The Dantzig Selector

presented by
Micheal Hutchinson, Brian Liu, Anna Menacher and Daniel Moss

March 20, 2020

The Dantzig Selector

- Introduced by Candes, Tao 2007 (Annals).
- Define a *Dantzig selector* to be a solution to the constrained optimization problem

$$\min_{\tilde{\beta} \in \mathbb{R}^p} \|\tilde{\beta}\|_1 \quad \text{subject to} \quad \|X^T r\|_\infty \leq \lambda_p \cdot \sigma \tag{1}$$

  where r is the vector of *residuals*, defined for each $\tilde{\beta}$ by $r = Y - X\tilde{\beta}$ and where $\lambda_p = (1 + t^{-1})\sqrt{2 \log p}$ (for the DS).
- Main result is a high-probability oracle inequality (non-asymptotic)

The Uniform Uncertainty Principle (UUP)

- This is a condition that can be placed a non squared matrix, and in a sense requires that the matrix is close enough to being (having orthogonal columns?) orthogonal.
- This is a class of constraint, not a single one.
- The one we are interested in here is of the form:

$$\delta_{2s} + \theta_{s,2s} < 1$$

where s is the sparsity level and $\delta_{(\cdot)}$, $\theta_{(\cdot,\cdot)}$ are to be defined.

δ_s, s-restricted isometric constant

- Let us consider a matrix $X \in \mathbb{R}^{n \times p}$
- Consider the matrix X_T to be the matrix formed by taking the columns in the set $T \subset \{1, ..., p\}$ and forming a matrix from them.
- δ_s is the smallest constant, such that for all T, $|T| < s$ and for all $c_T \in \mathbb{R}^{|T|}$,

$$(1-\delta_s)||c_T||_{l_2}^2 \leq ||X_T c_T||_{l_2}^2 \leq (1+\delta_s)||c_T||_{l_2}^2$$

- For an orthogonal matrix this would be 0.
- $[(1-\delta_s), (1+\delta_s)]$ gives upper and lower bound on singular values of X, i.e. the smallest angle between any two columns in X.

$\theta_{s,s'}$, s, s'-restricted orthogonality constants

- Consider the same matrix as before.
- Consider now 2 disjoint column selections from X, $X_T, X_{T'}$
- $\theta_{s,s'}$ is the smallest constant such that, for all T, T', $|T| < s$, $|T'| < s'$, and for all $c_T \in \mathbb{R}^{|T|}$, $c_{T'} \in \mathbb{R}^{|T'|}$,

$$|\langle X_T c_T, X_{T'} c_{T'} \rangle| \leq \theta_{s,s'} ||c_T||_{l_2}^2 ||c_{T'}||_{l_2}^2$$

- Columns are scaled to have unit norm so this condition corresponds to closeness to orthogonality of columns

The Uniform Uncertainty Principle

- So our constraint of

  $$\delta_{2s} + \theta_{s,2s} < 1$$

  is one possible expression of closeness to orthogonality.
- Note you can swap out δ and θ terms using a series of inequalities between the two. Also, $3s \leq p$.
- In the noiseless case, obeying this would allow for perfect reconstruction of the sparsity
- In the noisy case, this allows for a strong oracle inequality, which is the main theoretical attraction of the Dantzig selector.

Oracle inequality

Let $\beta \in \mathbb{R}^p$ be supported on S and $s = |S|$. Let $\delta_s + \theta_{s,2s} < 1$. Choose $t > 0$ such that $\delta_s + \theta_{s,2s} < 1 - t$. Then, with high probability (see report), the Dantzig selector satisfies

$$\|\hat{\beta} - \beta\|_2^2 \leq C_2^2 \lambda_p^2 \left(\sigma^2 + \sum_{i=1}^{p} \min(\beta_i^2, \sigma^2) \right)$$

where $\lambda_p = (1 + t^{-1})\sqrt{2 \log p}$ and where the exact constant C_2 depends only on δ_s and $\theta_{s,2s}$.

Note: The sum term is the optimal expected risk achieved by an oracle who knows which elements of β are below the noise threshold. It signifies an optimal bias-variance tradeoff.

LASSO

+ high predictive accuracy
+ equivalent solution of Dantzig selector and LASSO for λ
  $\rightarrow$ transfer theoretical properties
− theoretical justification not as extensive as the Dantzig selector

Robert Tibshirani (1996). *"Regression Shrinkage and Selection via the Lasso"*

The Dantzig Selector

- $+$ theoretical properties
  $\rightarrow$ loss within a logarithmic factor of the ideal MSE ($\sqrt{2\log(p)}\ \sigma$)
- $+$ reliable & accurate parameter estimation for $p \gg n$
- $-$ computational complexity of calculating entire coefficient path over a fine grid of λ
- $-$ lower prediction accuracy than for the LASSO for $p \gg n$ and a sparse parameter vector

Emmanuel Candes and Terence Tao (2007).*"The Dantzig selector: Statistical estimation when p is much larger than n"*

DASSO

+ computational efficiency of LARS algorithm
+ theoretical properties of Dantzig selector
− no computational implementation of the algorithm
− floating point numerical errors

Gareth M. James, Peter Radchenko and Jinchi Lv (2009). *"DASSO: Connections Between the Dantzig Selector and Lasso"*

Other Solutions

- Gauss-Dantzig Selector

  → higher statistical accuracy
- Adaptive (Doubly Weighted) Dantzig Selector

  → more computationally efficient

  → lessens overshrinkage of parameter estimates

Emmanuel Candes and Terence Tao (2007). *"The Dantzig selector: Statistical estimation when p is much larger than n"*
Gareth M. James and Peter Radchenko (2009). *"A generalized Dantzig selector with shrinkage tuning"*

Implementing DASSO

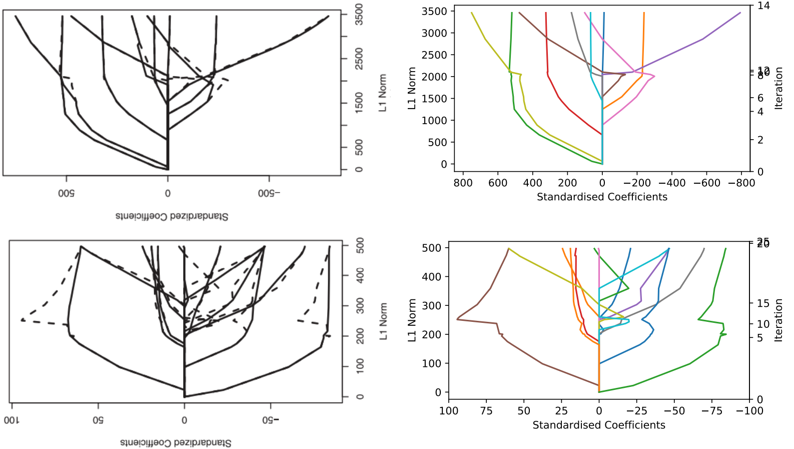

Figure: Coefficient paths generated by DASSO on (top) diabetes data and (bottom) Boston housing data. The paths for the datasets in our implementation (right) are comparable to that of the original paper (left, dashed lines).

Implementing DASSO

Challenge 1: Does `2.35561920212286` = `2.35561920212304` ?

Two values that are equal in theory can appear to be different in practice due to floating-point numerical errors.

This confuses our DASSO implementation, which features many equality/inequality tests.

We say two values are equal if they are less than ε apart:

- $\forall x \;\; x \neq 0$ if $|x| > \varepsilon$
- $\forall c \;\; c$ is maximal if $c > \max_c(c) - \varepsilon$

David Bindel and Jonathan Goodman (2009). *Principles of scientific computing.*

Implementing DASSO

Challenge 2: What is $\min(\varnothing)$ **?**

A step in the DASSO required calculating:

$$\gamma_1 = min^{+}_{j \in \mathcal{A}^c}(f(j), g(j)) \tag{2}$$

What if $\mathcal{A}^c = \varnothing$? James et al. (2009) forgot to define the quantity under this edge case.

Natural candidate: $\gamma_1 = +\infty \quad \rightarrow \quad$ infinite loop.
We have to refer to another paper to set a good value for this case.

Gareth M. James, Peter Radchenko and Jinchi Lv (2009). *"DASSO: Connections Between the Dantzig Selector and Lasso"*

References I

[1] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when *p* is much larger than *n*. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[2] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–451, 2004.

[3] Bradley Efron, Trevor Hastie, and Robert Tibshirani. Discussion: The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2358–2364, 2007.

[4] Gareth James, Peter Radchenko, and Jinchi Lv. Dasso: Connections between the dantzig selector and lasso. *Journal of the Royal Statistical Society Series B*, 71:127–142, 2009.

[5] Gareth M. James and Peter Radchenko. A generalized Dantzig selector with shrinkage tuning. *Biometrika*, 96(2):323–337, 2009.

References II

[6] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.