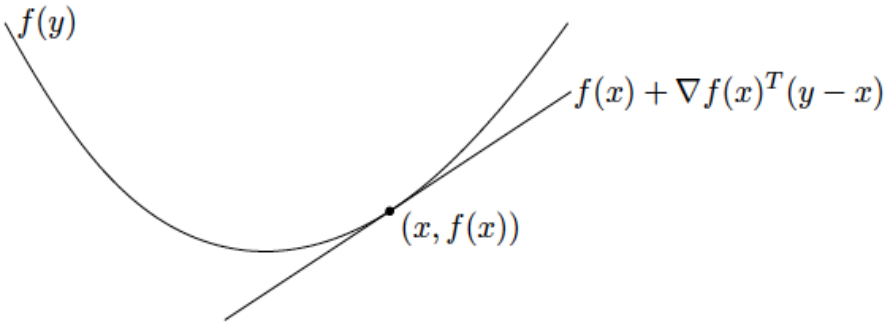
What happens when $d > 1$?

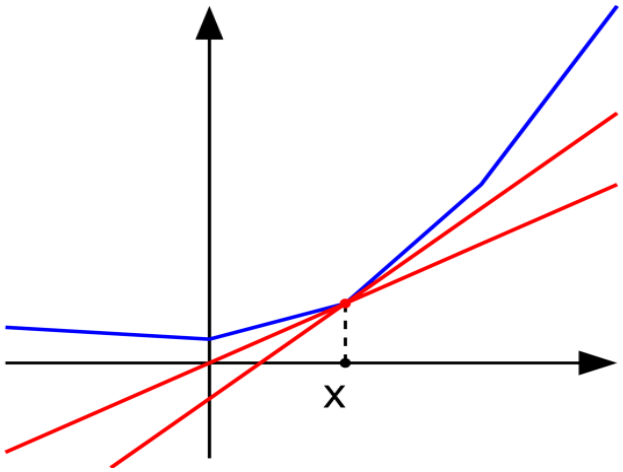
- If f is *differentiable*, then f is convex iff

$$f(x) + \nabla f(x)^\top (y - x) \leq f(y), \qquad \text{for all } x, y \in \mathbb{R}^d$$

where $\nabla f(x)$ is the *gradient* of f at x.

- The first-order Taylor approximation is a *global under-estimator* of f.





- In fact, for *every* convex function f, there exists $\xi_x \in \mathbb{R}^d$ such that

$$f(x) + \xi_x^\top (y - x) \leq f(y) \qquad \text{for all } x, y \in \mathbb{R}^d.$$

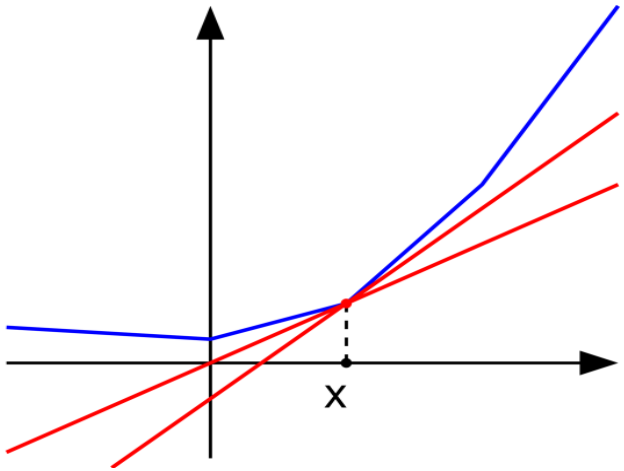
- ξ_x is called a *sub-gradient* at x.

Sub-differentials and sub-gradients

- The *sub-differential* of f at $x \in \mathbb{R}^d$ is given by

$$\partial f(x) := \{\xi \in \mathbb{R}^d : f(x) + \xi^T(y - x) \leq f(y) \quad \forall y \in \mathbb{R}^d\}.$$

- The elements of the sub-differential are called *sub-gradients*.
- Example: If $f(x) = \|x\|$, $\partial f(0) = \{\xi \in \mathbb{R}^d : \|\xi\| \leq 1\}$.
- f is *differentiable* at x iff $\partial f(x) = \{\nabla f(x)\}$.



Computation of $\hat{\theta}$

- The vector $\hat{\boldsymbol{\theta}}$ can now be computed by solving the following *quadratic program* with linear constraints:

$$\begin{array}{rc} \min & \|\mathbf{Y} - \boldsymbol{\theta}\|^2 \\ \text{subject to} & \theta_i + \xi_i^\top (X_j - X_i) \leq \theta_j \quad \forall\ i, j = 1, \ldots, n \\ & \xi_1, \ldots, \xi_n \in \mathbb{R}^d, \boldsymbol{\theta} \in \mathbb{R}^n \end{array}$$

- This problem has a *unique solution in $\boldsymbol{\theta}$* but not in the ξ_j's (i.e., for any two solutions $(\xi_1, \ldots, \xi_n, \boldsymbol{\theta})$ and $(\tau_1, \ldots, \tau_n, \boldsymbol{\zeta})$ we have $\boldsymbol{\theta} = \boldsymbol{\zeta}$)
- Quadratic program with $n(n-1)$ linear constraints and $n(d+1)$ variables.
- Can use MOSEK, *cvx* or any off-the-shelf solver for $n \sim 200$
- *Scalable* algorithm? (work in progress; can work with $n \sim 5000$)

- *Estimator* at a point X: $\hat{\phi}(X) = \max_{i=1,\ldots,n}\{\hat{\theta}_i + (X - X_i)^\top \hat{\xi}_i\}$

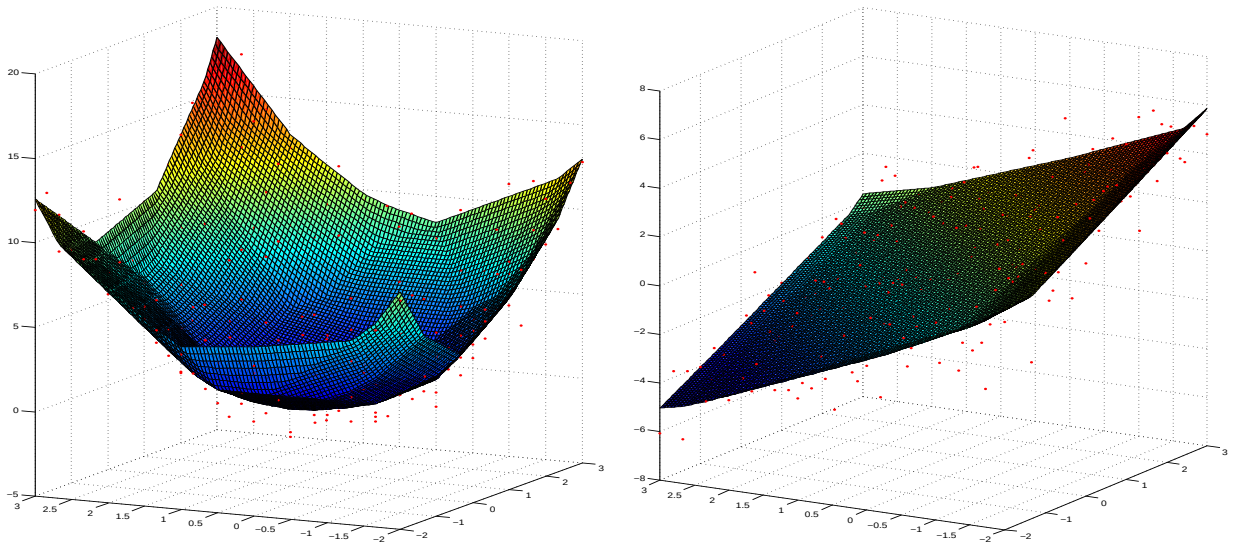


Figure: The scatter plot and LSE: (a) $\phi_0(X) = \|X\|^2$ (left panel); (b) $\phi_0(X) = -X_1 + X_2$ (right panel). For both cases, $n = 256$ and $\epsilon \sim \mathcal{N}(0, 1)$.

- $\hat{\phi}$ is a *polyhedral convex function*.
- Belgium labour firms example (MATLAB figures)

Consistency

The LSE is *consistent* in *fixed* and *stochastic* design settings.

Consistency theorem (Seijo and Sen (2011))

Under appropriate conditions we have:

(i) $\mathbb{P}(\sup_{X \in \mathtt{X}} \{|\hat{\phi}(X) - \phi_0(X)|\} \to 0 \ \text{ for any compact set } \mathtt{X} \subset \mathfrak{X}^\circ) = 1.$

(ii) Denoting by $\mathbf{B}$ the unit ball (w.r.t. the Euclidian norm) we have

$$\mathbb{P}(\partial\hat{\phi}(X) \subset \partial\phi_0(X) + \epsilon\mathbf{B} \ \text{ almost always}) = 1 \quad \forall\, \epsilon > 0,\ \forall\, X \in \mathfrak{X}^\circ.$$

(iii) If ϕ_0 is differentiable on $\mathfrak{X}^\circ$, then

$$\mathbb{P}\left(\sup_{\substack{\xi \in \partial\hat{\phi}(X) \\ X \in \mathtt{X}}} \|\xi - \nabla\phi_0(X)\| \to 0 \ \text{ for any compact set } \mathtt{X} \subset \mathfrak{X}^\circ \right) = 1.$$

Testing against a linear model

- Let $\mathbf{Y} := (Y_1, Y_2, \ldots, Y_n) \in \mathbb{R}^n$ and

$$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\theta}_0, \sigma^2 \mathbf{I}_n), \qquad \sigma^2 > 0$$

- Problem: Test $H_0 : \boldsymbol{\theta}_0 \in \mathcal{S}$ where

$$\mathcal{S} := \{\boldsymbol{\theta} \in \mathbb{R}^n : \boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\beta} \in \mathbb{R}^k, \text{ for some } k \geq 1\}$$

  and $\mathbf{X}$ is the *design* matrix.

- Important problem with a long history

- Most methods fit the *parametric* and a *nonparametric* model to estimate the regression function

- *Compare* the *two fits* to come up with a test statistic

Drawbacks of most methods

- Involves the *choice* of delicate tuning parameters (e.g., bandwidths)
- Calibration (i.e., the level) of the test is usually done by resampling the test statistics (bootstrap), and thus only *approximate*.

We develop a *likelihood ratio test* for H_0 using ideas from shape restricted regression.

- Fully *automated*, no tuning parameters required
- Test has *exact* level
- *Power* of the test goes to 1, under (almost) *all* alternatives

Testing against linear regression

$$Y_i = \phi_0(X_i) + \epsilon_i, \qquad i = 1, 2, \ldots, n.$$

- $\phi_0 : \mathbb{R}^d \to \mathbb{R}$, $d \geq 1$, is an unknown *smooth* function
- Design points $X_1, X_2, \ldots, X_n$ in $\mathbb{R}^d$; $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$
- Test $H_0 : \phi_0$ is *affine*, i.e., $\phi_0(X) = a + b^\top X$, for some unknown $a \in \mathbb{R}, b \in \mathbb{R}^d$
- Thus, $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\theta}_0, \sigma^2 \mathbf{I}_n)$ where $\boldsymbol{\theta}_0 := (\phi_0(X_1), \ldots, \phi_0(X_n))^\top \in \mathbb{R}^n$
- Test $H_0 : \boldsymbol{\theta}_0 \in \mathcal{S}$ where $\mathcal{S} := \{\boldsymbol{\theta} \in \mathbb{R}^n : \boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\beta} \in \mathbb{R}^{d+1}\}$, and $\mathbf{X} = [\mathbf{e}|\mathbf{x}]$ where $\mathbf{x} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times d}$, $\mathbf{e} = (1, \ldots, 1)^\top \in \mathbb{R}^n$.

Idea

- A function is *affine* iff it is *both* convex and concave.
- Let $\mathcal{I} = \{\boldsymbol{\theta} \in \mathbb{R}^n : \exists\ \psi \text{ convex s.t. } \psi(X_i) = \theta_i,\ \forall i = 1, \ldots, n\}$
- Let $\mathcal{D} = -\mathcal{I} = \{\boldsymbol{\theta} \in \mathbb{R}^n : \exists\ \psi \text{ concave s.t. } \psi(X_i) = \theta_i\}$
- $\mathcal{I}, \mathcal{D}$ are *cones* that contain $\mathcal{S}$
- If H_0 is true, both $\Pi(\mathbf{Y}|\mathcal{I})$ and $\Pi(\mathbf{Y}|\mathcal{D})$ must be *close* to $\Pi(\mathbf{Y}|\mathcal{S})$.

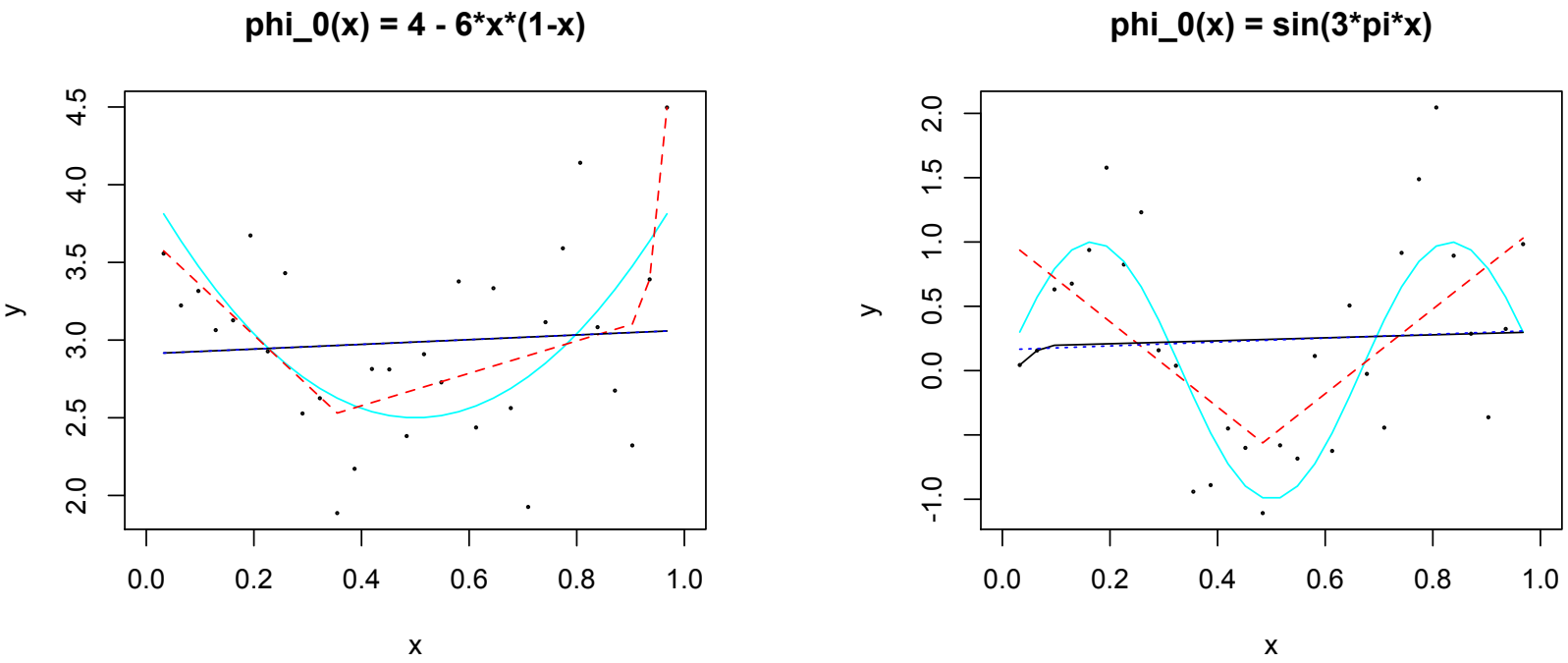


Figure: Constant (dotted blue), convex (dashed red) and concave (solid black) fits to a scatterplot generated from $Y = \phi_0(x) + \epsilon$ with independent standard normal errors.

Method

- Model: $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\theta}_0, \sigma^2 \mathbf{I}_n)$
- Consider testing $H_0 : \boldsymbol{\theta}_0 \in \mathcal{S}$ versus $H_1 : \boldsymbol{\theta}_0 \in \mathcal{I} \cup \mathcal{D} \backslash \mathcal{S}$
- $\mathcal{S}$ is a *linear subspace*.
- $\mathcal{I} \subset \mathbb{R}^n$ is a closed *convex cone* containing $\mathcal{S}$; $\mathcal{D} = -\mathcal{I}$.
- $\mathcal{I} \cap \mathcal{D} = \mathcal{S}$
- The log-likelihood function (up to a constant) is

$$\ell(\boldsymbol{\theta}, \sigma^2) = -\frac{1}{2\sigma^2} \|\mathbf{Y} - \boldsymbol{\theta}\|^2 - \frac{n}{2} \log \sigma^2.$$

- The *likelihood ratio* statistic is

$$\Lambda = 2 \left\{ \max_{\boldsymbol{\theta} \in \mathcal{I} \cup \mathcal{D}, \sigma^2 > 0} \ell(\boldsymbol{\theta}, \sigma^2) - \max_{\boldsymbol{\theta} \in \mathcal{S}, \sigma^2 > 0} \ell(\boldsymbol{\theta}, \sigma^2) \right\}$$

- An equivalent test is to *reject* H_0 if

$$T(\mathbf{Y}) := \frac{\max\{\|\hat{\boldsymbol{\theta}}_S - \hat{\boldsymbol{\theta}}_I\|^2, \|\hat{\boldsymbol{\theta}}_S - \hat{\boldsymbol{\theta}}_D\|^2\}}{\|\mathbf{Y} - \hat{\boldsymbol{\theta}}_S\|^2} \quad \text{is } \textit{large}$$

- $\hat{\boldsymbol{\theta}}_I = \Pi(\mathbf{Y}|\mathcal{I})$, $\hat{\boldsymbol{\theta}}_D = \Pi(\mathbf{Y}|\mathcal{D})$ and $\hat{\boldsymbol{\theta}}_S = \Pi(\mathbf{Y}|\mathcal{S})$.

Questions

- How do we find the *critical value* of the test?
- Power against all alternatives?

How to find the critical value of the test?

Result: The distribution of T is *invariant* to translations in $\mathcal{S}$.

Lemma (Sen and Meyer (2013))

For any $\mathbf{s} \in \mathcal{S}$ and $\mathbf{z} \in \mathbb{R}^n$,

$$T(\mathbf{z} + \mathbf{s}) = T(\mathbf{z}).$$

As a consequence, for any $\boldsymbol{\theta}_0 \in \mathcal{S}$,

$$T(\mathbf{Y}) \stackrel{d}{=} T(\mathbf{Z}),$$

where $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\theta}_0, \sigma^2 \mathbf{I}_n)$ and $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.

Proof

For $\mathbf{z} \in \mathbb{R}^n$ and $\mathbf{s} \in \mathcal{S}$,

$$\Pi(\mathbf{z} + \mathbf{s}|\mathcal{I}) = \Pi(\mathbf{z}|\mathcal{I}) + \mathbf{s}, \quad \text{and} \quad \Pi(\mathbf{z} + \mathbf{s}|\mathcal{S}) = \Pi(\mathbf{z}|\mathcal{S}) + \mathbf{s}.$$

Note that $\mathbf{Y} = \sigma \mathbf{Z} + \boldsymbol{\theta}_0$, where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.

- The test statistics T is *pivotal*, under H_0
- Suppose $T(\mathbf{Z}) \sim F$. Then, we *reject* H_0 if

$$T(\mathbf{Y}) > F^{-1}(1-\alpha),$$

where $\alpha \in (0,1)$ is the desired level of the test

- The distribution F can be approximated, up to any desired precision, by *Monte Carlo* simulations
- The test procedure has *exact* level α
- The test is also *unbiased*

Theorem (Sen and Meyer (2013))

Under mild conditions, the *power* of the test converges to *1*.

Example: Testing against simple linear regression

- $Y_i = \phi_0(x_i) + \epsilon_i; \quad 0 \leq x_1 < \cdots < x_n \leq 1$; $\epsilon_1, \ldots, \epsilon_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$
- $\phi_0 : [0, 1] \to \mathbb{R}$ is an unknown *smooth* function
- Test $H_0 : \phi_0$ is an *affine* function
- Recall $T(\mathbf{Y}) = \frac{\max\{\frac{1}{n}\|\hat{\boldsymbol{\theta}}_S - \hat{\boldsymbol{\theta}}_I\|^2, \frac{1}{n}\|\hat{\boldsymbol{\theta}}_S - \hat{\boldsymbol{\theta}}_D\|^2\}}{\frac{1}{n}\|\mathbf{Y} - \hat{\boldsymbol{\theta}}_S\|^2}$
- Under H_0, $T = O_p(\log n^{5/4}/n)$.
- Otherwise, $T \to_p c$, $c > 0$.
- Thus, $\mathbb{P}_{\phi_0}(T > F^{-1}(1 - \alpha)) \to 1$, if $\phi_0 \notin H_0$

Convex regression in 1-dimension

- $\phi_0 : [0,1] \to \mathbb{R}$ is an unknown *convex* function.
- Observe
$$Y_i = \phi_0(x_i) + \epsilon_i, \qquad \text{for } i = 1, 2, \ldots, n.$$
- $x_1 < x_2 < \cdots < x_n$ are fixed uniform grid points in $[0,1]$.
- $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$.
- Want to study the *LSE* $\hat{\phi}_{ls}$ of ϕ_0, i.e.,
$$\hat{\phi}_{ls} = \underset{\psi \text{ is convex}}{\operatorname{argmin}} \sum_{i=1}^{n} \{Y_i - \psi(x_i)\}^2.$$
- $\hat{\boldsymbol{\theta}} = (\hat{\phi}_{ls}(x_1), \ldots, \hat{\phi}_{ls}(x_n))$ is unique.

Theoretical results: Convex LSE when $d = 1$

Known results

- Asymptotic Behavior on Interior Compacta:
  - Consistency (Hanson and Pledger (1976))
  - Rate of convergence (Dümbgen et al. (2004))
- Asymptotic Behavior at a Point:
  - Rate of convergence (Mammen (1991))
  - Limiting distribution (Groeneboom et al. (2001))

- Not much is known about the *global* behavior of the LSE.

- Natural global loss function:

$$\ell^2(\psi, \phi) = \frac{1}{n} \sum_{i=1}^{n} \{\psi(x_i) - \phi(x_i)\}^2.$$

Worst case risk upper bound

- Let $\boldsymbol{\theta}_0 = (\phi_0(x_1), \ldots, \phi_0(x_n))$, $\hat{\boldsymbol{\theta}} = (\hat{\phi}_{ls}(x_1), \ldots, \hat{\phi}_{ls}(x_n))$
- For every ϕ_0, the *risk*

$$\mathbb{E}_{\phi_0} \ell^2(\hat{\phi}_{ls}, \phi_0) = \frac{1}{n} \mathbb{E}_{\boldsymbol{\theta}_0} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2,$$

is bounded from above by $n^{-4/5}$ up to logarithmic factors in n.

Risk upper bound (Guntuboyina and Sen (2013b))

We have

$$\mathbb{E}_{\phi_0} \ell^2(\hat{\phi}_{ls}, \phi_0) \lesssim n^{-4/5} \log n$$

- *No smoothness* assumptions on ϕ_0 necessary. Only convexity needed.
- *Finite sample* result, holds for every n.

- Question: *Can $\mathbb{E}_{\phi_0}\ell^2(\hat{\phi}_{ls}, \phi_0)$ be much smaller than $n^{-4/5}$?*

(1) **NO**: If $\phi_0''(x)$ exists and is bounded from above and below by positive constants on a subinterval of $(0, 1)$.

- Then, in a very strong sense, *no estimator* (not just the LSE) can have risk *smaller* than $n^{-4/5}$.

- The *local minimax risk* at ϕ_0:

$$R_n(\phi_0) := \inf_{\hat{\phi}} \sup_{\phi \in N(\phi_0)} \mathbb{E}_{\phi}\ell^2(\phi, \hat{\phi})$$

where $N(\phi_0) := \left\{\phi \text{ convex} : \|\phi - \phi_0\|_\infty \leq n^{-2/5}\right\}$.

Local minimax lower bound (Guntuboyina and Sen (2013b))

$$R_n(\phi_0) \gtrsim n^{-4/5}$$

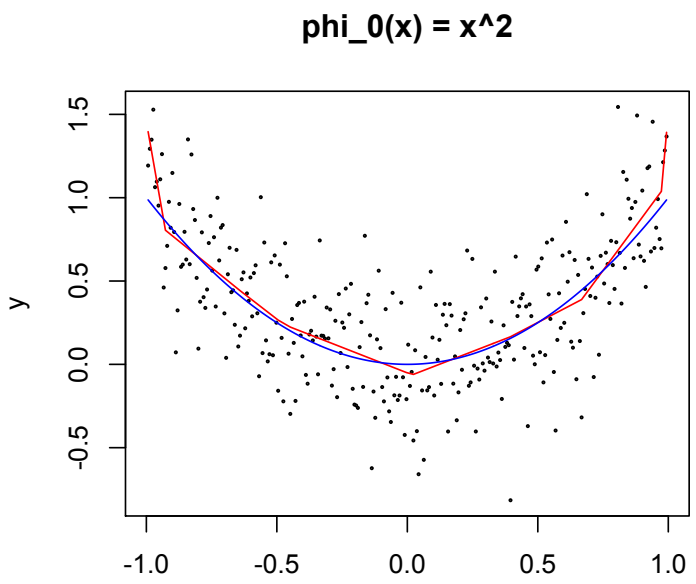
- Question: *Can $\mathbb{E}_{\phi_0}\ell^2(\hat{\phi}_{ls}, \phi_0)$ be much smaller than $n^{-4/5}$?*

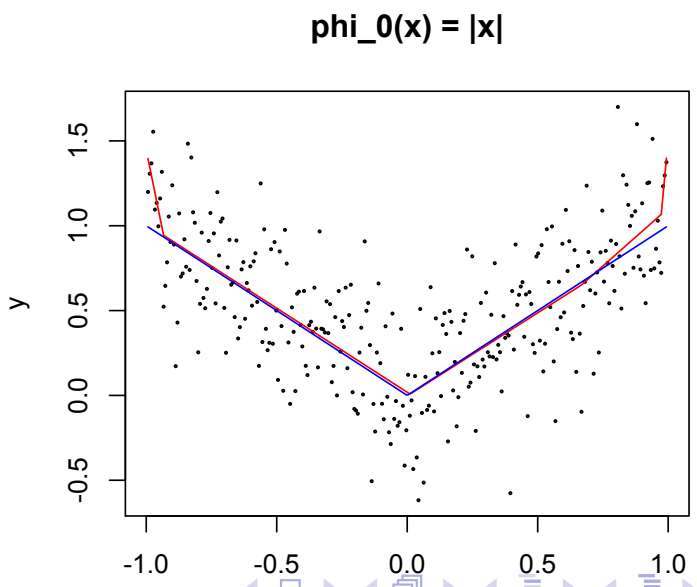
(2) **YES**: ϕ_0 is a piecewise affine convex function.

- *Adaptation*: Suppose ϕ_0 has *k affine pieces*, then

$$\mathbb{E}_{\phi_0}\ell^2(\hat{\phi}_{ls}, \phi_0) \lesssim \frac{k}{n}(\log n)^{5/4}.$$

- *Parametric rate* up to logarithmic factors.
- *Surprising* as the LSE minimizes the LS criterion over *all* convex fns.





Key proof ideas - Risk upper bounds

- Convex LSE is an *Empirical Risk Minimization* (ERM) procedure (Birgé and Massart (1993), Van de Geer, Van der Vaart and Wellner, Massart, Koltchinskii).

- Risk behavior of $\hat{\phi}_{ls}$ is determined by

$$\mathbb{E}\left[\sup_{\substack{\phi \text{ convex} \\ \ell^2(\phi,\phi_0)\leq r^2}} \sum_{i=1}^{n} \epsilon_i\{\phi(x_i)-\phi_0(x_i)\}\right]$$

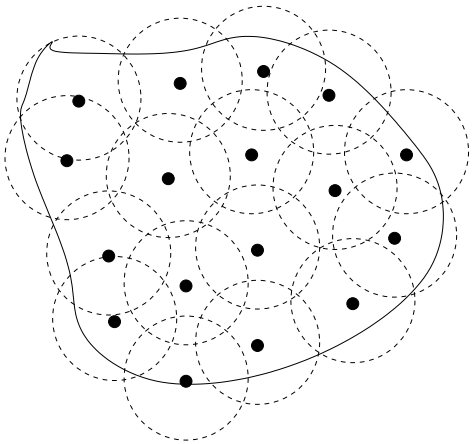
- *Local balls* of convex functions, defined as

$$S(\phi_0, r) := \{\phi \text{ convex} : \ell^2(\phi,\phi_0) \leq r^2\}$$

- Need good upper bounds on *expected maxima* of Gaussian processes.

- Using Dudley's theorem, we can bound the above if we can compute the *metric entropy* of these *local balls*

Key proof ideas - Risk upper bounds



- Recall: For a subset $\mathcal{F}$ of a metric space $(\mathcal{X}, \rho)$ the ϵ-*covering number* $N(\epsilon, \mathcal{F}, \rho)$ is defined as the *smallest* number of balls of radius ϵ whose union contains $\mathcal{F}$.
- Covering numbers capture the *size* of the underlying set $\mathcal{F}$.
- Metric entropy: *logarithm* of the covering numbers $\log N(\epsilon, \mathcal{F}, \rho)$
- We need good bounds on the *metric entropy* $\log N(\epsilon, S(\phi_0, r), \ell)$

Metric Entropy

- Bronšteĭn (1976) showed that the metric entropy of *uniformly Lipschitz* and *uniformly bounded* convex functions on $[0,1]^d$ under the supremum metric grows as $\epsilon^{-d/2}$.
- But functions in these local balls are neither uniformly bounded nor uniformly Lipschitz.
- We modify Bronshtein's result through a series of peeling arguments to obtain the metric entropy of these balls (Dryanov (2009) and Guntuboyina and Sen (2013a).

Ongoing research and future work

- Determine the *global* and *local* (point-wise) rates of convergence and the asymptotic distribution of the LSE in convex regression.
- Determine the *optimal* rates of convergence in these problems.
- How do we estimate and compute other multi-dimensional shape restricted functions, e.g., *multi-dimensional* monotone regression, quasi convex functions, etc.?
- Risk bounds for a general *polyhedral cone* (connection to non-negative least squares)?
- Investigate *additive models* with unknown monotone/convex functions.
- Study *single index models* with unknown monotone/convex links.

Collaborators

Emilio Seijo
former Ph.D. student
Start-up

Mary Meyer
Professor
Colorado State Uni.

Aditya Guntuboyina
Assistant Professor
UC Berkeley

and others ...

References

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.*, 26:641–647.

Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150.

Bronšteĭn, E. M. (1976). ε-entropy of convex sets and functions. *Sibirsk. Mat. Ž.*, 17(3):508–514, 715.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, 26:607–616.

Dryanov, D. (2009). Kolmogorov entropy for classes of convex functions. *Constructive Approximation*, 30:137–153.

Dümbgen, L., Freitag, S., and Jongbloed, G. (2004). Consistency of concave regression with an application to current-status data. *Math. Methods Statist.*, 13(1):69–81.

Groeneboom, P., Jongbloed, G., and Wellner, J. A. (2001). Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.*, 29(6):1653–1698.

Guntuboyina, A. and Sen, B. (2013a). Covering numbers for convex functions. *IEEE Trans. Inf. Th.*, 59(4):1957–1965.

Guntuboyina, A. and Sen, B. (2013b). Global risk bounds and adaptation in univariate convex regression. arXiv preprint arXiv:1305.1648.

Hanson, D. L. and Pledger, G. (1976). Consistency in concave regression. *Ann. Statist.*, 4(6):1038–1050.

Hildreth, C. (1954). Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.*, 49:598–619.

Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.*, 19(2):741–759.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.

Seijo, E. and Sen, B. (2011). Nonparametric least squares estimation of a multivariate convex regression function. *Ann. Statist.*, 39(3):1633–1657.

Sen, B. and Meyer, M. (2013). Testing against a parametric regression function using ideas from shape restricted estimation. arXiv preprint arXiv:1311.6849.

Thank you very much!

Questions?