# Totally Convex Regression

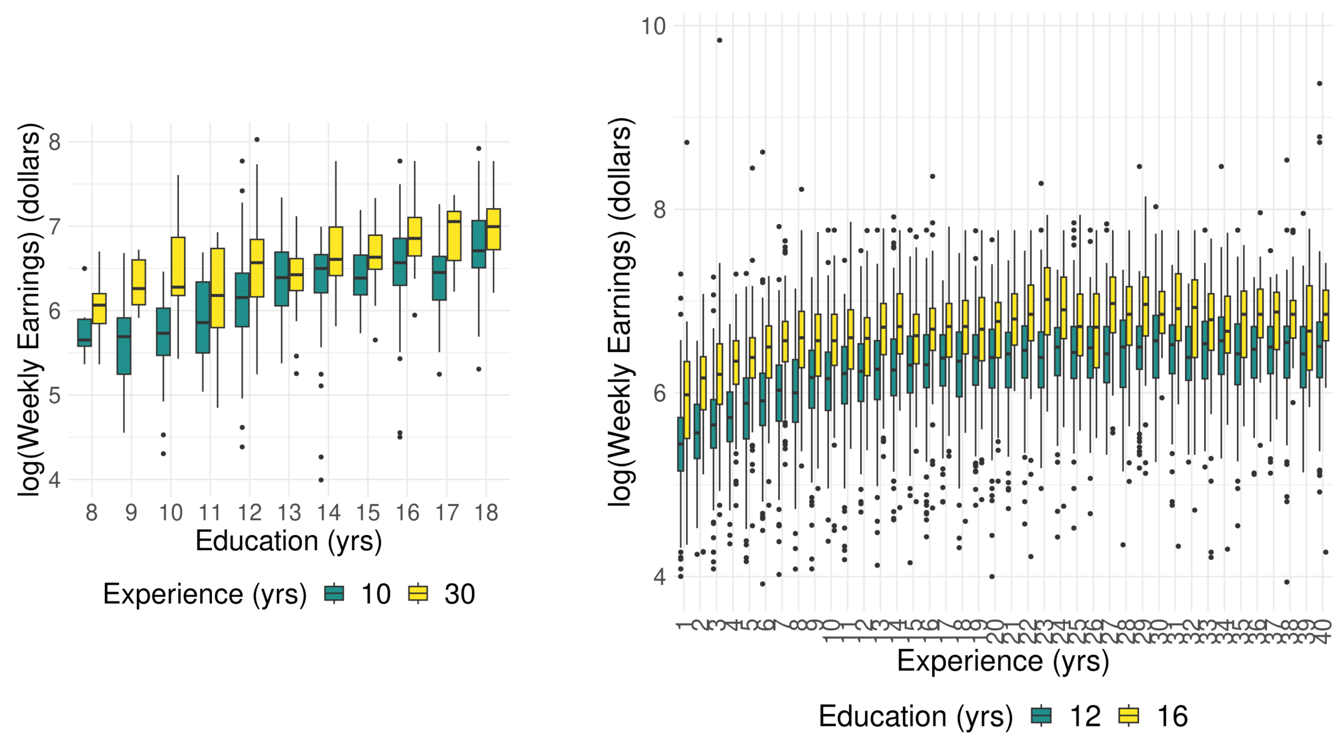
Dohyeong Ki & Adityanand Guntuboyina
Department of Statistics, University of California, Berkeley

## Motivating Example

Earnings dataset (ex1029 from R package Sleuth3):

The weekly earnings of 20,967 full-time non-black male workers in 1987 along with their years of education ( $\geq 8$) and years of experience ($1 \leq \cdot \leq 40$).

We predict the log of Earnings ($y$) from Education ($x_1$) and Experience ($x_2$) using convex/concave relations between them.




There is a clear concave relation between log of Earnings and Experience. But it is not clear whether we have a convex or concave relation between log of Earnings and Education.

Also, we can see that we need to consider interaction between Education and Experience (see also [Lemieux 2006] and references therein).

If log of Earnings ($y$) is an additive function of Education ($x_1$) and Experience ($x_2$), i.e., $y = f(x_1, x_2) = f_1(x_1) + f_2(x_2)$, then $f(x_1, z_2) - f(x_1, y_2) = f_2(z_2) + f_2(y_2)$ is constant in $x_1$. But we can see from the left plot, it is not the case for our data.

Question. How can we fit to the data a function $y = f(x_1, x_2)$ that is coordinate-wise concave in $x_1$ and coordinate-wise convex or concave in $x_2$ and also models interaction between $x_1$ and $x_2$?

## Additive Convex Regression

Suppose we are given $(x^{(1)}, y_1), \ldots, (x^{(n)}, y_n)$ ($x^{(i)} \in [0,1]^d, y_i \in \mathbb{R}$) and we want to fit a coordinate-wise convex (or concave) function $y = f(x)$ to the data.

If we don't need to consider interaction between predictors, a natural choice is additive convex regression, which restricts to functions of the form

$$f(x_1, \ldots, x_d) = f_1(x_1) + \cdots + f_d(x_d)$$

where $f_1, \ldots, f_d$ are univariate convex (or concave) functions.

Also, it is known that we just need to search each $f_k$ among linear combinations of 1 and

$$(\cdot - t)_+ := \max(\cdot - t, 0),\ t \in [0,1)$$

whose coefficient is nonnegative unless $t = 0$ (see, e.g., [Guntuboyina 2015]).

However, as we need to consider interaction between predictors, additive convex regression is not enough for our purpose.

## Interaction

How do we extend linear regression to take interaction into consideration?

→ We simply add products of predictors to linear regression models.

Example) Two-way interactions, Three-way interaction

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{23} x_2 x_3 + \beta_{31} x_3 x_1 + \beta_{123} x_1 x_2 x_3$$

Recall that for univariate convex (resp., concave) regression, basis functions are

$$1 \text{ and } (\cdot - t)_+,\ t \in [0,1)$$

and the coefficients of $(\cdot - t)_+$ for $t \in (0,1)$ is nonnegative (resp., nonpositive).

We can thus model $m$-way interactions with

$$\prod_{k \in S} (x_k - t_k)_+ \quad \text{for } S \subseteq \{1, \ldots, d\} \text{ with } |S| = m.$$

What about the signs of coefficients?

Our top priority is to keep coordinate-wise convexity (or concavity).

Observe that

$$\beta \cdot \prod_{k \in S} (x_k - t_k)_+ = \begin{cases} \text{coordinate-wise convex} & \text{if } \beta \geq 0 \\ \text{coordinate-wise concave} & \text{if } \beta \leq 0. \end{cases}$$

Hence, for example, if $d = 2$, in each case below, we can model interaction between $x_1$ and $x_2$ only with

- Coordinate-wise convexity (resp., concavity) both in $x_1$ and $x_2$

$$\beta x_1 x_2,\ \beta \in \mathbb{R} \text{ and } \beta (x_1 - t_1)_+ (x_2 - t_2)_+,\ \beta \geq 0 \text{ (resp., } \beta \leq 0)$$

- Mixed coordinate-wise convexity and concavity in $x_1$ and $x_2$

$$\beta x_1 x_2,\ \beta \in \mathbb{R}$$

→ As the second case is relatively simple, we focus on the first case from now on.

## Function Class

Our function class $\mathcal{F}^{d,s}_{TC}$ is defined as the collection of functions

$$(x_1, \ldots, x_d) \mapsto \sum_{\substack{S \subseteq \{1, \ldots, d\} \\ |S| \leq s}} a_S \cdot \prod_{k \in S} x_k + \sum_{\substack{S \subseteq \{1, \ldots, d\} \\ 0 < |S| \leq s}} \int_{[0,1)^{|S|} \setminus \{(0,\ldots,0)\}} \prod_{k \in S} (x_k - t_k)_+ \, d\nu_S(t_k, k \in S)$$

infinite linear combinations via measures

where $a_S \in \mathbb{R}$ and $\nu_S$ is a positive measure on $[0,1)^{|S|} \setminus \{(0,\ldots,0)\}$ for each subset $S$ of $\{1, \ldots, d\}$ with $|S| \leq s$. Here, $s$ is a restriction on interaction order.

$\mathcal{F}^{2,2}_{TC}$ is essentially the class of totally convex functions, originally introduced by T. Popoviciu and more recently described in [Gal 2010].
This is why we call our method totally convex regression.

## Estimator

Our estimator is then defined by

$$\hat{f}^{d,s}_n \in \underset{f}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - f(x^{(i)}))^2 : f \in \mathcal{F}^{d,s}_{TC} \right\}.$$

↖ no tuning parameter

Alternative characterization via constraints on derivatives:

$$\hat{f}^{d,d}_n \in \underset{f}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - f(x^{(i)}))^2 : \frac{\partial^{p_1 + \cdots + p_d} f}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \geq 0 \text{ for every } (p_1, \ldots, p_d) \in \{0,1,2\}^d \text{ with } \max_k p_k = 2 \right\}.$$

Observe that the total order $p_1 + \cdots + p_d$ of derivatives is as high as $2d$, but we take at most two derivatives along each coordinate.

This characterization is in fact not fully rigorous as second-order derivatives may not even exist. A rigorous version can be obtained by instead restricting first-order derivatives to be monotonic.

## Computation

We can search $\hat{f}^{d,s}_n$ over finite linear combinations of

$$(x_1, \ldots, x_d) \mapsto \prod_{k \in S} (x_k - t_k)_+,\ |S| \leq s,$$

where

$$t_k \in \{0\} \cup \{x^{(i)}_k : 1 \leq i \leq n\}$$

↖ observed $k^{\text{th}}$ components

and the corresponding coefficient is nonnegative unless $(t_k, k \in S) = (0, \ldots, 0)$.

We can also approximate $\hat{f}^{d,s}_n$ by restricting $t_k$ instead to a set of manageable size; e.g., $t_k \in \{0, .05, .10, \ldots, 1\}$.

## Theoretical Results

Under the standard set of model assumptions:

(1) $y_i = f^*(x^{(i)}) + \epsilon_i$ where $f^* \in \mathcal{F}^{d,s}_{TC}$ and $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$

(2) $x^{(1)}, \ldots, x^{(n)}$ form an equally-spaced lattice,

we have

$$\mathbb{E}\left[ \frac{1}{n} \sum_{i=1}^{n} \left( \hat{f}^{d,s}_n(x^{(i)}) - f^*(x^{(i)}) \right)^2 \right] = O\left( n^{-\frac{4}{5}} (\log n)^{\frac{3(2s-1)}{5}} \right).$$

→ Our estimator $\hat{f}^{d,s}_n$ can avoid the usual curse of dimensionality to some extent.

## References

[Gal 2010] Gal, S. G. (2010) *Shape-preserving approximation by real and complex polynomials.* Springer.

[Lemieux 2006] Lemieux, T. (2006) The "Mincer equation" thirty years after schooling, experience, and earnings. In *Jacob Mincer a pioneer of modern labor economics* (pp. 127-145). Springer.

[Guntuboyina 2015] Guntuboyina, A. & Sen, B. (2015) Global risk bounds and adaptation in univariate convex regression. *Probability Theory and Related Fields*, 163(1), 379-411.

dohyeong_ki@berkeley.edu