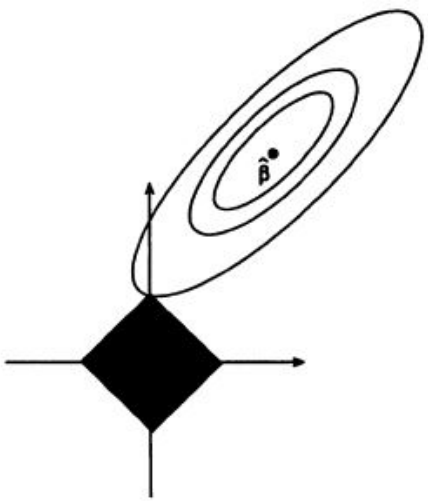
Root systems in modern statistics

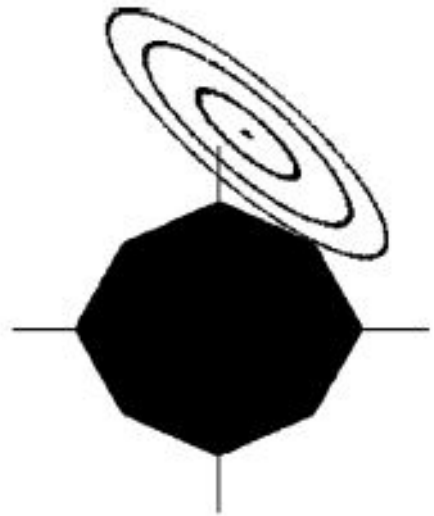
Tomasz Skalski

(joint research with M. Bogdan, P. Graczyk,
B. Kołodziejek, P. Tardivel and M. Wilczyński)

Politechnika Wrocławska

Université d'Angers





Linear regression model

Linear regression model: $Y = X\beta + \varepsilon$:

- $Y \in \mathbb{R}^n$: response vector
- $X \in \mathbb{R}^{n \times p}$: design matrix
- $\beta \in \mathbb{R}^p$: (unknown) vector of regression coefficients
- $\varepsilon \in \mathbb{R}^n$: noise term

Ordinary Least Squares (OLS) estimator

The Ordinary Least Squares estimator (Legendre, 1805, Gauss, 1809) minimizes the Euclidean distance between Y and Xb, namely:

- $\hat{\beta}^{OLS} := \arg\min_{b \in \mathbb{R}^p} \|Y - Xb\|_2^2$.
- when $\ker(X) = \{0\}$ (implying thus $p \leq n$) we have $\hat{\beta}^{OLS} = (X'X)^{-1}X'Y$.
- The OLS estimator is not defined when $p > n$.

Penalized estimator

Consider the following penalized estimator

$$\hat{\beta} := \arg\min_{b \in \mathbb{R}^p} \frac{1}{2}\|Y - Xb\|_2^2 + \lambda J(b), \text{ where } J \text{ is a norm.}$$

- $\hat{\beta}$ is well defined when $n \geq p$ as well as when $n < p$.
- $\hat{\beta}$ is characterized by its subdifferential ∂_J.
- The dual norm J^* is given by $J^*(x) = \sup\{z'x : J(z) \leq 1\}$.
- The null (convex) set is: $N = \{z \in \mathbb{R}^n : (X'z) \in B^*\}$
  ($\hat{\beta} = 0$ if and only if $Y \in N$).

Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO estimator (Chen & Donoho, 1994, Tibshirani, 1996) minimizes the ℓ^1-penalized Euclidean distance between Y and Xb, namely

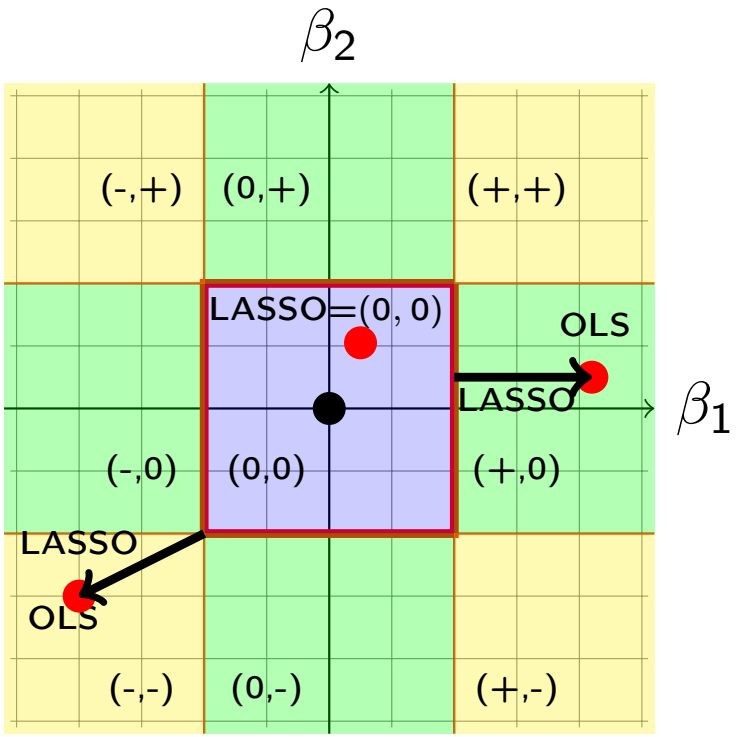
$$\hat{\beta}^{LASSO} := \arg\min_{b \in \mathbb{R}^p} \frac{1}{2} \|Y - Xb\|_2^2 \; + \lambda \|b\|_1, \quad \lambda > 0.$$

- $\hat{\beta}^{LASSO}$ is well defined when $n \geq p$ as well as when $n < p$.
- $\hat{\beta}^{LASSO}$ is characterized by

$$\begin{aligned} & 0 \in -X'(Y - X\hat{\beta}) + \lambda \partial_{\|\cdot\|_1}(\hat{\beta}) \\ \Leftrightarrow \quad & \|X'(Y - X\hat{\beta})\|_\infty \leq \lambda \text{ and } \hat{\beta}' X'(Y - X\hat{\beta}) = \lambda \|\hat{\beta}\|_1. \end{aligned}$$

Dual ball: $B^* = \mathrm{Conv}(W_\Sigma \lambda \mathbb{1})$, $\Sigma = A_1^p$.

LASSO vs. OLS



$\hat{\beta}^{LASSO}$ and $\hat{\beta}^{OLS}$ in orthogonal design: $X'X = \text{Id}$.

This picture explains that $\hat{\beta}^{LASSO}$ often has many $\hat{\beta}_i = 0$ (it is sparse).

Sign models of LASSO : $(0,0), (\pm,0), (0,\pm), (\pm,\pm)$. There as as many sign models as faces of $Conv(W_{A_1^p}[\lambda, \ldots, \lambda]')$.

Sorted ℓ^1 Penalized Estimator (SLOPE)

SLOPE estimator (Bogdan, van den Berg, Sabatti, Su, Candès, 2015) minimizes the sorted ℓ^1 penalized Euclidean distance between Y and Xb, namely:

$$\hat{\beta}^{SLOPE} := \arg\min_{b \in \mathbb{R}^p} \frac{1}{2}\|Y - Xb\|_2^2 + \gamma J_\Lambda(b).$$

- Sorted ℓ^1 norm: $J_\Lambda(b) := \sum_{i=1}^{p} \lambda_i |b|_{(i)}$, where $\lambda_1 > 0, \lambda_1 \geq \ldots, \lambda_p \geq 0$ and $|b|_{(1)} \geq \ldots \geq |b|_{(p)}$.
- $\hat{\beta}^{SLOPE}$ is well defined when $n \geq p$ as well as when $n < p$.
- $\hat{\beta}^{SLOPE}$ is characterized by

$$\begin{aligned} & 0 \in -X'(Y - X\hat{\beta}) + \lambda \partial J_\Lambda(\hat{\beta}) \\ \Leftrightarrow \quad & J_\Lambda^*(X'(Y - X\hat{\beta})) \leq 1 \text{ and } \hat{\beta}' X'(Y - X\hat{\beta}) = J_\Lambda(\hat{\beta}). \end{aligned}$$

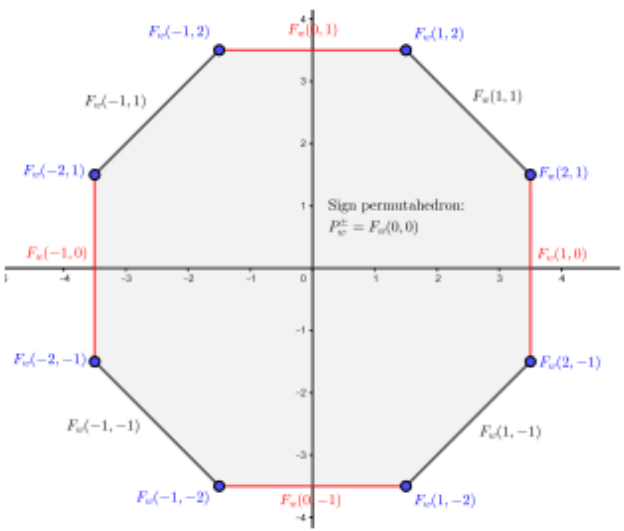
Dual ball: $B^* = \mathrm{Conv}(W_\Sigma \Lambda)$, $\Sigma = B_p$.

Signed permutahedron $P^{\pm}(\Lambda) = \mathrm{Conv}(W_{B_p}\Lambda)$

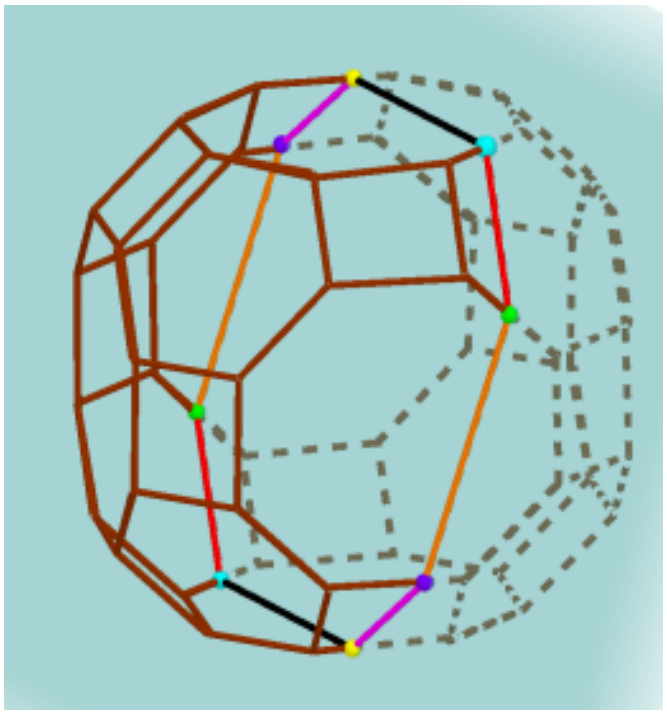
The dual sorted ℓ^1 norm is:

$$J_\Lambda^*(x) = \max\left\{\frac{|x|_{(1)}}{\lambda_1}, \frac{|x|_{(1)}+|x|_{(2)}}{\lambda_1+\lambda_2}, \ldots, \frac{|x|_{(1)}+\cdots+|x|_{(p)}}{\lambda_1+\cdots+\lambda_p}\right\}.$$

The unit ball of the J_Λ^* is the polytope $\mathrm{Conv}(W_{B_p}\Lambda)$.



$P^{\pm}(\Lambda)$ in $\mathbb{R}^2$



$P^{\pm}(\Lambda)$ in $\mathbb{R}^3$

SLOPE vs. OLS

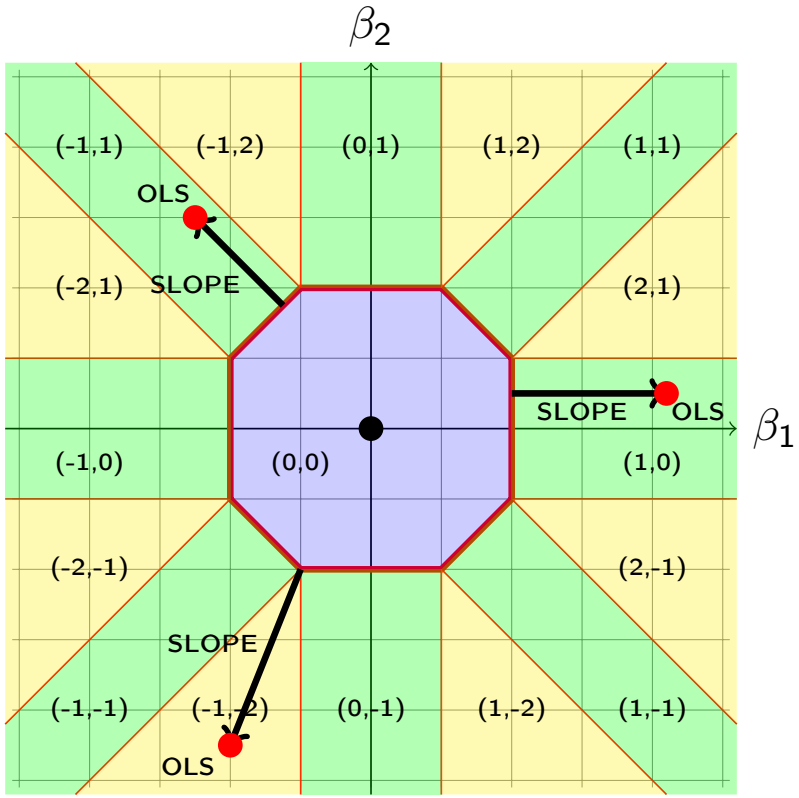


Figure: $\hat{\beta}^{SLOPE}$ and $\hat{\beta}^{OLS}$ in orthogonal design: $X'X = \mathrm{Id}$.

Models for SLOPE : $(0,0), \pm(1,0), \pm(0,1), \pm(1,1),$
$\pm(1,-1), \pm(2,1), \pm(2,-1), \pm(1,2), \pm(1,-2)$.

SLOPE model recovery

Recall: $Y = X\beta + \varepsilon$ and

$$\hat{\beta}^{SLOPE} := \arg\min_{b \in \mathbb{R}^p} \frac{1}{2}\|Y - Xb\|_2^2 + \gamma J_\Lambda(b);\ \gamma > 0.$$

There is a bijection between SLOPE models and faces of $P^{\pm}(\Lambda)$, thus $\#\{\text{SLOPE models in } \mathbb{R}^p\} = \#\{\text{faces } (P^{\pm}(\Lambda))\}$.

Theorem (Model recovery by SLOPE in the noiseless case ($\varepsilon = 0$) (TS, Bogdan, Graczyk, Kołodziejek, Tardivel, Wilczyński (2021+)))

For every $\Lambda = (\lambda_1 > \lambda_2 > \ldots \lambda_p > 0)$ there exists $\gamma > 0$ such that $\mathrm{mdl}(\hat{\beta}^{SLOPE}) = \mathrm{mdl}(\beta)$ if and only if :

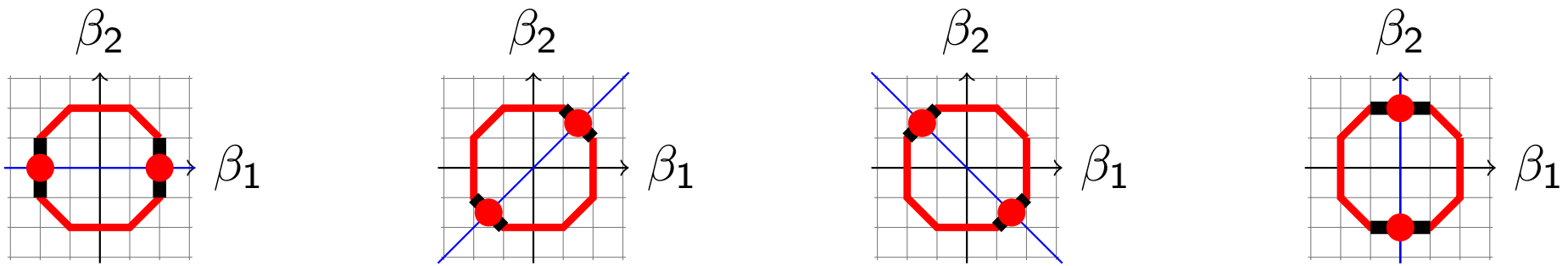
$$X'XV_{\mathrm{mdl}(\beta)} \cap F_{\mathrm{mdl}(\beta)} \neq \emptyset,$$

where

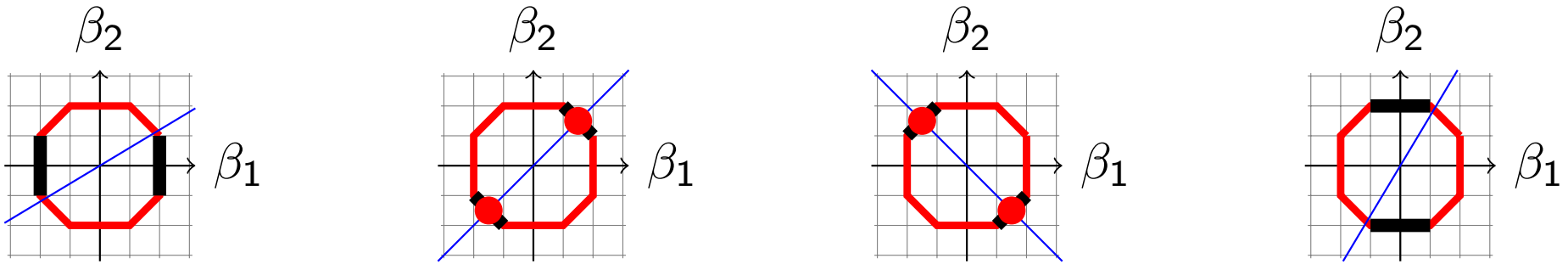
- $F_{\mathrm{mdl}(\beta)} := \partial J_\Lambda(\beta) = \partial J_\Lambda(\mathrm{mdl}(\beta))$,
- $V_{\mathrm{mdl}(\beta)} = \mathrm{lin}\{x \in \mathbb{R}^p : \mathrm{mdl}(x) = \mathrm{mdl}(\beta)\}$.

Model Recovery: $\mathrm{mdl}(\hat{\beta}^{SLOPE}) = \mathrm{mdl}(\beta)$

V_M vs. F_M (orthogonal case):



$X'XV_M$ vs. F_M (non-orthogonal case, $X'X = \begin{bmatrix} 1 & 3/5 \\ 3/5 & 1 \end{bmatrix}$):



References

M. Bogdan and E. van den Berg and C. Sabatti, W. Su and E. J. Candès, SLOPE – Adaptive Variable Selection Via Convex Optimization, Annals of Applied Statistics, vol.9, pp. 1103-1140, 2015.

D. Brzyski, Selecting relevant groups of explanatory variables via convex optimization methods with the false discovery rate control, PhD Thesis 2015.

S. Chen, D. Donoho. Basis pursuit. Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers, 1994, pp. 41-44 vol.1.

K. Ewald, U. Schneider. Uniformly Valid Confidence Sets Based on the Lasso. EJS, vol. 12, pp. 1358-1387, 2018.

C. F. Gauss. Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium, 1809.

A.-M. Legendre. Nouvelles méthodes pour la détermination des orbites des comètes, Paris, Firmin Didot, 1805.

U. Schneider and P. Tardivel. The geometry of uniqueness, sparsity and clustering in penalized estimation. Arxiv 2004.09106, submitted 2020.

T. Skalski, M. Bogdan, P. Graczyk, B. Kołodziejek, P. Tardivel, M. Wilczyński. Model Recovery by SLOPE. In preparation.

P. Tardivel, T. Skalski, P. Graczyk, U. Schneider. The Geometry of Model Recovery by Penalized and Thresholded Estimators. 2021. hal-03262087.

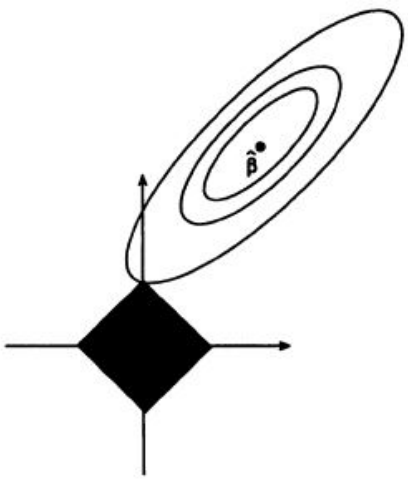
R. Tibshirani. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1, [Royal Statistical Society, Wiley], 1996, pp. 267–88.

R. J. Tibshirani, J. Taylor. The solution path of the generalized lasso. Annals of Statistics, vol. 39, no. 3, pp. 1335-1371, 2011.

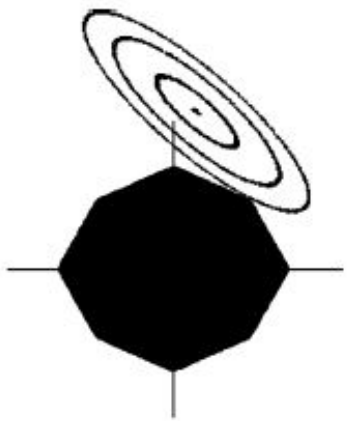
Thank you for your attention!

Appendix: Pictures from the Title Page

Meeting point of scaled B and scaled unit ball in ℓ^2 of $(Y - Xb)$ is equal to $\hat{\beta}$.



$$\mathrm{sign}(\hat{\beta}^{LASSO}) = (0, +)$$



$$\mathrm{mdl}(\hat{\beta}^{SLOPE}) = (1, 1)$$

Appendix: Subdifferential

Definition (Subgradient)

Let $f : \mathbb{R}^p \mapsto \mathbb{R}$. Then g is a subgradient of f at b if

$$\forall h \in \mathbb{R}^p \quad f(b+h) \geq f(b) + g'h.$$

Definition (Subdifferential)

The subdifferential $\partial f(b)$ of f at b is the set of all subgradients of f at b.

Appendix: Thresholding estimator

Definition (Thresholded penalized least squares estimator)

Let pen be a penalizer, $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$ and $\lambda > 0$. Given $\hat{\beta} \in S_{X,\lambda\mathrm{pen}}(y)$, we say that $\hat{u}$ is a thresholded estimator of $\hat{\beta}$ if $\partial_{\mathrm{pen}}(\hat{\beta}) \subset \partial_{\mathrm{pen}}(\hat{u})$.

Definition (Thresholded LASSO)

$$\hat{\beta}_i^{\mathrm{LASSO},\tau} = \begin{cases} \hat{\beta}_i^{\mathrm{LASSO}}, & \text{if } |\hat{\beta}_i^{\mathrm{LASSO}}| > \tau, \\ 0, & \text{otherwise.} \end{cases}$$

Appendix: LASSO in orthogonal case

Theorem (Tibshirani, 1996)

Exact formula in orthogonal ($X'X = I$) *case:*

$$\hat{\beta}_i^{LASSO} = \text{sign}(\hat{\beta}_i^{OLS}) \max\{|\hat{\beta}_i^{OLS}| - \lambda, 0\}.$$

Thank you for your attention!