Auto-tuned high-dimensional regression with the TREX: theoretical guarantees and non-convex global optimization

Jacob Bien[1], Irina Gaynanova[1], Johannes Lederer[1], Christian L. Müller[2,3]
[1]Cornell University, Ithaca [2]New York University, [3]Simons Center for Data Analysis, New York

Summary

- **Lasso** [1] is a popular method for high-dimensional variable selection, but **difficult** to tune in practice.
- We introduce **TREX** [2], an alternative to Lasso, that **does not require tuning parameters**.
- **TREX** can **outperform cross-validated Lasso** in terms of variable selection and computational efficiency.
- We derive **proofs** for the **prediction error** of TREX under mild assumptions on the linear regression model.
- The **non-convex** TREX objective can be **globally optimally** solved using **Second-Order Cone Programming (SOCP)**.
- The geometry of the TREX objective function provides further valuable insights for the variable selection process.

From Lasso to TREX

We aim at variable selection in linear regression.
We therefore consider models of the form

$$Y = X\beta^* + \sigma\epsilon, \qquad \text{(Model)}$$

where $Y \in \mathbb{R}^n$ is a response vector, $X \in \mathbb{R}^{n\times p}$ a design matrix, $\sigma > 0$ a constant, and $\varepsilon \in \mathbb{R}^n$ a noise vector.

Lasso

$$\widehat{\beta}_{\text{Lasso}}(\lambda) \in \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \left\{ \frac{\|Y - X\beta\|_2^2}{n} + \lambda\|\beta\|_1 \right\}. \qquad \text{(Lasso)}$$

Lasso [1] requires the tuning of the regularization parameter λ via **heuristic** methods such as cross-validation (CV) or the Bayesian information criterion (BIC).

Theory suggest to choose:

$$\lambda \sim \frac{\sigma\|X^\top\epsilon\|_\infty}{n}.$$

Sqrt-Lasso

$$\widehat{\beta}_{\sqrt{\text{Lasso}}}(\gamma) \in \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \left\{ \frac{\|Y - X\beta\|_2}{\sqrt{n}} + \gamma\|\beta\|_1 \right\}. \qquad \text{(Square-Root Lasso)}$$

Sqrt-Lasso [3,4] simultaneously estimates the unknown noise variance σ but still needs to select a tuning parameter γ via CV or BIC.

Theory suggest to choose:

$$\gamma \sim \frac{\|X^\top\epsilon\|_\infty}{n},$$

TREX idea

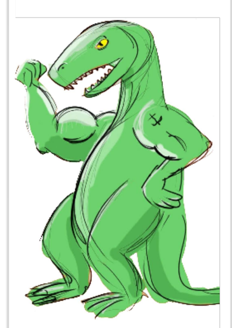
Incorporate an inherent estimation of the **entire** quantity of interest

$$\sigma\|X^\top\epsilon\|_\infty/n$$

into the estimator!

TREX objective and solution

TREX objective



We use the fact that if $\hat\beta$ is a consistent estimator of β^* then $\sigma\|X^T(Y - X)\hat\beta)\|_\infty/n$ is a consistent estimator of $\sigma\|X^T\epsilon\|_\infty/n$. We thus define the TREX:

$$\widehat{\beta}_{\text{TREX}} \in \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \left\{ \frac{\|Y - X\beta\|_2^2}{\frac{1}{2}\|X^\top(Y - X\beta)\|_\infty} + \|\beta\|_1 \right\}. \qquad \text{(TREX)}$$

Fast approximate numerical solution

- The data-fitting term $L(\beta) = \frac{\|Y-X\beta\|_2^2}{\frac{1}{2}\|X^T(Y-X\beta)\|_\infty}$ of the non-smooth TREX objective function $f_{\text{TREX}} = L(\beta) + \|\beta\|_1$ is approximated by the smooth term $\overline{L}(\beta) = \frac{\|Y-X\beta\|_2^2}{\frac{1}{2}\|X^T(Y-X\beta)\|_q}$.
- In practice, for any $q > 10$, the function $\overline{L}(\beta) + \|\beta\|_1$ is a sufficient approximation to f_{TREX} and can be efficiently minimized with projected scaled sub-gradient algorithms [5].

Exact solution using SOCP techniques

Reformulate the TREX objective (e.g., with a=1/2) as:

$$P^* := \min_{\beta\in\mathbb{R}^p} \left\{ \frac{\|Y - X\beta\|^2}{\max_{j\in\{1,\dots,p\}} a|x_j^\top(Y - X\beta)|} + \|\beta\|_1 \right\}$$

$$= \min_{\beta\in\mathbb{R}^p} \min_{j\in\{1,\dots,p\}} \left\{ \frac{\|Y - X\beta\|^2}{a|x_j^\top(Y - X\beta)|} + \|\beta\|_1 \right\}.$$

This leads to p pairs of problems of the form:

$$\min_{\beta\in\mathbb{R}^p} \left\{ \frac{\|Y - X\beta\|^2}{a x_j^\top(Y - X\beta)} + \|\beta\|_1 \quad \text{s.t.} \quad x_j^\top(Y - X\beta) \geq 0 \right\}$$

and

$$\min_{\beta\in\mathbb{R}^p} \left\{ \frac{\|Y - X\beta\|^2}{-a x_j^\top(Y - X\beta)} + \|\beta\|_1 \quad \text{s.t.} \quad -x_j^\top(Y - X\beta) \geq 0 \right\}.$$

which have the common form for p-dimensional vectors v:

$$P^*(v) := \min_{\beta\in\mathbb{R}^p} \left\{ \frac{\|Y - X\beta\|^2}{v^\top(Y - X\beta)} + \|\beta\|_1 \quad \text{s.t.} \quad v^\top(Y - X\beta) \geq 0 \right\}.$$

This is a standard quadratic over linear problem which can be solved by SOCP techniques. We currently use the *embedded conic solver* (ECOS) [6] to compute all 2p TREX problems.

Statistical guarantees for the TREX

We are able to derive statistical guarantees for the TREX. We provide bounds for the *prediction performance* in relationship to the Lasso and derive *slow-rate bounds* with no assumptions on the design matrix X in [7].

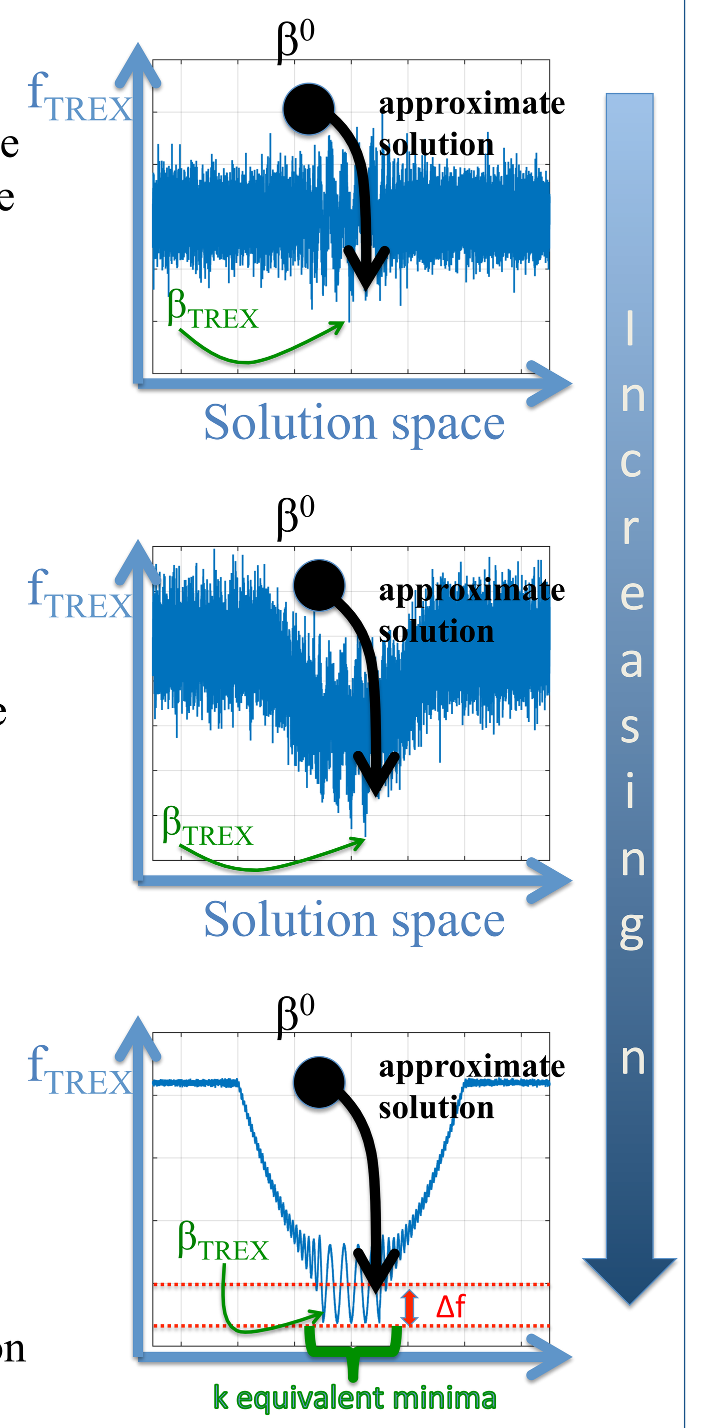
Illustrating the geometry of the TREX

We illustrate the global **funnel structure** of the TREX objective function with in-creasing sample size n, fixed p, and sparsity k, where $|\beta_j|=1$ for all k non-zero indices j.

The x axis represents the p-dimensional solution space and the y axis the TREX objective function value f_{TREX}.

Using the SOCP formulation we enumerate **all** 2p local minima, including the global minimum β_{TREX}. The approximate solver starts at a sparse solution (e.g., the **all zeros** vector β^0) and proceeds to a **local** minimum (black arrow).

Already at moderate n, we observe a function gap Δf between k equivalent TREX solutions (that are conditioned on the corresponding non-zero j's) and all other local solutions.



The structure of the objective function provides an alternative variable selection method.

TREX phase transition

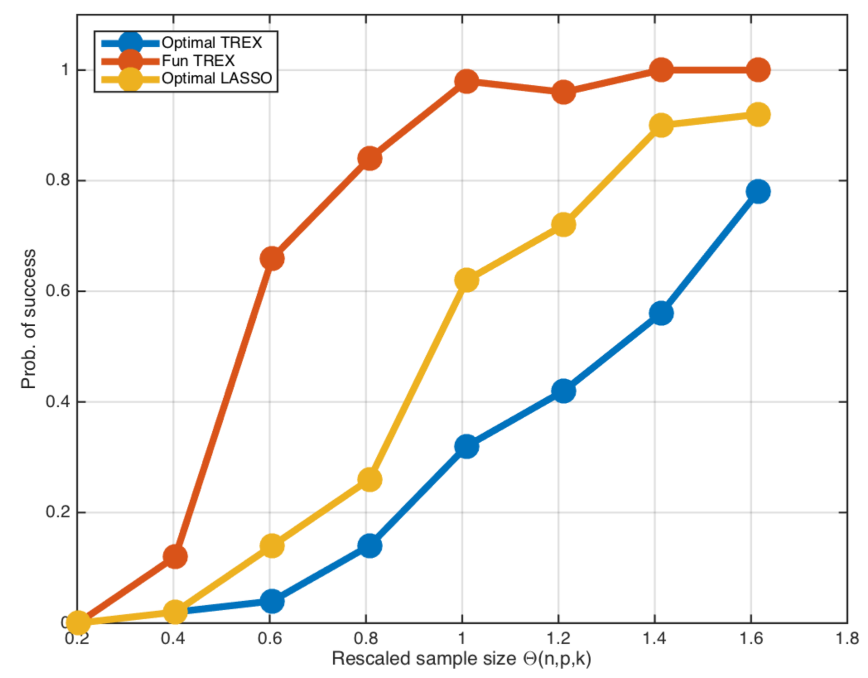


Figure 1: Success probability $P[S\pm(\beta) = S\pm(\beta^*)]$ of obtaining the correct signed support versus the rescaled sample size $\theta(n, p, k) = n/[2k \log(p - k)]$ for problem size p=64 with sparsity $k = \lceil 0.20\, p^{0.75} \rceil$. The number of repetitions is 50. The optimal a in TREX is in [0.45,0.5]. The lambda in LASSO is automatically determined by MATLAB. Variable selection using the function gap property (Fun TREX) is shown in red.

Ongoing work and improvements

- **Theoretical guarantees** for variable selection with **TREX**.
- Theoretical analysis of the TREX function gap phenomenon
- Improving the efficiency of SOCP solvers for Big Data
- **TREX** as building block for **GTREX,** an adaptive neighbor-hood selection scheme for graphical model inference [8].

Contact

Christian L. Müller, PhD
Simons Center for Data Analysis, New York
Email: cmueller@simonsfoundation.org
Website: www.simonsfoundation.org
SIMONS FOUNDATION

References

1. Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* 58(1):267–288.
2. Lederer, J., and Müller, C.L. 2015. Don't Fall for Tuning Parameters: Tuning-Free Variable Selection in High Dimensions With the TREX. Proc. 29th AAAI Conference on Artificial Intelligence
3. Belloni, A.; Chernozhukov, V.; and Wang, L. 2011. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* 98(4):791–806.
4. Sun, T.; and Zhang, C. 2012. Scaled sparse linear regression. *Biometrika* 99(4):879–898.
5. Schmidt, M. 2010. *Graphical Model Structure Learning with L1-Regularization*. Ph.D. Dissertation, University of British Columbia.
6. Domahidi, A.; Chu, E.; and Boyd S. 2013. ECOS: An SOCP Solver for Embedded Systems, *European Control Conference (ECC), 2013*
7. Bien, J.; Gaynanova, I.; Lederer, J.; and Müller, C.L. 2015. Auto-tuned high-dimensional regression: theoretical guarantees and non-convex global optimization (in preparation)
8. Lederer, J., and Müller, C.L. 2014. Topology Adaptive Graph Estimation in High Dimensions. *preprint, arxiv:1410.7279.*