Penalized robust regression in high-dimension

Derek Bean, Peter Bickel[*], Noureddine El Karoui[†], Chinghway Lim and Bin Yu[‡]

October 14th, 2011

[*] Support from NSF grant DMS-0907362 is gratefully acknowledged.

[†] Support from an Alfred P. Sloan Research Fellowship and NSF grant DMS-0847647 (CAREER) is gratefully acknowledged

[‡] Support from NSF Grant DMS-0906808 is gratefully acknowledged. **Key words and Phrases :** Concentration of measure, robust regression, high-dimensional statistics,

Penalized robust regression in high-dimension

Abstract

We discuss the behavior of penalized robust regression estimators in high-dimension and compare our theoretical predictions to simulations. Our results show the importance of the geometry of the dataset and shed light on the theoretical behavior of LASSO and much more involved methods.

1 Introduction

We consider the problem of understanding the properties of

$$\widehat{\beta} = \text{argmin}_{\beta\in\mathbb{R}^p} \sum_{i=1}^n \rho(y_i - X_i'\beta) + \tau P(\beta) \ ,$$
$$y_i = X_i'\beta_0 + \epsilon_i \ ,$$

where X_i is a p-dimensional vector of observed predictors, ϵ_i is a vector of (unobserved) errors which we assume are independent of $\{X_i\}_{i=1}^n$, y_i is a 1-dimensional response, ρ is a given and known function. P is a penalty function. We are concerned with the high-dimensional case where p/n has a finite non-zero limit. $P(\beta)$ is a penalty function and β_0 is an unknown vector which we are trying to estimate.

The problem we are considering is very general, as it includes generic robust regression (Huber (1973)) as a subcase (case where $\tau = 0$), as well as methods which have recently received much attention such as LASSO (Tibshirani (1995)). In that latter case, $\rho(x) = x^2/2$ and $P(\beta) = \|\beta\|_1$.

There is currently much interest in statistics in the high-dimensional case, where p/n (the ratio of number of predictors over number of observations) is not close to zero. In that setting, it is interesting but non trivial to try to understand the behavior of $\widehat{\beta} - \beta_0$, as well as for instance that of the residuals. In particular, natural questions concern the impact of the loss function, ρ, on the results, the interaction of the distribution of ϵ_i's and ρ, as well as the role of the penalty function P and τ.

We are particularly interested in understanding the expected squared prediction error (conditional on the data we have seen). Hence, for new observations (X_{new}, y_{new}) (independent of the ones we have used to construct the estimator) we would like to understand

$$EPE = \mathbf{E}\left((y_{new} - X_{new}'\widehat{\beta})^2 | (y_1, X_1), \ldots, (y_n, X_n)\right) \ .$$

Clearly, if y_{new} follows our model, we have

$$EPE = \text{var}(\epsilon_{new}) + (\widehat{\beta} - \beta_0)'\text{cov}(X_{new})(\widehat{\beta} - \beta_0) \ .$$

In particular, when $\text{cov}(X_{new}) = \text{Id}_p$, this amounts to understanding $\|\widehat{\beta} - \beta_0\|$.

In this note, we propose an advanced heuristic (grounded in rigorous theory) to answer these questions. In particular, under assumptions that will be detailed below we are able to:

- find a system of 3 non-linear equations for 3 parameters to characterize $\|\widehat{\beta} - \beta_0\|$
- characterize the behavior of the residuals.
- understand the impact of the loss function, the distribution of the errors and the penalty
- make clear that low-dimensional intuition is upended in high-dimension. For instance, if the distribution of the errors is known, the natural loss function to use in low-dimension is related to the log of the error density. In high-dimension, it is sometime the case that another loss performs better. We illustrate this with double exponential errors and median vs least-squares regression penalized in both cases with an ℓ_2 penalty.

2 Characterization of the solution

2.1 Assumptions

For the purpose of this paper, we make the following assumptions.

- X_i are i.i.d can be written $X_i = \lambda_i \mathcal{X}_i$, where λ_i is a random variable independent of $\mathcal{X}_i$ and $\mathcal{X}_i$ is

Preliminary work. Under review by AISTATS 2012. Do not distribute.

$\mathcal{N}(0, \mathrm{Id}_p)$. Furthermore, $\mathbf{E}\left(\lambda_i^2\right) = 1$. In particular, $\mathrm{cov}(X_i) = \mathrm{Id}_p$.

- ρ is convex (and closed and proper).
- $P(\beta) = \sum_{k=1}^p f_k(\beta_k)$, i.e P works coordinate-wise on β. f_k is convex (and closed and proper).
- $\{\epsilon_i\}_{i=1}^n$ are independent of $\{X_i\}_{i=1}^n$.

The convexity of ρ and f_k's is needed to insure that certain functions that will appear below exist, and to guarantee uniqueness of the solution of our original M-estimation problem.

2.2 Results

Before we state the results, we need to present a short reminder concerning the prox function of convex optimization.

The prox function Suppose that ρ is a proper, closed, convex function (see Rockafellar (1997)). Then, following Moreau (1965), we denote by prox_c the unique solution of

$$\mathrm{prox}_c(\rho)(x) = \mathrm{argmin}_u \frac{1}{2c}(u-x)^2 + \rho(u) \; .$$

A key property of the prox is that if ρ has a subdifferential ψ,

$$\mathrm{prox}_c(\rho)(x) = (\mathrm{Id} + c\psi)^{-1}(x) \; .$$

In other words, despite the fact that ψ is potentially multivalued, $(\mathrm{Id} + c\psi)^{-1}$ is single valued. We refer the reader to Rockafellar (1997) and Lemaréchal and Sagastizábal (1997) for more information about the prox function.

2.2.1 The system

Our aim is to characterize $\|\widehat{\beta} - \beta_0\|$. To do so, we need to introduce two extra parameters, which we call $c_\tau > 0$ and $\nu(\tau)$. We will also need another parameter $K_\tau = \frac{\tau c_\tau}{n\nu(\tau)}$.

We call R_i the i-th residual (i.e $R_i = y_i - X_i'\widehat{\beta}$), $r_{i,[k]}$ the i-th residual we obtain by not using the k-th predictor $\{X_i(k)\}_{i=1}^n$ in a regression where we have changed the response y_i by removing the influence of $X_i(k)$ on it (namely the "new" response in this regression is $y_i - \beta_0(k)X_i(k)$). And we call ψ the subdifferential of ρ.

We will also need to introduce the random variables

$$Z_k = \left[\beta_0(k) + \frac{1}{\nu(\tau)}\frac{1}{n}\sum_{i=1}^n X_i(k) c_\tau \psi(r_{i,[k]})\right] \; .$$

Result 1. *The heuristic manipulations made in the supplementary material lead to the following results. We assume that p/n tends to a finite non-zero limit. We also assume that all the entries of β_0 are of comparable size.*

$\|\widehat{\beta} - \beta_0\|$, c_τ, $\nu(\tau)$ are asymptotically deterministic. Furthermore,

- *Call $\tilde{r}_{i,(i)} = \epsilon_i + \lambda_i \mathcal{N}(0, \|\widehat{\beta} - \beta_0\|)$, where λ_i is independent of the normal variable that multiplies it.*
- *We have $R_i \simeq r_{i,[k]}$ for all k and $R_i \simeq prox_{c_\tau \lambda_i^2}(\rho)(\tilde{r}_{i,(i)})$. Furthermore, $c\psi(R_i) \simeq \tilde{r}_{i,(i)} - R_i$.*
- *Call Z_k the random variables defined above and note that their marginal distribution is, when ϵ_i are i.i.d, ($\overset{\mathcal{L}}{=}$ is used to denote equality in law)*

$$Z_k \overset{\mathcal{L}}{=} \mathcal{N}\left(\beta_0(k), \frac{1}{n\nu(\tau)^2}\mathbf{E}\left(\frac{([prox_{c_\tau \lambda_i^2}(\rho)](\tilde{r}_{i,(i)}) - \tilde{r}_{i,(i)})^2}{\lambda_i^2}\right)\right) \; .$$

We have the relationships

$$\begin{cases} \frac{1}{n}\sum_{i=1}^n [prox_{c_\tau \lambda_i^2}(\rho)]'(\tilde{r}_{i,(i)}) & \simeq 1 - \nu(\tau) \; , \\ \frac{p}{n}\frac{1}{p}\sum_{k=1}^p [prox_{K_\tau}(f_k)]'(Z_k) & \simeq \nu(\tau) \; , \\ \forall 1 \leq k \leq p \; , prox_{K_\tau}(f_k)\,[Z_k] & \simeq \widehat{\beta}_k \; . \end{cases}$$

The last p equations relate $\|\widehat{\beta} - \beta_0\|^2$ (and $\mathbf{E}\left(\|\widehat{\beta} - \beta_0\|^2\right)$) and to $\nu(\tau)$ and c_τ, giving us a system of 3 equations in the three unknowns $\|\widehat{\beta} - \beta_0\|$, $\nu(\tau)$ and c_τ.

Examples of solutions to this system are given in Section 3.

2.2.2 Comments on the system and model

Choice of the model, impact of geometry We chose to model the predictors as $X_i = \lambda_i \mathcal{X}_i$ to study the sensitivity of our results to the geometry of the design matrix. As a matter of fact, when $\mathcal{X}_i$ is normal, standard results in concentration of measure theory tell us that for any given deterministic matrix A, we have $\sup_{i=1,\ldots,n} |\mathcal{X}_i' A \mathcal{X}_i / p - \mathrm{trace}\,(A)\,/p| \simeq 0$, provided the largest eigenvalue of A is much smaller than $\sqrt{p}$ (see Ledoux (2001)), $\lim p/n = \kappa > 0$ is bounded and p and n are sufficiently large. Taking $A = \mathrm{Id}_p$, we realize that this means that the points $\mathcal{X}_i$ are close to the unit sphere in $\mathbb{R}^p$, a well-known fact about high-dimensional Gaussian data. In fact they are close to the surface of any given ellipsoid (since those are defined by a positive-definite A). Hence, the Gaussian predictor model is extremely constraining geometrically. The model we look at allows to break somewhat these constraints. Indeed, our points are not close to a sphere if λ_i's are allowed to vary. However, they are still nearly orthogonal to one another, just like high-dimensional Gaussian random vectors. What the system shows is that our asymptotic results are quite sensitive to the underlying geometric assumptions made by the model about the data.

Interaction ϵ_i and ρ The system makes clear that in general the results will depend on the distribution of ϵ_i and ρ. An exception to this rule is the case where $\rho(x) = x^2/2$, i.e penalized least squares. An important subcase is LASSO, which is ℓ_1 penalized least squares. This follows from properties of the prox function that are detailed below.

Residuals One of the results of our manipulations is the fact that $R_i = \text{prox}_{c_\tau \lambda_i^2}(\tilde{r}_{i,(i)})$. The distribution of $\tilde{r}_{i,(i)}$ is extremely simple, since it is a sum of two independent random variables (though in general it is not Gaussian). Hence, this relationship between R_i and $\tilde{r}_{i,(i)}$ tells us pretty much everything about the marginal distribution of the residuals. It should also be noted that if $\widehat{\beta}_{(i)}$ denotes the estimator we get when the i-th observation has been removed, we have $\tilde{r}_{i,(i)} = \epsilon_i + X_i'(\widehat{\beta}_{(i)} - \beta_0)$. (Note that asymptotically, $\|\widehat{\beta}_{(i)} - \beta_0\| \simeq \|\widehat{\beta} - \beta_0\|$, hence the approximation made in the statement of our result.)

Beyond $\mathbf{cov}(X_i) = \mathbf{Id}_p$ The methods we use to get at the system above should work a priori for more complicated models for the data. However, the computations are much more involved and we do not present them here. We note however that the reasonably simple structure of the first two equations of our system likely to be shattered when $\text{cov}(X_i) \neq \text{Id}_p$. Also, we rely very strongly for the representation of $\widehat{\beta}_p$ very strongly on the fact that the penalty acts coordinate-wise. More complicated functions do not yield as simple a representation.

Robustness to distributional assumptions In our arguments, the normality of $\mathcal{X}_i$ plays a minor role. However, the concentration properties of the normal distribution are very important. We expect that the same results hold when we replace the normality assumption by assumptions guaranteeing concentration of quadratic forms.

About the entries of β_0 It seems that the assumption that the entries of β_0 are all of comparable size is not a strong requirement for the argument to go through. In particular, if β_0 is sparse, the same analysis should go through. (Our analysis is really about the entries of $\widehat{\beta}$ that correspond to small entries of β_0. The large entries of β_0 need to be treated separately but if there are few of them, it seems that they will not (in general) create much problem for global quantities such as $\|\widehat{\beta} - \beta_0\|$.)

Final simplifications The sums taken in the first two equations are large sums taken over weakly correlated random variables, so we expect that they stabilize and converge to their expectation. In the case of i.i.d errors, the system can therefore be simplified to

$$\begin{cases} \mathbf{E}\left([\text{prox}_{c_\tau \lambda_i^2}(\rho)]'(\tilde{r}_{i,(i)})\right) & \simeq 1 - \nu(\tau) \ , \\ \frac{p}{n}\frac{1}{p}\sum_{k=1}^{p} \mathbf{E}\left([\text{prox}_{K_\tau}(f_k)]'(Z_k)\right) & \simeq \nu(\tau) \ , \\ \forall 1 \leq k \leq p \ , \text{prox}_{K_\tau}(f_k)\left[Z_k\right] & \simeq \widehat{\beta}_k \ . \end{cases}$$

Further comments At this point, we have empirical evidence that the previous system can be solved in certain cases. However, we do not have a proof of existence and uniqueness of a solution to our 3-parameter system. We are actively working on making all of our assertions fully rigorous (including of course the validity of the system).

Particular values The case of no penalty is interesting, too. In this case, $\tau = 0$, and $f_k = 0$. Therefore, the second equation gives us $\nu(\tau) = p/n$. We also see that $K_\tau = 0$. and we are left with a system with two parameters only that is the same as the one we found in our previous work.

3 Examples

Penalized regression methods have become increasingly popular in statistics for handling high dimensional data. Understanding the behavior of these estimates in high dimensions is therefore of great theoretical and practical interest. A key quantity to understand is $\|\hat{\beta} - \beta_0\|_2^2$. We show that the functional system can be solved to yield results for $\|\hat{\beta} - \beta_0\|_2^2$ in these cases:

1. $f_k(x) = x^2/2$, $\rho(x) = x^2/2$ (Ridge regression);
2. $f_k(x) = x^2/2$, $\rho(x) = |x|$ (ℓ_2-penalized median regression);
3. $f_k(x) = |x|$, $\rho(x) = x^2/2$ (LASSO);
4. $f_k(x) = |x|$, $\rho(x) = |x|$ (ℓ_1-penalized median regression).

We then validate these results through simulations. We make extensive use of the following ideas:

Notes on the prox function. The prox function has a simple form for both the square and absolute value function, and the zero function:

1. For $f(x) = x^2/2$ we have $[prox_c(f)](x) = \frac{x}{1+c}$ and hence the derivative of the prox is the constant $\frac{1}{1+c}$
2. For $f(x) = |x|$, $[prox_c(f)](x) = \text{sgn}(x)(c - |x|)_+$; i.e. the prox of the absolute value function corresponds to soft-thresholding at c. Also, $[prox_c(f)]'(x) = \mathbf{1}_{[|x|>c]}$.
3. For $f \equiv 0$, $[prox_c(f)](x) = x$.

Stein's identity. We make frequent use of Stein's identity: for $Z \sim \mathcal{N}(0,1)$ and differentiable function f, we have (under mild regularity conditions) $\mathbf{E}(Zf(Z)) = \mathbf{E}(f'(Z))$. (In fact, this holds for any $\mathcal{N}(0,\sigma^2)$ random variable). These special cases are particularly important:

1. $\mathbf{E}(Z^2\mathbf{1}_{[|Z|\geq c]}) = 2c\phi(c) + \mathbb{P}(|Z| \geq c)$, where ϕ is the standard normal density
2. $\mathbf{E}(Z\mathbf{1}_{[Z\geq c]}) = \phi(c)$.

3.1 ℓ_2 Penalization, Gaussian Predictors

Simplify the system into two equations. In the case of Gaussian predictors $\lambda_i = 1$ for all i, considerably simplifying the functional system. For the case of ℓ_2 penalties, we aim to further simplify the system into a system of two equations in two unknowns: c_τ and $\|\hat\beta - \beta_0\|_2^2$. Recall that the prox for $x^2/2$ at c is the function $x/(1+c)$. This makes the second equation in the functional system $n\nu(\tau) = p/(1+K_\tau)$, from which we obtain the simple relation $\nu(\tau) = p/n - \tau c_\tau/n$. Now let $v^2 = \mathbb{E}\left([prox_{c_\tau}(\rho)](\tilde r) - \tilde r\right)^2$. Note that $Z_k \overset{\mathcal{L}}{=} beta_0(k) + vZ/sqrtn$ for $Z \sim \mathcal{N}(0,1)$. Thus by equation 3 of the system, $\mathbb{E}(\hat\beta_k - \beta_0(k))^2 = \frac{r^2/n}{(p/n)^2} + (\tau c_\tau/p)^2\beta_0(k)^2$. This yields a functional system of two equations for ℓ_2-penalized M-estimates:

$$\begin{cases} \mathbf{E}\left([prox_{c_\tau}(\rho)]'(\tilde r)\right) = 1 - \frac{p}{n} - \frac{\tau}{n}c_\tau, \\ \mathbf{E}\|\hat\beta - \beta_0\|_2^2 = \frac{n}{p}r^2 + \left(\frac{\tau c_\tau}{p}\right)^2\|\beta_0\|_2^2 \end{cases}$$

3.1.1 Case: $\rho(x) = x^2/2$; ridge regression

Though ridge regression can be analyzed more directly with random matrix methods (see **?** (citation masked)), we investigate our functional system's prediction for $\|\hat\beta - \beta_0\|$. Use the fact that the derivative of the prox of $x^2/$ is constant to obtain c_τ as the positive root of $1/(1+c_\tau) = 1 - p/n + \tau c_\tau/n$. Rewrite this equation as $\tau c_\tau/p = 1 - \frac{n}{p}\frac{c_\tau}{1+c_\tau}$. Moreover we can simplify v^2 to the form $\left(\frac{c_\tau}{1+c_\tau}\right)^2(\sigma_\epsilon + \mathbb{E}\|\hat\beta - \beta_0\|_2^2)$. Plugging these relations into the other equation yields:

$$\mathbf{E}\|\hat\beta - \beta_0\|_2^2 = \frac{\frac{p}{n}\left(\frac{c_\tau}{1+c_\tau}\right)^2\sigma_\epsilon^2 + \left(1 - \frac{n}{p}\frac{c_\tau}{1+c_\tau}\right)^2\|\beta_0\|_2^2}{1 - \frac{p}{n}\left(\frac{c_\tau}{1+c_\tau}\right)^2}.$$

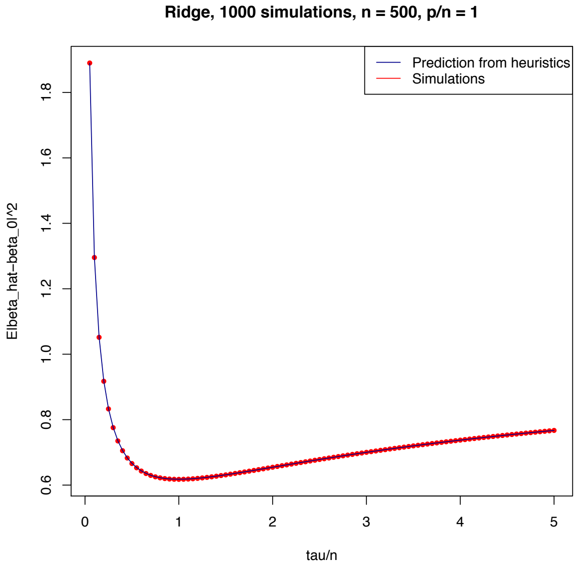

Figure 1: Ridge regression over a grid of 50 values of τ/n. Predictions vs realized values of $\|\hat\beta - \beta\|_2^2$ based on 1000 simulations. Errors are $\mathcal{N}(0,1)$. β_0 was drawn uniformly from the unit sphere.

Simulation results validate the above analytic expression, as seen in Figure (1). We note that $\|\hat\beta - \beta_0\|$ for ridge regression is the same for any error distribution with variance σ^2.

3.1.2 Case: $\rho(x) = |x|$; ℓ_2-penalized median regression

Simplify to Gaussian errors. Because of the simple form of the prox of $f(x) = x^2/2$, when the loss is ℓ_2 the distribution of the errors ϵ_i enters into the functional system only through $\mathbf{E}(\epsilon^2) = \sigma_\epsilon^2$. However, for general loss functions ρ, including $\rho(x) = |x|$, higher order parameters of the distribution of ϵ appear in the system. While it is certainly of theoretical interest to investigate the effect of non-Gaussian errors on the system, we know from experience that this can take significant time and resources. In the interest of presenting worked examples, we make the simplifying assumption of Gaussian errors. Therefore, the distribution of the random variables $\tilde{r}$ is $\mathcal{N}(0, s^2)$ where $s^2 = \sigma_\epsilon^2 + \mathbf{E}\|\hat{\beta} - \beta_0\|_2^2$.

In light of the above simplification, we have $\mathbb{P}(|Z| \geq c_\tau/s) = 1 - p/n + \tau c_\tau/n$ (recall that the derivative of the prox of $f(x) = |x|$ is an indicator function, so its expectation is a probability). Writing Φ^{-1} for the standard normal quantile function, we get $\frac{c_\tau}{s} = \Phi^{-1}\left(\frac{1}{2}(1 + \frac{p}{n} - \frac{\tau}{n}c_\tau\right)$, a relation between s and c_τ. Furthermore it is easy to show $v^2 = c_\tau^2 \mathbb{P}\left(|Z| \geq c_\tau/s\right) + s^2 \mathbf{E}\left(Z^2 \mathbf{1}_{[|Z| \leq c_\tau/s]}\right)$. Apply Stein's identity to the right hand side and write $\mathbf{E}\|\hat{\beta} - \beta\|_2^2 = s^2 - \sigma_\epsilon^2$; then we have:

$$s^2 - \sigma_\epsilon^2 = \left(\frac{\tau c_\tau}{p}\right)^2 \|\beta_0\|_2^2 + \frac{n}{p}\left\{c_\tau^2 \mathbb{P}\left(|Z| \geq \frac{c_\tau}{s}\right) + s^2\left[1 - \mathbb{P}\left(|Z| \geq \frac{c_\tau}{s}\right) - 2\frac{c_\tau}{s}\phi(\frac{c_\tau}{s})\right]\right\},$$

Plugging into the above display $1 - \frac{p}{n} - \frac{\tau}{n}c_\tau$ for $\mathbb{P}\left(|Z| \geq \frac{c_\tau}{s}\right)$, $\Phi^{-1}\left(\frac{1}{2}(1 + \frac{p}{n} - \frac{\tau}{n}c_\tau)\right)$ for c_τ/s, and $\left(c_\tau/\Phi^{-1}\left(\frac{1}{2}(1 + \frac{p}{n} - \frac{\tau}{n}c_\tau)\right)\right)^2$ for s^2 yields an equation which depends only on c_τ. Numerically find the root value c_τ^* of this equation, then plug back into the equation defining s to obtain:

$$\mathbf{E}\|\hat{\beta} - \beta_0\|_2^2 = \left(c_\tau^*/\Phi^{-1}\left(\frac{1}{2}(1 + \frac{p}{n} - \frac{\tau}{n}c_\tau^*)\right)\right)^2 - \sigma_\epsilon^2.$$

Figure (2) shows the agreement between simulations and our heuristic predictions.

3.2 ℓ_1-penalization, Gaussian predictors.

Regularization by the ℓ_1 norm has received considerable attention since this penalty produces sparse solutions in regression settings. This is highly desirable from a model selection standpoint. Applying our functional system to the ℓ_1 penalty yields some interesting results:

We can interpret $\nu(\tau)$. Recall that $[prox_c(|\cdot|)](y)$ was the soft-threshold of y at c and that $[prox_c(|\cdot|)]'(y) = \mathbf{1}_{[(]}|y| \geq c)$. Returning to the full functional system of three equation in three unknowns (for the Gaussian case), we see by the second equation that $n\nu(\tau) = \left|\{k : \hat{\beta}_k \neq 0\}\right|$; that is, $n\nu(\tau)$ *counts* the number of nonzero coefficients in $\hat{\beta}$.

Simplification of β_0 to the Gaussian case. In light of the remark above, we see that the likelihood that $\hat{\beta}_j \neq 0$ for a given j depends on the size of the true $\beta_0(j)$. At this point many investigations of ℓ_1 penalties impose sparsity constraints on β_0 but such assumptions seem to increase the challenge of solving the functional system. We instead make the simplifying assumption that $\beta_0(j) \overset{iid}{\sim} \frac{\eta}{\sqrt{p}}Z$. Then for each k $Z_k \sim \mathcal{N}(0, t^2)$ where $t^2 = v^2/n\nu(\tau)^2 + \eta^2/p$. Assuming concentration, the second equation in the functional system becomes $n\nu(\tau) = p\mathbb{P}(t|Z| \geq K_\tau)$.

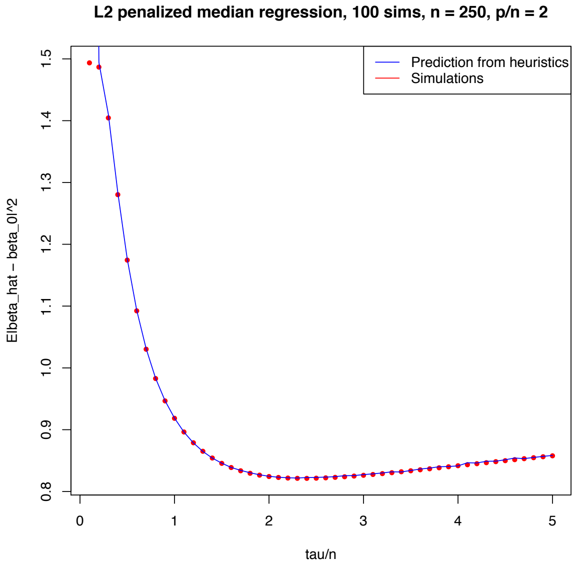

Figure 2: Median regression with L2 penalty. Predictions vs realized values of $\|\hat{\beta} - \beta\|_2^2$ based on 100 simulations. β_0 was drawn uniformly from the unit sphere.

Consequently, with high probability β_0 will be distributed diffusely near the unit sphere in $\mathbb{R}^p$. It is known that ℓ_1 penalization performs poorly in this situation (ℓ_2 regularization is preferable). But in making this assumption on β_0 we make the functional system more analytically tractable, and we show that our predictions are valid even in this "worst case" setting for ℓ_1 penalties.

3.2.1 Case: $\rho(x) = x^2/2$; LASSO

Some recent work has been done on the theory of LASSO - see e.g Donoho et al. (2009). The approached used there is completely different from ours but our results are of course consistent with our findings. We note that we could not extend the approach taken there to cover more general loss functions (ρ). This is in part what lead us to this work. In that sense our results are more general.

Recall that the derivative of the prox of $x^2/2$ is a constant, therefore the functional system gives $1 -$

$\nu(\tau) = 1/(1 + c_\tau)$. This admits the simple relationship $\nu(\tau) = \frac{c_\tau}{1+c_\tau}$. Therefore, $v^2/\nu(\tau)^2 = s^2$ and $t^2 = \frac{s^2}{n} + \frac{\eta^2}{p}$. Moreover, $K_\tau/t = \Phi^{-1}\left(1 - n\nu(\tau)/2p\right)$. Since $K_\tau = \frac{\tau}{n(1-\nu(\tau))}$ this means we can express t in terms of $\nu(\tau)$. Turning to the form of the distribution of $\hat{\beta}_k$ given in the system we have, after several applications of Stein's identity,

$$\mathbf{E}\left(\hat{\beta}_k - \beta_0(k)\right)^2 = t^2\mathbf{E}\big([prox_{\frac{K_\tau}{t}}(|\cdot|)](Z)\big)^2 \\ -\frac{\eta^2}{p}\left(1 - 2\frac{n}{p}\nu(\tau)\right)$$

If we define $g(z) = \mathbf{E}\big([prox_z(f_k)](Z)\big)^2$ we can write $s^2 - \sigma_\epsilon^2 = pt^2 g\left(\Phi^{-1}\left(1 - n\nu(\tau)/2p\right)\right) - \eta^2(1 - 2n\nu(\tau)/p)$. Moreover we have $s^2 = nt^2 - n\eta^2/p$. Plugging this into the above display and substituting $t = K_\tau/\Phi^{-1}\left(1 - n\nu(\tau)/2p\right)$ yields an equation in the single unknown $\nu(\tau)$. Numerically find the root $\nu^*(\tau)$ and plug back in to obtain:

$$\mathbf{E}\|\hat{\beta} - \beta_0\|_2^2 = n\left[\frac{\tau/n\left[(1-\nu^*(\tau))\right]}{\Phi^{-1}\left(1 - \frac{n}{p}\frac{\nu^*(\tau)}{2}\right)}\right]^2 - \frac{n}{p}\eta^2 - \sigma_\epsilon^2.$$

Note that the above computations are the same for any error distribution with variance σ_ϵ^2. Thus our heuristics predict that the norm of $\hat{\beta} - \beta_0$ for LASSO is the same for any error distribution with variance σ_ϵ^2. This conjecture is is supported by simulations using the Gaussian and Double Exponential error distributions. Figure (3) confirms the accuracy of the predictions of the functional system.

3.2.2 Case: $\rho(x) = |x|$; ℓ_1-penalized median regression

For further simplification assume the errors are Gaussian and that $\eta = 0$; that is, the model is fully sparse. Then $c_\tau/s = \Phi^{-1}\left((1 + \nu(\tau))/2\right)$. Recall $K_\tau/t = \Phi^{-1}\left(1 - n\nu(\tau)/2p\right)$. Dividing the first equation by the second and using the definition of K_τ gives:

$$\frac{t}{s} = \frac{\tau}{n\nu(\tau)}\frac{\Phi^{-1}\left(\frac{1}{2}(1+\nu(\tau))\right)}{\Phi^{-1}\left(1 - \frac{n}{p}\frac{\nu(\tau)}{2}\right)}.$$

Now write $v^2 = s^2\mathbf{E}\left([prox_{\frac{c_\tau}{s}}(\rho)](Z) - Z\right)^2$. Defining the function $h(z) = \mathbf{E}\left([prox_z(\rho)](Z) - Z\right)^2$ we have the equation:

$$\frac{v^2}{s^2} = h\left(\Phi^{-1}\left((1+\nu(\tau))/2\right)\right).$$

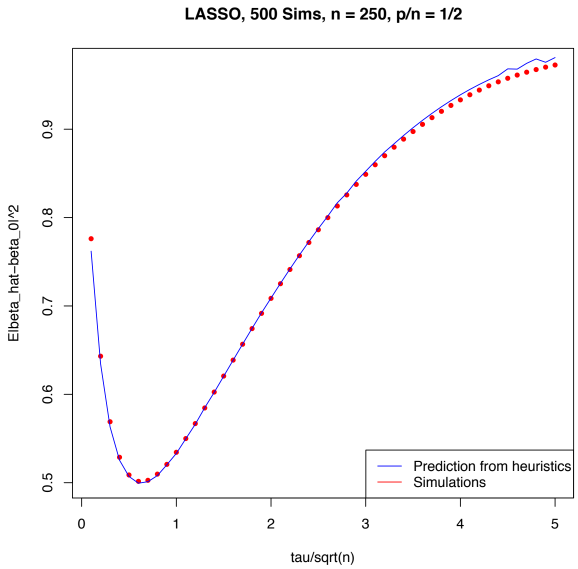

Figure 3: LASSO. Prediction vs realized values of $\|\hat{\beta} - \beta_0\|_2^2$ for 500 simulations. For each run $\beta_0 \sim \boldsymbol{Z}/\sqrt{p}$ where $\boldsymbol{Z}$ is a vector of i.i.d. $\mathcal{N}(0,1)$ random variables.

Since $\eta = 0$ we have $(t/v)^2 = 1/(n\nu(\tau)^2)$; on the other hand, $(t/v)^2 = (t/s)^2\,(s/v)^2$. We showed both ratios on the right hand side depend on $\nu(\tau)$ only. This gives an equation that depends solely on $\nu(\tau)$. Solve numerically to find the root $\nu^*(\tau)$. Noting that the identity $s^2 - \sigma_\epsilon^2 = pt^2 g\left(\Phi^{-1}\left(1 - n\nu(\tau)/2p\right)\right)$ derived in the LASSO case holds in this situation as well, it follows that we can obtain s by solving:

$$1 - \frac{\sigma_\epsilon^2}{s^2} = \frac{p}{n}\left[\frac{\tau}{\sqrt{n}\nu^*(\tau)}\frac{\Phi^{-1}\left(\frac{1}{2}(1+\nu^*(\tau))\right)}{\Phi^{-1}\left(1 - \frac{n}{p}\frac{\nu^*(\tau)}{2}\right)}\right]^2 \\ \times g\left(\Phi^{-1}\left(1 - \frac{n\nu^*(\tau)}{2p}\right)\right).$$

Upon solving for s we obtain: $\mathbf{E}\|\hat{\beta} - \beta_0\|_2^2 = s^2 - \sigma_\epsilon^2$. Simulation results (omitted) confirm the accuracy of our predictions.

3.2.3 LASSO with elliptical predictors

As our system indicates, the theoretical predictions we are making are sensitive to our implicit assumptions about the underlying geometry of the dataset. In particular, it is clear that doing computations in the Gaussian predictor case is not enough: these results will not generalize to datasets with a more complicated geometry, though they might generalize to distributions that share the concentration properties of the Gaussian distribution. These distributions may be in appearance

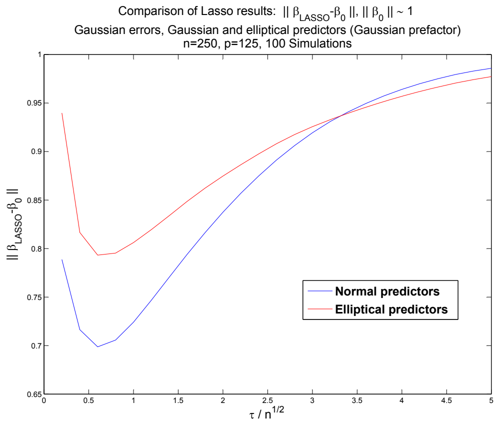

Figure 4: **Impact of geometry on Lasso predictions** In these simulations, λ_i are i.i.d $\mathcal{N}(0,1)$. $n = 250$, $p = 125$ and the errors were normal. There were $N = 100$ simulations. β_0 was drawn independently for each simulations, each entry being i.i.d $\mathcal{N}(0, 1/p)$. Each curve represent the average over our 100 simulations of $\|\widehat{\beta}_{LASSO} - \beta_0\|$. Our simulations clearly illustrate the impact of the geometry on the LASSO results: the behavior of the solution is different according to whether the predictors are elliptical or Gaussian.

different from the normal (for instance, we might be able to replace the requirement that the entries of X_i be i.i.d Gaussian by the requirement that they be i.i.d entries with mean 0 and be bounded), but effectively share a common geometry for the dataset (in the case of i.i.d bounded entries we can see it by applying Talagrand's concentration inequality (see Ledoux (2001), Corollary 4.10).

So we investigate the LASSO situation when the predictors are elliptical. We assume (to make computations simpler) that β_0 has i.i.d entries with $\beta_0(1) = \frac{\eta}{\sqrt{p}}\mathcal{N}(0,1)$. The same caveats as the ones discussed above in a similar situation apply for this model. The computations in this case are more complicated than in the Gaussian predictor case, but nonetheless quite doable. Using the fact that $\text{prox}_c(|\cdot|^2/2)(x) = x/(1+c)$, the first equation of our system becomes $\mathbf{E}\left(1/(1+c_\tau\lambda_i^2)\right) = 1 - \nu(\tau)$. Hence, if c_τ is given, we can numerically solve for $\nu(\tau)$. (This also gives us the value of K_τ.) We also have $\tilde{r}_{i,(i)} - \text{prox}_{c_\tau\lambda_i^2}(\tilde{r}_{i,(i)}) = [c_\tau\lambda_i^2/(1+c_\tau\lambda_i^2)]\tilde{r}_{i,(i)}$ and therefore $Z_k = \beta_0(k)/\sqrt{p} + \mathcal{N}(0, \sigma^2_{\text{Ellip}})$ with $\sigma^2_{\text{Ellip}} = \frac{\|\widehat{\beta}-\beta_0\|^2}{n\nu^2(\tau)}\mathbf{E}\left(\frac{c_\tau^2\lambda_i^4}{(1+c_\tau\lambda_i^2)^2}\right) + \frac{\sigma_\epsilon^2}{n\nu^2(\tau)}\mathbf{E}\left(\frac{c_\tau^2\lambda_i^2}{(1+c_\tau\lambda_i^2)^2}\right)$. As was noted above, with our assumptions on β_0, if we call $s_1^2 = \frac{\eta^2}{p} + \sigma^2_{\text{Ellip}}$, we have the approximation

$$s_1 \simeq K_\tau/\Phi^{-1}(1 - n\nu(\tau)/2p) \ .$$

Once again, if c_τ is given (and hence $\nu(\tau)$ is known), we can extract s_1 from the above equation - since we already know K_τ.

Working with the equations defining the $\widehat{\beta}_k$'s we also arrive at

$$\|\widehat{\beta} - \beta_0\|^2 \simeq ps_1^2 G(K_\tau/s_1) - 2\eta^2\frac{n}{p}\nu(\tau) + \eta^2 \ ,$$

where $G(T) = \mathbf{E}\left((|Z| - T)_+^2\right)$, where Z is $\mathcal{N}(0,1)$.

After we remark that $\mathbf{E}Z_k^2 = s_1^2$ (where this expectation is unconditional on $\beta_0(k)$, we can relate s_1^2 and $\|\widehat{\beta} - \beta_0\|^2$. Working on this relation and injecting $s_1 \simeq K_\tau/\Phi^{-1}(1 - n\nu(\tau)/2p)$ in it, we get

$$\mathbf{E}\left(\frac{\lambda_i^4}{(1+c_\tau\lambda_i^2)^2}\right)\|\widehat{\beta} - \beta_0\|^2 \simeq$$

$$\left[\frac{\tau^2}{n\left[\Phi^{-1}(1 - \frac{n\nu(\tau)}{2p})\right]^2} - \frac{\tau^2\eta^2}{npK_\tau^2} - \sigma_\epsilon^2\mathbf{E}\left(\frac{\lambda_i^2}{(1+c_\tau\lambda_i^2)^2}\right)\right]$$

We can now replace $\|\widehat{\beta} - \beta_0\|^2$ and s_1 in the previous equation by their value in terms of $\nu(\tau)$ which we have found earlier. This gives us a single equation in terms of $\nu(\tau)$ which we can solve numerically and propagate the results to find $\|\widehat{\beta} - \beta_0\|^2$. For space reasons, these results will be presented elsewhere.

3.3 Results

Main results. In the special case that the predictors are Gaussian and that the loss function, or the penalty function, is either the ℓ_2 norm or the ℓ_1 norm, simulation results strongly suggest that the behavior of $\|\hat{\beta} - \beta_0\|_2^2$ is governed by our functional system. At the heart of this system is the behavior of the prox transformation of convex functions. Both $x^2/2$ and $|x|$ have simple prox transformations which make the functional system more tractable. Since the ℓ_2 and ℓ_1 norms are popular amongst statisticians and data practitioners, our findings in these examples may be of significant interest.

ℓ_2 penalties. The especially simple form of the prox of $x^2/2$ allows us to simplify the original functional system of three equations in three unknown to a system of two equations in two unknowns. Simulations validated the predictions of the functional system in the case of β_0 diffuse on the sphere, a situation in which it is known that ℓ_2 penalization performs well.

1. **Ridge.** Putting $\rho(x) = x^2/2$ for the loss function further simplifies the functional system. We obtained an analytic expression for $\|\hat{\beta} - \beta\|_2^2$. The behavior of $\|\hat{\beta} - \beta_0\|_2^2$ is the same for all error distributions with equal variance.

2. **ℓ_2-penalized median regression.** It is possible to numerically solve for $\|\hat{\beta} - \beta_0\|_2^2$.

ℓ_1 penalties. This case is of considerable interest due to the increasing popularity of ℓ_1 penalization methods. The unknown constant $\nu(\tau)$ in our functional system becomes interpretable as the number of $\hat{\beta}_j$'s not equal to zero divided by n. However the equations are difficult to solve for general β_0. We assumed $\beta_0(j) \sim \eta Z/\sqrt{p}$ for each j, which leads to diffuse β_0's, a scenario in which ℓ_1 penalization fares poorly. However, simulations validated the predictions of the functional system even in these cases.

1. **LASSO:**It is possible to solve for $\|\hat{\beta} - \beta\|_2^2$ numerically. Our heuristic suggest that the behavior of this quantity is unchanged for error distributions with equal variance, a somewhat surprising result confirmed in simulations.

2. **ℓ_1-regularized median regression:** The functional system is very complicated for this method. It is possible to numerically solve the system in the fully sparse case, i.e. $\beta_0 = 0$. Simulation results, not presented, confirmed the predictions of the functional system in this case.

Implications. Our investigation showed that robust regression estimates behave very differently in high dimension than in low dimensions. To fix ideas, suppose the errors have a double exponential distribution. It is well known that median regression is more efficient than least squares in low dimensions (i.e. $p/n \to 0$). But this is not true for penalized robust regression estimates in high dimensions. Figure (5) shows Ridge Regression performing uniformly better across τ than ℓ_2-regularized median regression when $p/n = 2$, even though the errors are double exponential. Our proposed functional system has the potential to explain these differences.

4 Conclusion

Our analysis sheds light on the difficult problem of understanding the behavior of robust regression estimators in high-dimension. Our work makes clear that the geometry of the dataset drives the results. We also linked our M-estimation problem with various branches of optimization theory and probability.

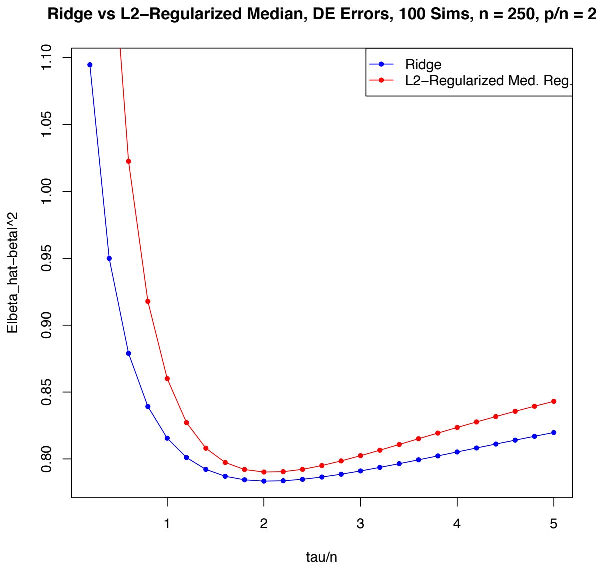

Figure 5: Ridge vs. L2-Regularized Median Regression with Double Exponential errors. Ridge outperforms the regularized median method even though the errors have heavier tails than a Gaussian. Based on 100 simulations and a grid of 25 values of τ/n. β_0 drawn uniformly from the sphere.

We have shown that our heuristics (which are grounded in rigorous theory) are well verified in simulations. We should note that we have adapted them to handle heteroskedastic errors, weighted robust regression and many such variants.

The simulations also show that relying on low-dimensional intuition to decide which method to apply to high-dimensional data can lead to erroneous conclusions. Strikingly, relying on ℓ_1 loss when the errors are double exponential (the correct thing to do in low-dimension) leads to worst performance in high-dimension than using ℓ_2 loss.

Acknowledgements

We thank funding agencies and a private foundation for their support.

References

Donoho, D., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *PNAS* **106**, 18914–18919.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799–821.

Ledoux, M. (2001). *The concentration of measure* *phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.

Lemaréchal, C. and Sagastizábal, C. (1997). Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries. *SIAM J. Optim.* **7**, 367–385. URL http://dx.doi.org/10.1137/S1052623494267127.

Moreau, J.-J. (1965). Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France* **93**, 273–299.

Rockafellar, R. T. (1997). *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ. Reprint of the 1970 original, Princeton Paperbacks.

Tibshirani, R. (1995). Regression shrinkage with the lasso. *JRSS B* .

Appendix

In this appendix, we explain how to derive the system of equations we obtained for characterizing $\|\widehat{\beta}-\beta_0\|$. Our derivation is heuristic, so we make assumptions as needed as we go along.

Recall that $y_i = X_i'\beta_0 + \epsilon_i$.

We call

$$\widehat{\delta} = \widehat{\beta} - \beta_0 \ ,$$

and, for a scalar c,

$$g_c(x) = x + c\psi(x) \ .$$

In this derivation, we are assuming that β_0 has all its entries close to 0.

The gradient definition of $\widehat{\beta}$ is

$$\sum_{i=1}^n -X_i\psi(\epsilon_i - X_i'\widehat{\delta}) + \tau\nabla P_{\widehat{\beta}} = 0 \ . \tag{1}$$

Throughout, we assume that X_i are elliptical, i.e in matrix form

$$X = D_\lambda \mathcal{X} \ .$$

$\mathcal{X}$ is a matrix whose entries are i.i.d $\mathcal{N}(0,1)$. D_λ is a diagonal matrix independent of $\mathcal{X}$.

When we leave out one observation, we have

$$\sum_{j\neq i} -X_j\psi(\epsilon_j - X_j'\widehat{\delta}_{(i)}) + \tau\nabla P_{\widehat{\beta}_{(i)}} = 0 \ .$$

Call $R_j = \epsilon_j - X_j'\widehat{\delta}$ and $\tilde{r}_{j,(i)} = \epsilon_j - X_j'\widehat{\delta}_{(i)}$.

Take the difference with above, Taylor expand around $\tilde{r}_{j,(i)}$ $(j\neq i)$ and get

$$-X_i\psi(R_i) + \sum_{j\neq i} X_j\psi'(\tilde{r}_{j,(i)})(\tilde{r}_{j,(i)} - R_j) \\ +\tau H(\widehat{\beta} - \widehat{\beta}_{(i)}) \simeq 0 \ .$$

Here H is the Hessian of P - it is computed at $\widehat{\beta}$. Of course, for the purpose of this derivation, we are expecting that $\nabla P_{\widehat{\beta}} - \nabla P_{\widehat{\beta}_{(i)}} \simeq H_{\widehat{\beta}_{(i)}}(\widehat{\beta}-\widehat{\beta}_{(i)}) \simeq H(\widehat{\beta}-\widehat{\beta}_{(i)})$. (For many losses this will be the case, for others it will be harder to see.)

Now, since $R_j - \tilde{r}_{j,(i)} = X_j'(\widehat{\beta}_{(i)} - \widehat{\beta})$, we have

$$-X_i\psi(R_i) + (\sum_{j\neq i}\psi'(\tilde{r}_{j,(i)})X_jX_j' + \tau H_{\widehat{\beta}})(\widehat{\beta} - \widehat{\beta}_{(i)}) = 0 \ .$$

So, calling $S_i = \sum_{j\neq i}\psi'(\tilde{r}_{j,(i)})X_jX_j'$, we have

$$\widehat{\beta} - \widehat{\beta}_{(i)} \simeq (S_i + \tau H)^{-1}X_i\psi(R_i) \ .$$

Multiply on the left by X_i' and we get, if λ_i is the (i,i) entry of D_λ,

$$\tilde{r}_{i,(i)} - R_i \simeq \lambda_i^2 \text{trace}\left((S_i + \tau H)^{-1}\right) \psi(R_i) \;,$$

by appealing to the well known fact that if A is matrix that is independent of X_i and whose eigenvalues are well-behaved, we have $X_i' A X_i \simeq \text{trace}\,(A)$ (see for instance Ledoux (2001)). We are also implicitly using the fact that H is diagonal and $H \simeq H_{\widehat{\beta}_{(i)}}$, so we can replace H by another matrix independent of X_i.

Calling $c_\tau = \text{trace}\left((S_i + \tau H)^{-1}\right)$, we see that

$$(R_i + \lambda_i^2 c_\tau \psi(R_i)) \simeq \tilde{r}_{i,(i)} \;.$$

Our experience in random matrix theory suggests that c_τ is asymptotically deterministic. (Rigorous arguments can be made but it is not the point of this conference paper.)

Here, we are going to assume that the penalty acts coordinate wise. We call $X_i = \left[[V_i // X_i(p)\right]$, where V_i is $p-1$ dimensional. We call $\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_{S_p} \\ \widehat{\beta}_p \end{bmatrix}$. We call $\widehat{\gamma}$ the estimator we get by using only V_i's as predictors (so we drop the last predictor). And we call γ_0 the restriction of β_0 to its first $p-1$ coordinates.

Using these notations, we have, if $P^{(2)}$ denotes the restriction of P to its first $p-1$ coordinates,

$$\sum -X_i \psi(\epsilon_i - V_i' \widehat{\delta}_{S_p} - X_i(p)\widehat{\delta}_p) + \tau \nabla P_{\widehat{\beta}} = 0 \;,$$

and

$$\sum -V_j \psi(\epsilon_i - V_i'(\widehat{\gamma} - \gamma_0)) + \tau \nabla P^{(2)}_{\widehat{\gamma}} = 0 \;.$$

In the second equation we are working with "data" $y_i = \epsilon_i + V_i' \beta_{0,S_p}$. We call $\gamma_0 = \beta_{0,S_p}$ and $\widehat{\zeta} = \widehat{\gamma} - \gamma_0$.

Because we assume that $\beta_0(p)$ is small, we expect $r_{i,[p]} = \epsilon_i - V_i' \widehat{\zeta}$ to be close to $R_i = \epsilon_i - X_i' \widehat{\delta}$.

Now since we are assuming that on its first $p-1$ coordinates, P acts like $P^{(2)}$, the functional form of ∇P is the same as that of ∇P_2. We now expand the p-th coordinate of the first equation and take difference from the others.

We get, for the last coordinate: (e_p is the p-th canonical basis vector)

$$\sum -X_i(p) \Big(\psi(r_{i,[p]}) + \psi'(r_{i,[p]})[-\widehat{\delta}_p X_i(p) \\ + V_i'(\widehat{\zeta} - \widehat{\delta}_{S_p})] \Big) + \tau \nabla P_{\widehat{\beta}} e_p = 0$$

For the others, we get (making approximations on the value of the gradients and the fact that we can take the Hessian to approximate)

$$\sum_i V_i(\psi(r_{i,[p]}) - \psi(R_i)) + \tau H_{\widehat{\gamma}}(\widehat{\beta}_{S_p} - \widehat{\gamma}) \simeq 0 \;.$$

In other words,

$$\sum_i \psi'(R_i) V_i [V_i'(\widehat{\delta}_{S_p} - \widehat{\zeta}) + \widehat{\delta}_p X_i(p)] + \tau H_{\widehat{\gamma}}(\widehat{\delta}_{S_p} - \widehat{\zeta}) \simeq 0 \;.$$

Hence, calling $\mathfrak{S}_p = \sum_i \psi'(R_i) V_i V_i'$, $u_p = \sum \psi'(R_i) V_i X_i(p)$,

$$(\mathfrak{S}_p + \tau H_{\widehat{\gamma}})(\widehat{\delta}_{S_p} - \widehat{\zeta}) + u_p \widehat{\delta}_p \simeq 0 \;.$$

The equation for the first line gave

$$-\sum X_i(p)\psi(r_{i,[p]}) + \widehat{\delta}_p \sum X_i(p)\psi'(r_{i,[p]}) \\ -u_p'(\widehat{\zeta} - \widehat{\delta}_{S_p}) + \tau \nabla P_{\widehat{\beta}} e_p \simeq 0 \;.$$

So we conclude that

$$-\sum X_i(p)\psi(r_{i,[p]}) + \widehat{\delta}_p \Big[\sum X_i(p)\psi'(r_{i,[p]}) \\ -u_p'(\mathfrak{S}_p + \tau H_{\widehat{\gamma}})^{-1} u_p \Big] + \tau \nabla P_{\widehat{\beta}} e_p \simeq 0 \;.$$

In other words,

$$\widehat{\delta}_p \simeq \frac{\sum X_i(p)\psi(r_{i,[p]}) - \tau \nabla P_{\widehat{\beta}} e_p}{\sum X_i(p)\psi'(r_{i,[p]}) - u_p'(\mathfrak{S}_p + \tau H_{\widehat{\gamma}})^{-1} u_p} \;.$$

Note that we can write $u_p = V' D X(p)$, where D is a diagonal matrix containing $\psi'(r_{i,[p]})$, V is a $n \times (p-1)$ matrix with the V_i's as its rows and $X(p)$ is an $n \times 1$ vector. Note that we can write $V = D_\lambda \mathcal{V}$, where $\mathcal{V}$ is a $n \times (p-1)$ dimensional matrix with i.i.d entries.

The denominator has a simple structure: it is of the form

$$X(p)' D^{1/2} M D^{1/2} X(p) \;. \tag{2}$$

with

$$M = \left[\text{Id} - D^{1/2} V (V' D V + \tau H)^{-1} V' D^{1/2} \right] \;.$$

Since $X(p) = D_\lambda \mathcal{X}(p)$, we can also rewrite this as, using the fact that D_λ is diagonal,

$$\mathcal{X}(p)' D_\lambda D^{1/2} M D^{1/2} D_\lambda \mathcal{X}(p) \;.$$

Because $\mathcal{X}(p)$ is an n dimensional vector with i.i.d Gaussian entries, we have, using the fact that M, D_λ and D are independent of $\mathcal{X}$,

$$\mathcal{X}(p)' D_\lambda D^{1/2} M D^{1/2} D_\lambda \mathcal{X}(p) \simeq \\ \text{trace}\left(D_\lambda D^{1/2} M D^{1/2} D_\lambda \right) = \\ \sum_{i=1}^{n} \lambda_i^2 \psi'(\tilde{r}_{i,[p]}) M_{i,i} \;.$$

It is easy to verify that the diagonal entries of M are of the form (recall that $S = V'DV$)

$$\begin{aligned} M_{i,i} &= 1 - d_{ii}V_i'(S+\tau H)^{-1}V_i \\ &= \frac{1}{1+d_{ii}V_i'(S_i+\tau H)^{-1}V_i} , \end{aligned}$$

and therefore,

$$\begin{gathered} \sum_i \frac{1}{1+d_{ii}V_i'(S_i+\tau H)^{-1}V_i} = \\ n - \sum_{i=1}^n d_{ii}V_i'(S+\tau H)^{-1}V_i = \\ n - \operatorname{trace}\left(S(S+\tau H)^{-1}\right) = \\ n - p + \operatorname{trace}\left(\tau H(S+\tau H)^{-1}\right) . \end{gathered}$$

From the previous expression we also see that

$$\begin{gathered} \sum_i \frac{d_{ii}V_i'(S_i+\tau H)^{-1}V_i}{1+d_{ii}V_i'(S_i+\tau H)^{-1}V_i} \\ = p - \tau\operatorname{trace}\left(H(S+\tau H)^{-1}\right) . \end{gathered}$$

Using the usual concentration of quadratic forms trick, we get, assuming that $H_{\widehat{\gamma}}$ and D are sufficiently independent of V_i,

$$\begin{gathered} \sum_i \frac{d_{ii}V_i'(S_i+\tau H)^{-1}V_i}{1+d_{ii}V_i'(S_i+\tau H)^{-1}V_i} \simeq \\ \sum_i \frac{d_{ii}\lambda_i^2\operatorname{trace}\left((S+\tau H)^{-1}\right)}{1+d_{ii}V_i'(S_i+\tau H)^{-1}V_i} \end{gathered}$$

and hence, assuming $c_\tau \simeq \operatorname{trace}\left((S+\tau H)^{-1}\right)$,

$$\begin{gathered} c_\tau \sum_i \frac{d_{ii}\lambda_i^2}{1+d_{ii}V_i'(S_i+\tau H)^{-1}V_i} \simeq \\ p - \tau\operatorname{trace}\left(H(S+\tau H)^{-1}\right) . \end{gathered}$$

Recall that

$$\begin{aligned} \sum_i \frac{d_{ii}\lambda_i^2}{1+d_{ii}V_i'(S_i+\tau H)^{-1}V_i} &= \sum_i d_{ii}\lambda_i^2 M_{i,i} \\ &= \operatorname{trace}\left(D_\lambda D^{1/2}MD^{1/2}D_\lambda\right) . \end{aligned}$$

We have therefore established that

$$\begin{gathered} c_\tau\left[\sum X_i(p)\psi'(r_{i,[p]}) - u_p'(\mathfrak{S}_p+\tau H_{\widehat{\gamma}})^{-1}u_p\right] \\ \simeq p - \tau\operatorname{trace}\left(H(S+\tau H)^{-1}\right) \end{gathered}$$

Let us call

$$\nu(\tau) = \frac{p}{n} - \frac{\tau}{n}\operatorname{trace}\left(H(S+\tau H)^{-1}\right)$$

We now have the approximation

$$\frac{\nu(\tau)}{c_\tau}\widehat{\delta}_p \simeq \frac{1}{n}\left(\sum_{i=1}^n X_i(p)\psi(r_{i,[p]}) - \tau\nabla P_{\widehat{\beta}}e_p\right) .$$

Hence,

$$\frac{\nu(\tau)}{c_\tau}\widehat{\beta}_p + \frac{\tau}{n}\nabla P_{\widehat{\beta}}e_p = \frac{\nu(\tau)}{c_\tau}\beta_0(p) + \frac{1}{n}\sum_{i=1}^n X_i(p)\psi(r_{i,[p]}) . \tag{3}$$

Relationship between $\boldsymbol{\nu(\tau)}$ and $\boldsymbol{c_\tau}$. The computations above essentially indicate that

$$\nu(\tau) \simeq 1 - \frac{1}{n}\sum_{i=1}^n \frac{1}{1+\lambda_i^2\psi'(r_{i,[p]})c_\tau} ,$$

which we rephrase as (this is hard to find but fairly easy to verify)

$$\mathbf{E}\left((g^{-1}_{c_\tau\lambda_i^2})'(\tilde{r}_{i,(i)})\right) \simeq 1 - \nu(\tau) .$$

Another relationship can be derived between c_τ and $\nu(\tau)$. Namely

$$1 \simeq \operatorname{trace}\left((\tau c_\tau H + n\nu(\tau)\mathrm{Id})^{-1}\right) .$$

We refer the reader to e.g **?** (citation masked) for a simple derivation (that includes more general cases). In general this requires that H be independent of S and be "well-behaved". So 3 parameters have appeared: $\|\widehat{\beta}-\beta_0\|$, $\nu(\tau)$ and c_τ. At this point we have 2 equations relating these three parameters.

Using the structure of the penalty Suppose that the penalty is of the form

$$P(\beta) = \sum_{i=1}^p f_i(\beta_i) .$$

Then the Hessian is diagonal and the (sub-)gradient is just

$$\nabla P_\beta e_p = f_p'(\beta_p) .$$

Hence, if we rewrite Equation (3) as

$$\begin{gathered} (\mathrm{Id} + \frac{\tau c_\tau}{\nu(\tau)n}\frac{\partial f_p(x)}{\partial x})\widehat{\beta}_p \\ = \beta_0(p) + \frac{c_\tau}{\nu(\tau)}\frac{1}{n}\sum_{i=1}^n X_i(p)\psi(r_{i,[p]}) , \end{gathered}$$

we get a prox function representation

$$\widehat{\beta}_p = \operatorname{prox}_{\frac{\tau c_\tau}{\nu(\tau)n}}(f_p)\left[\beta_0(p) + \frac{1}{\nu(\tau)}\frac{1}{n}\sum_{i=1}^n X_i(p)c_\tau\psi(r_{i,[p]})\right] .$$

We call

$$Z_p = \beta_0(p) + \frac{1}{\nu(\tau)}\frac{1}{n}\sum_{i=1}^n X_i(p)c_\tau\psi(r_{i,[p]}) .$$

Naturally, the random variable inside the prox is

$$Z_p \overset{\mathcal{L}}{=} \mathcal{N}\left(\beta_0(p), \frac{1}{n\nu(\tau)^2}\mathbf{E}\left(\frac{(g^{-1}_{c_\tau \lambda_i^2}(\tilde{r}_{i,(i)}) - \tilde{r}_{i,(i)})^2}{\lambda_i^2}\right)\right) .$$

Further simplification of the relationship between $\nu(\tau)$ and c_τ

Note also that when the Hessian is diagonal, the second equation, i.e

$$1 \simeq \operatorname{trace}\left(\tau c_\tau + n\nu(\tau)\mathrm{Id}\right)^{-1}$$

can again be interpreted as a prox relation: the equation reads (in the twice-differentiable case)

$$n\nu(\tau) = \sum_{i=1}^{p} \frac{1}{1 + \frac{\tau c_\tau}{\nu(\tau)n} f_k''(\widehat{\beta}_k)} .$$

Calling

$$Z_k = \left[\beta_0(k) + \frac{1}{\nu(\tau)}\frac{1}{n}\sum_{i=1}^{n} X_i(p) c_\tau \psi(r_{i,[p]})\right] ,$$

we have

$$\widehat{\beta}_k = (\mathrm{Id} + \frac{\tau c_\tau}{\nu(\tau)n} f_k')^{-1} Z_k ,$$

and therefore

$$\frac{1}{1 + \frac{\tau c_\tau}{\nu(\tau)n} f_k''(\widehat{\beta}_k)} = [\operatorname{prox}_{\frac{\tau c_\tau}{\nu(\tau)n}}(f_k)]'(Z_k) .$$

So the second equation can be rewritten as

$$n\nu(\tau) = \sum_{k=1}^{p} [\operatorname{prox}_{\frac{\tau c_\tau}{\nu(\tau)n}}(f_k)]'(Z_k) .$$

These are the heuristics we used to derive our system of equations. (Work is under way to make them rigorous.)