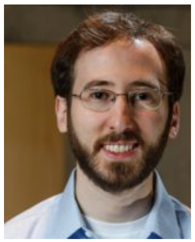
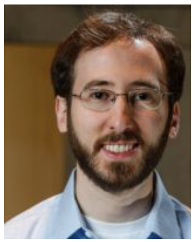
Auto-tuned high-dimensional regression with the TREX: theoretical guarantees and non-convex global optimization

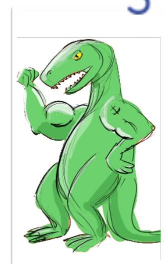
Jacob Bien[1], Irina Gaynanova[1], Johannes Lederer[1], Christian L. Müller[2,3]
[1]Cornell University, Ithaca [2]New York University, [3]Simons Center for Data Analysis, New York
SIMONS FOUNDATION



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

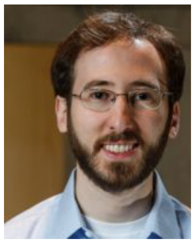


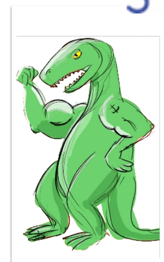



Auto-tuned high-dimensional regression with the TREX: theoretical guarantees and non-convex global optimization

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

SIMONS FOUNDATION



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

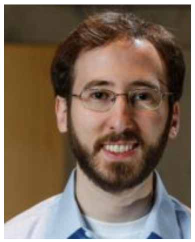
$$Y = X\beta^* + \sigma\epsilon, \qquad \text{(Model)}$$

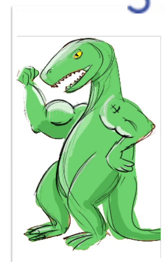
where $Y \in \mathbb{R}^n$ is a response vector, $X \in \mathbb{R}^{n\times p}$ a design matrix, $\sigma > 0$ a constant, and $\varepsilon \in \mathbb{R}^n$ a noise vector.

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

SIMONS FOUNDATION



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

$p >> n$

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

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.

High-dimensional variable selection in linear regression

$$n\left\{\boldsymbol{Y}\right. = \boldsymbol{X} \times \left.\boldsymbol{\beta}^*\right\}p + \boldsymbol{\sigma}\,\boldsymbol{\epsilon}$$

Standard approach: The LASSO (Tibshirani, 1996)

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

Standard approach: The LASSO (Tibshirani, 1996)

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

+ convex optimization problem
+ good statistical properties
- Tuning of regularization parameter required

Standard approach: The LASSO (Tibshirani, 1996)

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

+ convex optimization problem
+ good statistical properties
- Tuning of regularization parameter required

Standard approach: The LASSO (Tibshirani, 1996)

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

+ convex optimization problem
+ good statistical properties
- Tuning of regularization parameter required

Novel proposition: The TREX (Lederer and M., AAAI 2015)

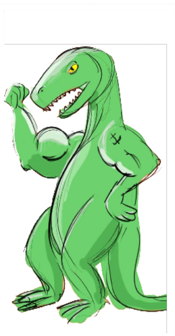
$$\widehat{\beta}_{\text{TREX}} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{\|Y - X\beta\|_2^2}{\frac{1}{2}\|X^\top(Y - X\beta)\|_\infty} + \|\beta\|_1 \right\}. \quad \textbf{(TREX)}$$

Standard approach: The LASSO (Tibshirani, 1996)

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

+ convex optimization problem
+ good statistical properties
- Tuning of regularization parameter required

Novel proposition: The TREX (Lederer and M., AAAI 2015)



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

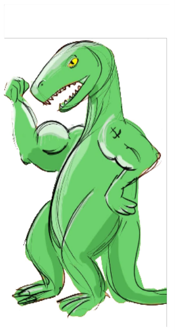
+ good statistical properties
+ Tuning-free method
- non-convex optimization problem

Standard approach: The LASSO (Tibshirani, 1996)

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

+ convex optimization problem
+ good statistical properties
- Tuning of regularization parameter required

Novel proposition: The TREX (Lederer and M., AAAI 2015)



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

+ good statistical properties
+ Tuning-free method
- non-convex optimization problem

Standard approach: The LASSO (Tibshirani, 1996)

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

+ convex optimization problem
+ good statistical properties
- Tuning of regularization parameter required

Novel proposition: The TREX (Lederer and M., AAAI 2015)

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

+ good statistical properties
+ Tuning-free method
- non-convex optimization problem

BUT...

The non-convex TREX objective function can be globally optimally solved by using Second Order Cone Programming.

The non-convex TREX objective function can be globally optimally solved by using Second Order Cone Programming.

1.) The TREX (with e.g. constant a=0.5) can be written as:

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

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

2.) For each index j this leads to a pair of problem 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

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

2.) For each index j this leads to a pair of problem 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

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

3.) or, in general, 2p problems of the quadratic over linear form:

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

2.) For each index j this leads to a pair of problem 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

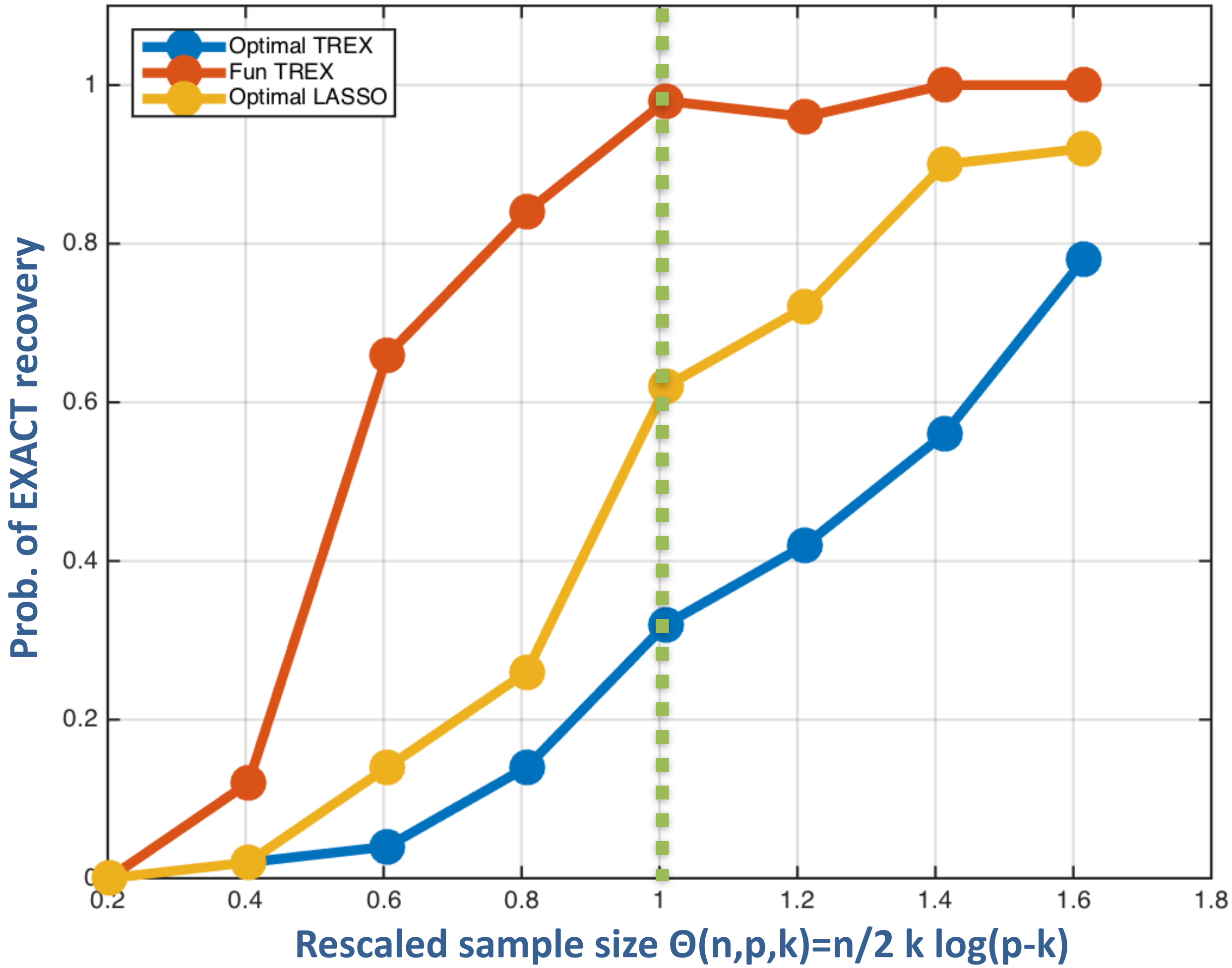
$$\min_{\beta\in\mathbb{R}^p} \left\{ \frac{\|Y - X\beta\|^2}{-a x_j^\top (Y - X\beta)} + \|\beta\|_1 \quad \text{s.t.} \quad -x_j^\top (Y - X\beta) \geq 0 \right\}.$$

3.) or, in general, 2p problems of the quadratic over linear form:

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

Each problem is a Second-Order Cone Program!

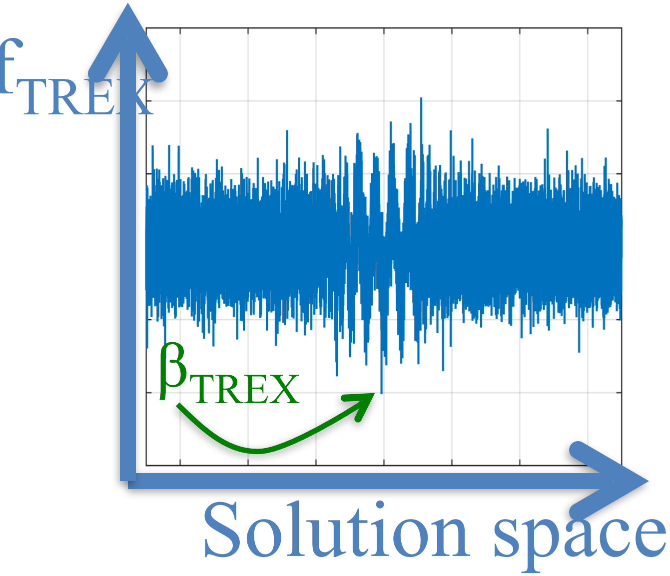
Phase transition of exact recovery with the TREX and the LASSO



p=64

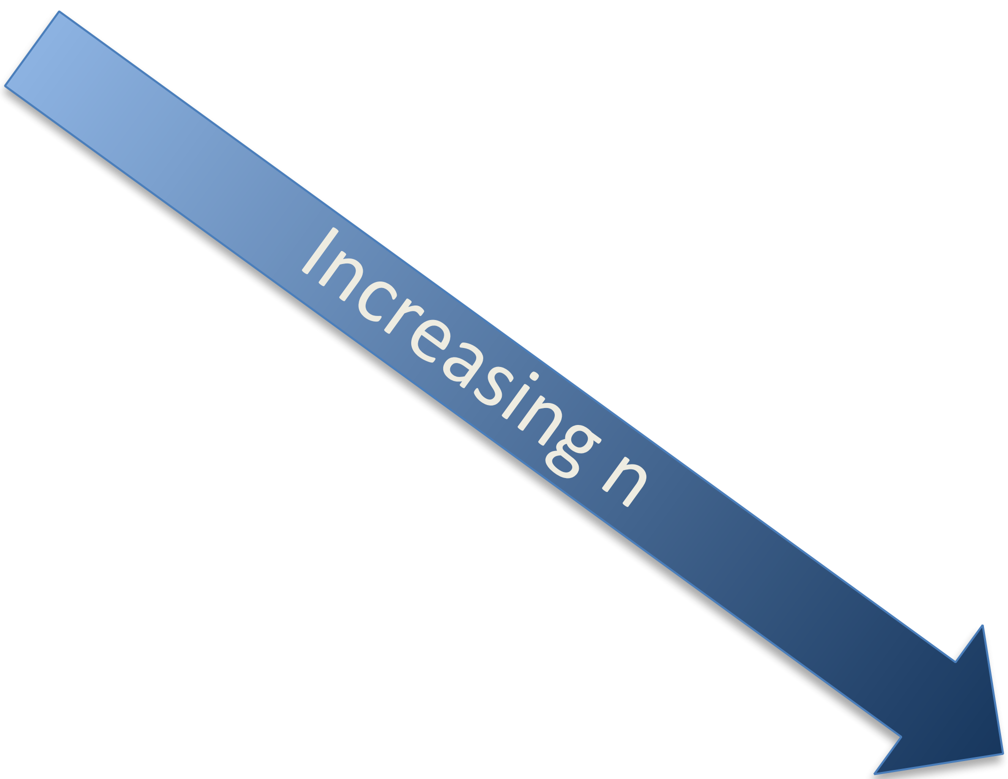
$k=0.2p^{0.75}$

Figure 1: Success probability P[S±(β) = S±(β∗)] of obtaining the correct signed support versus the rescaled sample size θ(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=0.5 in TREX. The lambda in LASSO is automatically determined by MATLAB. Variable selection using the function gap property (Fun TREX) is shown in red.



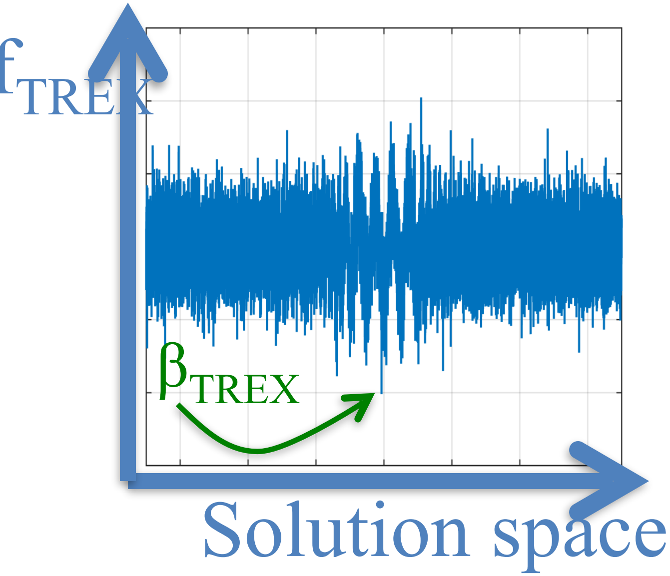
Sketching the topology of the TREX

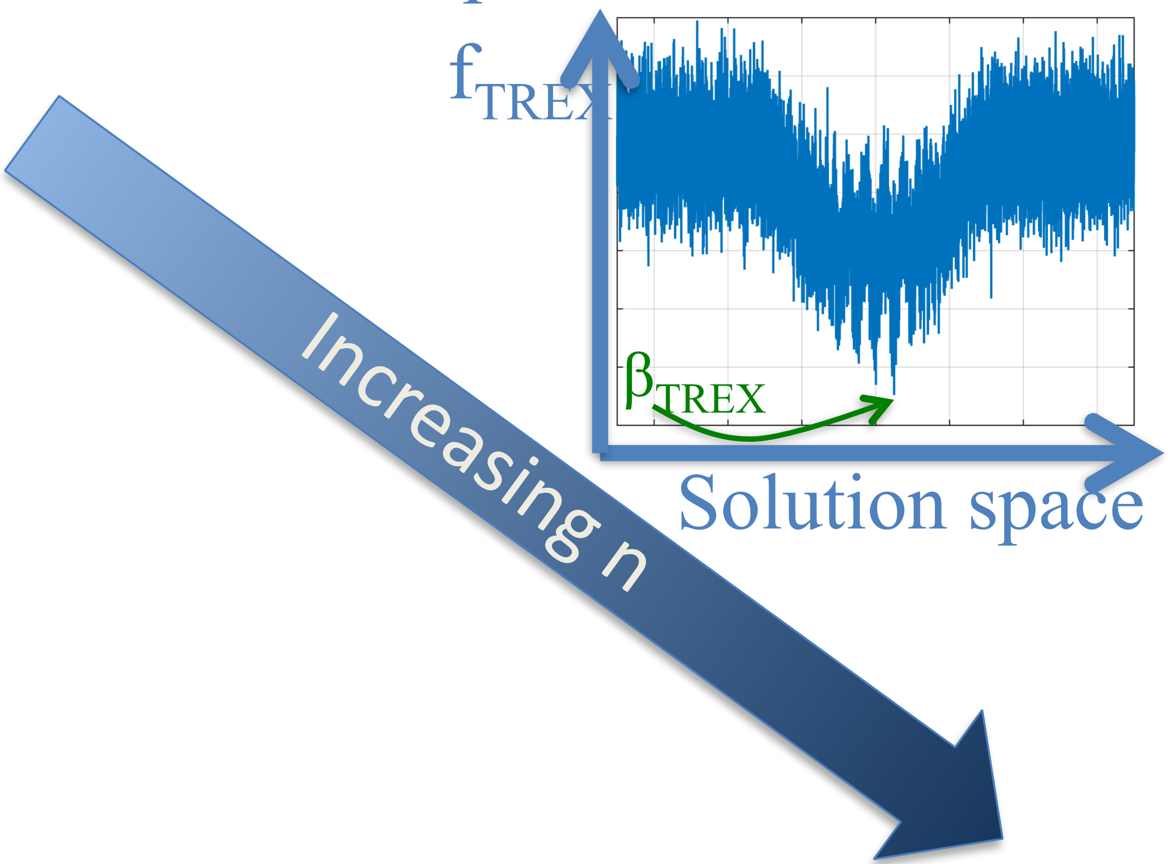
Consider the case where data (p>>n) are generated from a linear model with a sparse β vector with k<<p non-zero entries of equal absolute value.



Sketching the topology of the TREX

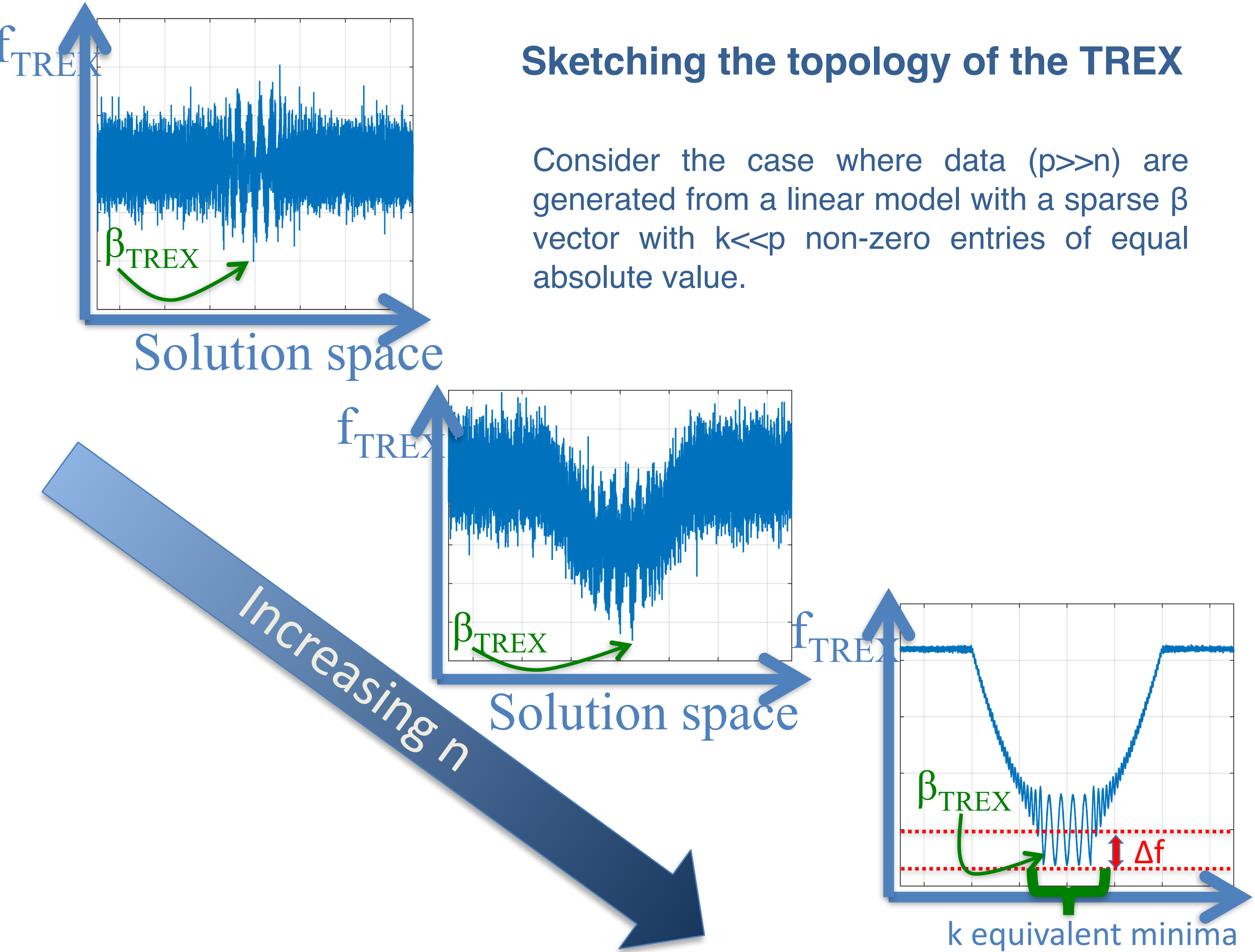
Consider the case where data (p>>n) are generated from a linear model with a sparse β vector with k<<p non-zero entries of equal absolute value.



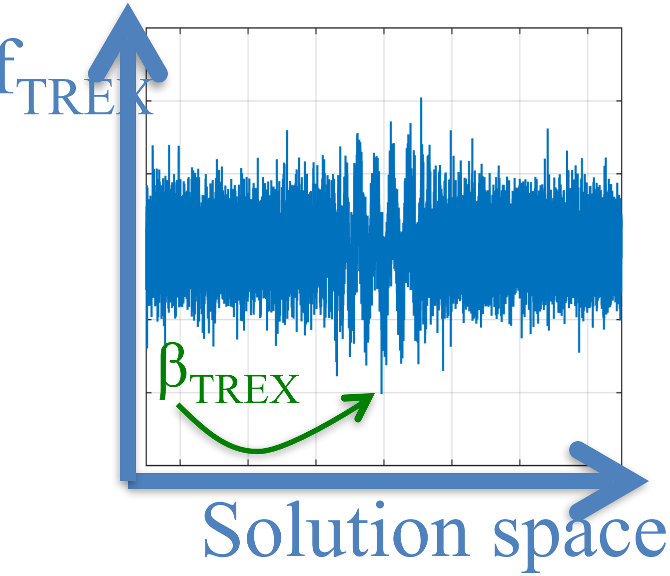


Sketching the topology of the TREX

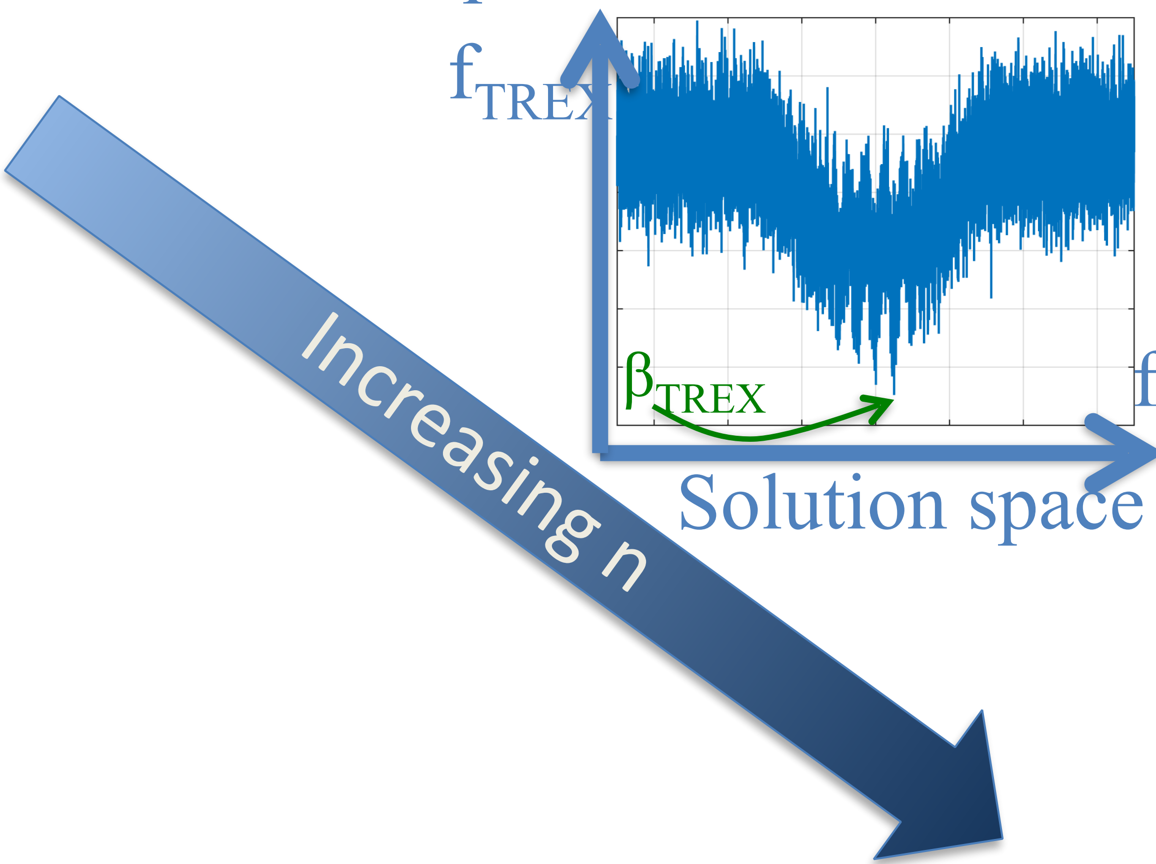
Consider the case where data (p>>n) are generated from a linear model with a sparse β vector with k<<p non-zero entries of equal absolute value.



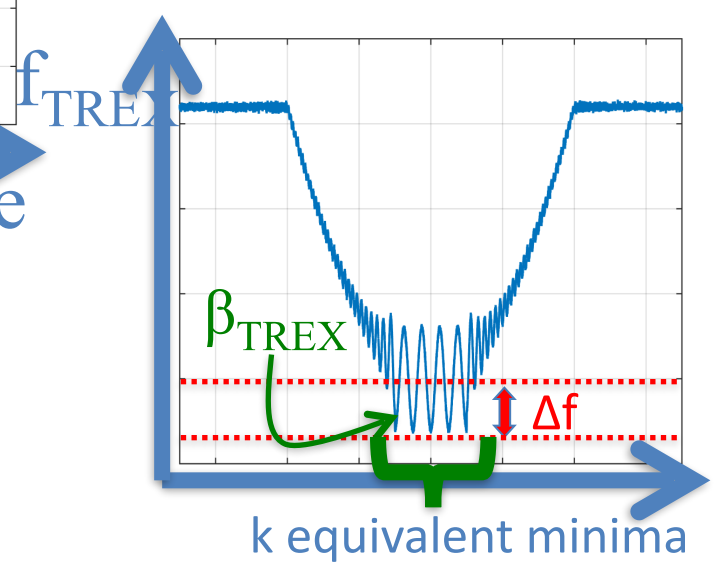
Sketching the topology of the TREX



Consider the case where data (p>>n) are generated from a linear model with a sparse β vector with k<<p non-zero entries of equal absolute value.



The topology of the objective function can be used as an alternative variable selection method.



How can we scale the TREX to BIG DATA?

How can we scale the TREX to BIG DATA?

Current solvers for SOCP

+ ECOS solver (Interior-Point method)
+ SCS solver (ADMM scheme)

How can we scale the TREX to BIG DATA?

Current solvers for SOCP

+ ECOS solver (Interior-Point method)
+ SCS solver (ADMM scheme)

Current solvers for local minimization of non-convex TREX function (smooth-non-convex + L1)

+ Projected scaled sub-gradient method (Mark Schmidt's code)
+ Orthant-wise L-BFGS
+ Proximal gradient (Jason Lee's package)

How can we scale the TREX to BIG DATA?

Current solvers for SOCP

+ ECOS solver (Interior-Point method)
+ SCS solver (ADMM scheme)

Current solvers for local minimization of non-convex TREX function (smooth-non-convex + L1)

+ Projected scaled sub-gradient method (Mark Schmidt's code)
+ Orthant-wise L-BFGS
+ Proximal gradient (Jason Lee's package)

ANY IDEA HOW TO SPEED THINGS UP?