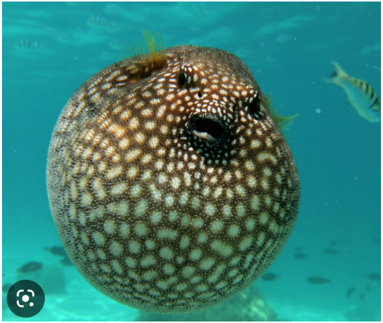
Preconditioning the Lasso for Sign Consistency: An overview

Yujing Lu and Tabitha Peter

The University of Iowa, College of Public Health

May 08, 2023

A problem, a paper, and a Puffer fish



Puffer fish

What makes a matrix ill-conditioned?

Consider systems

$$\begin{cases} x + y = 2 \\ x + 1.001y = 2 \end{cases} \quad \text{and} \quad \begin{cases} x + y = 2 \\ x + 1.001y = 2.001 \end{cases}$$

The system on the left has solution $x = 2, y = 0$ while the one on the right has solution $x = 1, y = 1$. The coefficient matrix is called *ill-conditioned* because a small change in the constant coefficients results in a large change in the solution. A *condition number*, defined in more advanced courses, is used to measure the degree of ill-conditioning of a matrix (≈ 4004 for the above).

Generalized least squares is not always a good preconditioner

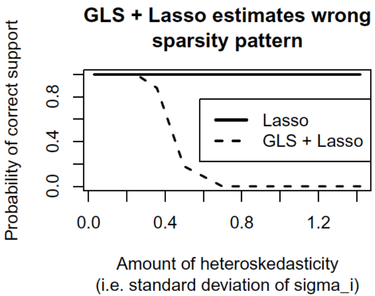


FIG 1. *GLS acts as a bad preconditioner, making the design matrix ill-conditioned. Thus, correcting for the heteroskedasticity degrades the estimation performance. In this simulation,* $n = 200$, $p = 1000$ *and there are 10 nonzero elements in* β^*. *Appendix D contains further details on this simulation.*

Sign consistency and the irrepresentable condition

Definition of sign consistency

The Lasso is sign consistent if there exists a sequence λ_n such that,

$$\mathbb{P}(sign(\hat{\beta}(\lambda_n)) = sign(\beta^*)) \to 1, \text{ as } n \to \infty.$$

The irrepresentable condition

The design matrix $\mathbf{X}$ satisfies the irrepresentable condition for β^* if, for some constant $\eta \in (0, 1]$,

$$\|\mathbf{X}_{S^c}^\top \mathbf{X}_S (\mathbf{X}_S^\top \mathbf{X}_S)^{-1} sign(\beta_S^*)\|_\infty \leq 1 - \eta,$$

where $S = \{j : \beta_j^* \neq 0\} \subset \{1, ..., p\}$

Finding a better preconditioner connects back to the irrepresentable condition

- Many methods have been proposed to circumvent the irrepresentable condition - concave penalty, adaptive lasso, etc.
- Preconditioning attempts to solve the problem from a different angle: altering the shape of $\|\mathbf{Y} - \mathbf{X}\beta\|^2$

Definition of the Puffer transformation

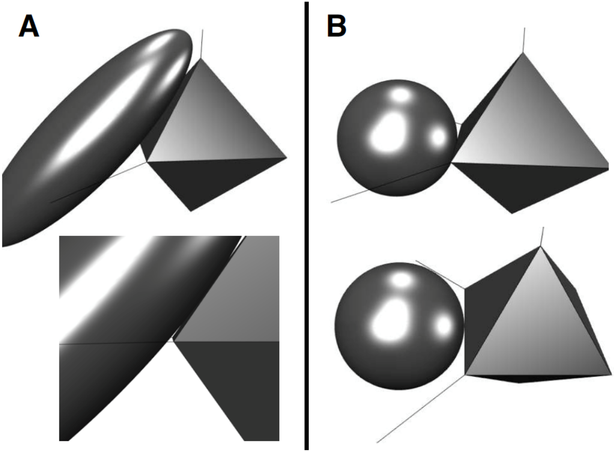
Suppose $\mathbf{X} \in \mathbb{R}^{n \times p}$ has rank $d = \min\{n, p\}$, then from SVD, we have $\mathbf{U} \in \mathbb{R}^{n \times d}$, $\mathbf{V} \in \mathbb{R}^{p \times d}$, and diagonal matrix $\mathbf{D} \in \mathbb{R}^{d \times d}$, then the Puffer transformation is $\mathbf{F} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^\top$

The Puffer transformation

- $\mathbf{FX} = \mathbf{UV}^\top$ - singular values of $\mathbf{FX}$ are all 1, which leads to orthonormality
- $\mathbf{FY} = (\mathbf{FX})\beta^* + \mathbf{F}\epsilon$, where $\mathbf{F}\epsilon \sim N(\mathbf{0}, \tilde{\boldsymbol{\Sigma}} = \sigma^2\mathbf{UD}^{-2}\mathbf{U}^\top)$
- There are issues when any singular values of $\mathbf{X}$ approach 0, a modified preconditioner will be introduced later

Geometrical representation

- $\hat{\beta}(c) = \arg\min_{\beta:\|\beta\|_1 \leq c} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2$



Low dimension results

- If $n \geq p$ and $\mathbf{X}$ is full rank, then $(\mathbf{FX})^\top \mathbf{FX} = \mathbf{I}$

Theorem of sign consistency after the Puffer transformation

Suppose that $\mathbf{Y} = \mathbf{X}\beta^* + \epsilon$ with $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Suppose that $n \geq p$ and $\mathbf{X}$ has rank p. Further assume that $\Lambda_{min}(\frac{1}{n}\mathbf{X}^\top\mathbf{X}) \geq \tilde{C}_{min} > 0$. Let $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{X}$, $\tilde{\mathbf{Y}} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{Y}$, and $\tilde{\mathbf{\Sigma}} = \sigma^2\mathbf{U}\mathbf{D}^{-2}\mathbf{U}^\top$).
If $\min_{j\in S}|\beta_j^*| > 2\lambda$, then $\tilde{\beta}(\lambda) =_s \beta^*$ with probability greater than

$$1 - 2p\exp\left\{-\frac{n\lambda^2\tilde{C}_{min}}{2\sigma^2}\right\}$$

Low dimension remarks

- Suppose that $\tilde{C}_{min} > 0$ is a constant. If $p, \min_{j \in S} |\beta_j^*|$ and σ^2 so not change with n, then choosing λ such that $\lambda \to 0$ and $\lambda^2 n \to \infty$ ensures that $\tilde{\beta}(\lambda)$ is sign consistent. One possible choice is $\lambda = \sqrt{\frac{\log n}{n}}$
- If pen_j's are identical functions that have a cusp at zero, then the solution selects the same sequence of models as preconditioned correlation screening: $\hat{\beta}_j \neq 0$, if $|cor(\mathbf{FY}, \mathbf{FX}_j)| > \lambda$
- In high-dimensional scenario, **FX** is no longer orthogonal

Generalized Puffer transformation uses a tuning parameter to hem in singular values

- $\tilde{\boldsymbol{\Sigma}} = \sigma^2 \mathbf{U}\mathbf{D}^{-2}\mathbf{U}^\top$

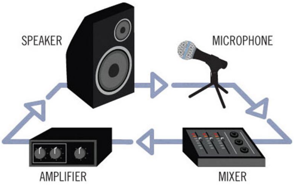
Definition of the Generalized Puffer transformation

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a design matrix with SVD $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. Define $g : \mathbb{R}^2 \to \mathbb{R}$, $\tau \in \mathbb{R}$, and $\hat{D}_{ii} = \frac{g(D_{ii},\tau)}{D_{ii}}$,

$$\mathbf{F}_{g,\tau} = \mathbf{U}\hat{\mathbf{D}}\mathbf{U}^\top$$

- **Note**: when g is the hard thresholding function $h(x,\tau) = \mathbb{1}(x \geq \tau)$, then the spectral norm of $\mathbf{F}_{h,\tau}$ is bounded by $\frac{1}{\tau}$

Dealing with the irrepresentable condition is like dealing with mic feedback



High dimension remarks

- There is a tension between 1) satisfying the irrepresentable condition and 2) limiting the amount of additional noise created by the preconditioner
- The generalized Puffer transformation can handle high degrees of correlation among features
- TL;DR of the main result for the generalized case: we can make the lower bound on the probability $\mathbb{P}(\tilde{\boldsymbol{\beta}}(\lambda) =_s \boldsymbol{\beta}^*)$ converge to 1 by choosing the tuning parameter so that $\lambda^2\tau_n^2$ grows faster than $\log(p)$

Simulations

- The rows of $\mathbf{X}$ are mean zero Gaussian vectors with constant correlation ρ
  - The Puffer preconditioned Lasso simultaneously achieves fewer false positives, fewer false negatives, and smaller MSE in β across all values of ρ
- $X_{ij} = (G_i/\alpha)Z_{ij}$, where Z_{ij} are iid standard normal, and G_i are independent Gamma r.v. with shape α and rate 1
  - As $\alpha \to 0$, the standard deviation of G_i/α grows
  - The generalized Puffer transformed Lasso yields a better sign estimator than both the Lasso and the Puffer preconditioned Lasso
- Other types of preconditioner

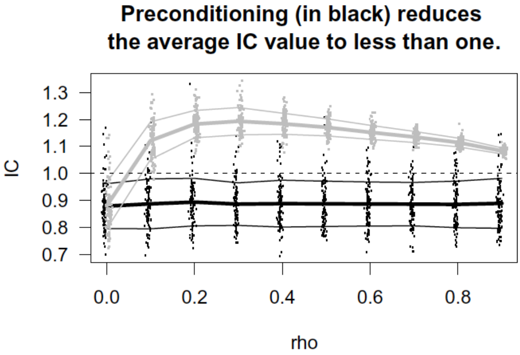
Wrap up: discussion and take-aways

- Preconditioning can circumvent the irrepresentable condition and achieve sign consistency
- In low dimensions, the Puffer transformation ensures the irrepresentable condition; In high dimensions, the generalized Puffer transformation satisfies the irrepresentable condition with a high probability
- Puffer fish video link

References

- Jia, J. and Rohe, K. "Preconditioning the Lasso for sign consistency" (2015). Electronic Journal of Statistics.
- Khan, Emtiyaz. "Ill-Conditioned Matrices" lecture from *Pattern Classification and Machine Learning* course. Accessed online May 05 2023. https://emtiyaz.github.io/pcml15/illconditioned.pdf
- Lall, Sanjay. "SVD and applications" lecture from *Introduction to Linear Dynamical Systems* course. Accessed online 05 May 2023. https://ee263.stanford.edu/lectures/svd.pdf

High dim. simulation shows Puffer can powerfully reduce correlation between features



where IC is the expression previously defined for the irrepresentable condition. $IC_{\beta^*}(\mathbf{X}) < 1 \rightarrow \mathbf{X}$ satisfies the irrepresentable condition