# Combining $l^1$ Penalization with Higher Moment Constraints in Regression Models

Austin David Brown

December 6, 2018

# Motivation

- In statistics and probability theory it is common to impose moment assumptions on a random variable $X : \Omega \to \mathbb{R}^n$ such as $E(\|X\|^k) < \infty$ for $k \in \mathbb{R}$.

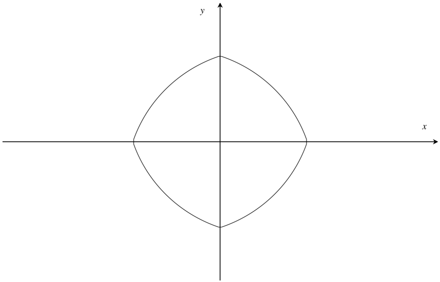
- These constraints correspond to the $L^p$ spaces which allow control over the width and the height of such random variables. This can be interpreted as imposing "stability" on the random variable $X$.

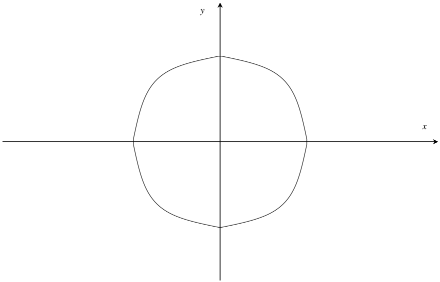
- If statisticians so freely impose such constraints then we should build a tool to allow scientists and researchers to impose such constraints on their real problems.

- In this project, we build a package implements this idea. The goal is to build upon the idea of ElasticNet [7] and create a tool available to scientists and researchers.

# Geometric Motivation

Consider for example an Elasticnet [7] penalty $Q(x) = \frac{1}{2}|x| + \frac{1}{2}|y| + \frac{1}{2}|x|^2 + \frac{1}{2}|y|^2 \leq 1$ shown on the left. Extending this idea, a new penalty $P(x) = \frac{1}{2}|x| + \frac{1}{2}|y| + \frac{1}{2}|x|^4 + \frac{1}{2}|y|^4 \leq 1$ shown on the right.




It seems reasonable that a scientist or researcher may want the option to "bow" out the feasible set.

# The Setup

Define a new penalty to try to impose more "stability" for scientists and researchers (try to extend the Elasticnet idea). Let

$$L_\lambda(\beta) = \frac{1}{2} \|y - X\beta\|_2^2 + \lambda P(\beta)$$

with penalty

$$\lambda P(\beta) = \lambda\alpha_0 \|\beta\|_1 + \lambda \sum_{k=1}^{5} \alpha_k \|\beta\|_{2k}^{2k}$$

where $y \in \mathbb{R}^n$, $X \in M_{n\times p}(\mathbb{R})$, and $\beta \in \mathbb{R}^p$, $\lambda \in \mathbb{R}_+$ and $\alpha$'s are convex combinations or use separate tuning parameters instead.

This is a convex, separable penalty, but the derivative is not Lipschitz.

# Algorithm Implementation 1

**Algorithm 1:** Subgradient Coordinate Method

Choose $\beta^0 \in \mathbb{R}^p$ and tolerance $\delta > 0$;
Set $k \leftarrow 0$
**repeat**
  Set the step size $h^k \leftarrow \frac{R}{\sqrt{1+k}}$ for some $R > 0$ or use a constant step size.
  Permute $I = \{1, \ldots, p\}$
  **for** $i \in I$ **do**
    $\beta_i^{k+1} \leftarrow \beta_i^k - h^i g^i$ where $g^i \in (\partial L)_i$
  **end**
  $k \leftarrow k + 1$
**until** *Until the loss difference* $\Delta L$ *is less than* $\delta$;

# Drawbacks

- Not a descent method.
- No good stopping criterion.
- Convergence theory is worse.
- Tends to produce really small values instead of truly sparse solutions.

# A Much Better Algorithm

**Algorithm 2:** Proximal Gradient Coordinate Descent

Choose $\beta^0 \in \mathbb{R}^p$ and tolerance $\delta > 0$;
Set $k \leftarrow 0$
**repeat**
    Set the step size $h^k > 0$ with diminishing or line search.
    Randomly permute $I = \{1, \ldots, p\}$
    **for** $i \in I$ **do**
        $\beta_i^{k+1} \leftarrow (\mathbf{prox}_{h^k L})_i(\beta_i^k - h^k \langle X_i, y - X\beta \rangle)$
    **end**
    $k \leftarrow k + 1$
**until** *Until the Moreau-Yoshida mapping* $M_{h_k, f} < \delta$;

# Benefits

- A descent method, so can do line search.
- Good stopping criterion at $O(p^2)$ flops.
- Convergence theory is unchanged from differentiable functions.
- Coordinate descent theory is possibly developed here by Nesterov (2012). The iterations are cheap so 1/k convergence is actually really good.

# Cross-Validation Implementation

**Algorithm 3:** Warm Start Cross-Validation

Choose a sequence of Langrangian dual variables $\lambda_1, \ldots, \lambda_N$, and initial value $\beta^0$.

Order $\lambda_{(1)}, \ldots, \lambda_{(N)}$ descending.

$\beta^{Warm} \leftarrow \beta^0$

**for** $k \in 1, \ldots, N$ **do**

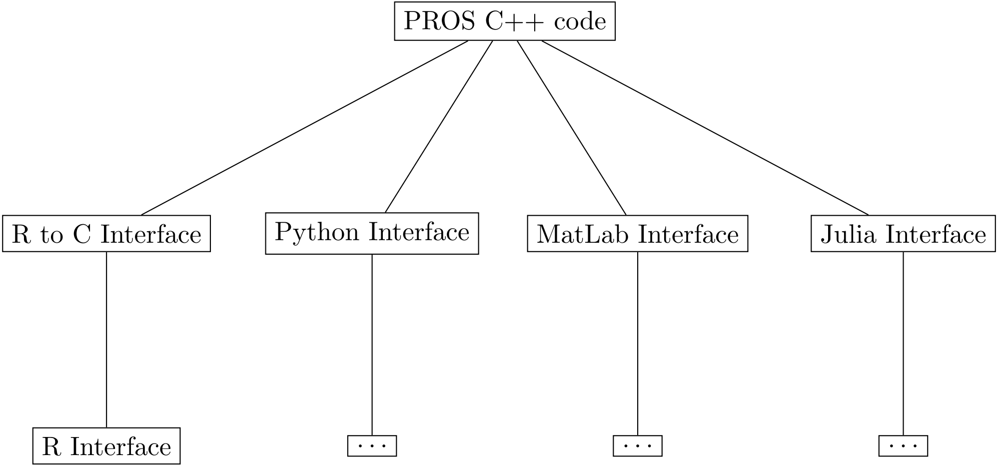
    $\beta^k \leftarrow$ Using Cross-Validation with $\lambda_{(k)}$ warm started with $\beta^{Warm}$.

    $\beta^{Warm} \leftarrow \beta^k$

**end**

# Package Architecture

Written entirely in C++ for speed using Eigen [5] and can be interfaced to many other popular languages.



# Penalized Regression on Steroids

Installation:

```
> devtools::install_github("austindavidbrown/pros/R-package")
```

A single fit function with prediction

```
> fit <- pros(X, y, alpha, lambda)
> predict(fit, new_X)
```

A cross-validation fit function with prediction

```
> cv <- cv.pros(X, y, alpha)
> predict(cv, new_X)
```

There is even a reference manual.

# The Prostate Cancer Dataset Analysis

This is a popular dataset from Stamey et al. [1] and analyzed in the ElasticNet paper Zou and Hastie [7]. There are 8 predictors and 1 response is the log of the prostate specific antigen.

- The data is split into a training set with 67 observations and a test set with 30 observations.
- The predictor data was standardized.
- 10-fold cross-validation and manual tuning due to time constraint of the project were used.
- The **glmnet** [6] library was used to fit and tune the Lasso and a naive un-tuned ElasticNet.
- **pros** [2] was used to fit the new penalty.
- This was not an extensive study due to time and the code is available to you at the **pros** repository [2]. I also don't really know the best way to tune my own method :(.

# The Prostate Cancer Dataset Results

Random seed = 8989. Varying results for glmnet: sometimes lower sometimes higher.

| **Penalty** | **Tuning** | **Test MSE** |
|---|---|---|
| Lasso Penalty (glmnet) | $\alpha = (1, 0)$ | 0.4558642 |
| ElasticNet Penalty (glmnet) | $\alpha = (1/2, 1/2)$ | 0.4554527 |
| New Penalty (pros) | $\alpha = (1/5, 0, 0, 0, 0, 4/5), \lambda = 54.02$ | 0.4442196 |
| New Penalty (pros) | $\alpha = (1/5, 0, 1/5, 1/5, 1/5, 1/5), \lambda = 54$ | 0.4474082 |
| New Penalty (pros) | $\alpha = (1/2, 0, 0, 0, 0, 1/2), \lambda = 21.05$ | 0.4441578 |

# Conclusion

- I think that there are many "good" solutions to real problems.
- We can get "statistics" on these solutions by imposing more norms on the feasible sets building upon the Elasticnet [7] idea.
- I would like to explore even more sparsity inducing penalizations such as

$$\lambda P'(\beta) = \lambda\alpha_1 \|\beta\|_1 + \lambda\alpha_2 \|\beta\|_\infty$$

$$\lambda P''(\beta) = \lambda\alpha_1 \|\beta\|_1 + \lambda \sum_{k=2}^{10} \alpha_k \|\beta\|_k$$

# Things to Address

- Currently, you have to tune a step size. I need to implement a fast line search. This is not too difficult actually just ran out of time.
- C++ offers many random number generators. I need to make sure I am handling this correctly and interopting with R properly. This is subtle and tricky. Also am I using a good one? I don't know.
- overflow fixes
- passing data between R and C in the fastest way.
- Look into convergence analysis.

# References


[1] Stamey, T.A., Kabalin, J.N., McNeal, J.E., Johnstone, I.M., Freiha, F., Redwine, E.A. and Yang, N. (1989) Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. radical prostatectomy treated patients, Journal of Urology 141(5), 1076–1083.

[2] PROS. github.com/austindavidbrown/pros

[3] Neal Parikh and Stephen Boyd. 2014. Proximal Algorithms. Found. Trends Optim. 1, 3 (January 2014), 127-239. DOI=10.1561/2400000003 http://dx.doi.org/10.1561/2400000003

[4] Stephen J. Wright. 2015. Coordinate descent algorithms. Math. Program. 151, 1 (June 2015), 3-34. DOI=10.1007/s10107-015-0892-3 http://dx.doi.org/10.1007/s10107-015-0892-3

[5] Guennebaud, Gaël (2013). Eigen: A C++ linear algebra library (PDF). Eurographics/CGLibs.

[6] Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.

[7] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67, 301–320.

[8] Yurii Nesterov. 2014. Introductory Lectures on Convex Optimization: A Basic Course (1 ed.). Springer Publishing Company, Incorporated.