How design matrix structure affects optimal tuning parameter values for the LASSO

Michael Komodromos
Department of Mathematics
Imperial College London

Zoi Tsangalidou
Department of Statistics
University of Oxford

Jose Pablo Folch
Department of Mathematics
Imperial College London

Alexander Larionov
Department of Mathematics
Imperial College London

Abstract

Modern statistical inference struggles with the problem of high-dimensionality, where the number of covariates, p, is often much higher than the number of observations, n. In practice, a common solution for high-dimensionality is to use the LASSO - which needs to be tuned by a parameter λ. Typically λ is taken to be proportional to $\sigma\sqrt{\frac{\log(p)}{n}}$ irrespective of correlation of parameters. We explore how the optimal λ varies with the design matrix construction, with specific focus on correlation between covariates, ultimately explaining and experimentally verifying the work of (Hebiri and Lederer, 2013) which showed that as correlation of covariates increases the optimal λ value decreases.

1 Introduction

Least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) is a popular method for problems with a large number of parameters. In particular the LASSO is well-suited to situations where the parameter vector size p is much larger than the sample size n, a situation which commonly arises in bioinformatics, astronomy and fiancne (Johnstone and Titterington, 2009).

Within our study we will focus on normal linear regression, recalling the model specification is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \sigma\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \overset{iid}{\sim} \mathcal{N}(0, \mathbb{I}_n), \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the response variable, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the design matrix and $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the parameter vector. When $p > n$, traditional methods such as ordinary least squares (OLS) are not suitable, as $\boldsymbol{X}^\top\boldsymbol{X}$ is non-invertable, leading to an infinite number of solutions for $\boldsymbol{\beta}^*$. Under the assumption that $\boldsymbol{\beta}^*$ is sparse, i.e. for $\mathcal{S} \subset \{1, \dots, p\}, \beta_j \neq 0$ if and only if $j \in \mathcal{S}$, it is natural to consider an estimator for $\boldsymbol{\beta}^*$ that penalizes the level of sparsity, given by

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_0 \right\}, \tag{2}$$

where the ℓ_0-norm, $\|\boldsymbol{v}\|_0 := \sum_j \mathbb{1}\{v_j \neq 0\}$, counts the number of non-zero entries in $\boldsymbol{v}$ for some vector $\boldsymbol{v}$. However, solving Equation (2) is non-trivial as the problem is not convex. We therefore turn to the LASSO, a convex relaxation of Equation (2), given by

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \qquad \lambda > 0. \tag{3}$$

The LASSO has seen uses in: variable selection (i.e. choosing which components are important in a sparse parameter vector), parameter estimation. (i.e. estimating the true value of $\boldsymbol{\beta}^*$ of the data generating process) and prediction. In practical applications, a rule of thumb is to choose $\lambda \propto \sigma\sqrt{\frac{\log(p)}{n}}$. Further, as λ determines the amount of penalisation it is crucial in the performance of the LASSO (Hebiri and Lederer, 2013).

The influence of correlations in the design matrix on the prediction error has been a much-researched topic in the literature during the past decade (Bunea et al., 2007; Bickel et al., 2009; Hastie et al., 2015). Multiple types of upper bounds have been derived for the prediction error under various assumptions on the correlation structure of $\boldsymbol{X}$. Some of these bounds require uncorrelated or very highly correlated matrices, while others are applicable to any correlation structure. An exhaustive presentation of such bounds and their theoretical derivations is beyond the scope of this article, but we endeavour to present some key ideas in Section 2.

Within our study we will focus on the predictive performance of the LASSO where we will examine the size of the prediction error and the optimal value of the tuning parameter λ. We will examine these properties of the LASSO for various constructions of a deterministic design matrix $\boldsymbol{X}$, with a particular focus on the correlations between columns $\boldsymbol{X}^i, \boldsymbol{X}^j$ for $i, j \in \{1, \ldots, p\}$. In Section 3 we will present an experimental study followed by a discussion of our findings in Section 4. Finally we conclude our report in Section 5.

2 Rates

When discussing rates we consider upper bounds on the prediction error of the LASSO and how these bounds vary with increasing number of observations, n. We quantify prediction error via the mean squared error, given by

$$\text{MSE}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\lambda^{\text{LASSO}}) = \frac{1}{n}\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\lambda^{\text{LASSO}} - \boldsymbol{\beta}^*)\|_2^2 \tag{4}$$

Rate bounds provide upper bounds on the prediction accuracy of Lasso and can be expressed in terms of the tuning parameter λ or the sample size n. In the literature authors often formulate the bounds in terms of λ or n; for our purposes we consider the two cases to be equivalent as λ is commonly recommended to be proportional to $n^{-1/2}$ (Dalalyan et al., 2017).

For clarity, we will consider two types of rate bounds defined in terms of the sample size n. On the one hand, there are *fast rate bounds* which scale with $\mathcal{O}(n^{-1})$ and are almost optimal. However, fast rate bounds require restricted eigenvalues or similar assumptions and hence apply only for weakly correlated designs. On the other hand, *slow rate bounds* scale with $\mathcal{O}(n^{-1/2})$. Slow rate bounds are valid for any degree of correlations and hence are more generally applicable but are sub-optimal.

Apart from providing a theoretical non-asymptotic guarantee on the size of the LASSO prediction error, these bounds are also useful in guiding the choice of tuning parameter in practice, particularly in cases where cross-validation may not be possible.

In typical applications, a rule of thumb is to consider $\lambda \propto \sigma\sqrt{\frac{\log(p)}{n}}$, but this is valid mainly for uncorrelated designs. It has been theoretically established (Hebiri and Lederer, 2013) that correlations allow for smaller tuning parameters. This result allows for improved slow rate bounds in the correlated setting (even though fast rates are still not applicable).

2.1 Fast Rates

These bounds are only valid for weakly correlated design matrices and require the *Restricted Eigenvalue Condition (REC)* given by:

$$\phi(\bar{s}) = \min_{J_0 \subset [p]: J_o \leq \bar{s}} \min_{\Delta \neq 0: \|\Delta_{J_0^c}\|_1 \leq 3\|\Delta_{J_0}\|_1} \frac{\|\boldsymbol{X}\Delta\|_2}{\sqrt{n}\|\Delta_{J_0}\|_2} > 0 \tag{5}$$

where $\bar{s}$ plays the role of a sparsity index. Notably when the design is correlated, $\phi(\bar{s}) \approx 0$. [1] . Under the REC assumption $s \leq \bar{s}$ where $s := |\{j : (\boldsymbol{\beta}_0)_j \neq 0\}|$ we have

$$\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\lambda^{\text{LASSO}} - \boldsymbol{\beta}^*)\|_2^2 \leq \frac{n\lambda^2 \bar{s}}{\phi^2(\bar{s})} \tag{6}$$

on the set

$$\mathcal{T} := \left\{ \sup_{\boldsymbol{\beta}} \frac{\sigma|\boldsymbol{\epsilon}^\top \boldsymbol{X}\boldsymbol{\beta}|}{n\|\boldsymbol{\beta}\|_1} \leq \lambda \right\} \tag{7}$$

and we want $\mathcal{T}$ to have high probability for the bound to be useful (Hebiri and Lederer, 2013). Multiplying both sides of Equation (6) by n^{-1}, we notice the LHS corresponds to the MSE and the RHS obtains the fast rate of $\mathcal{O}(n^{-1})$.

2.2 Slow Rates

Slow rates remain valid when considering highly correlated designs. In this setting we have

$$\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\lambda^{\text{LASSO}} - \boldsymbol{\beta}^*)\|_2^2 \leq 2n\lambda\|\boldsymbol{\beta}^*\|_1 \tag{8}$$

on the set $\mathcal{T}$. Note this bound depends on $\|\boldsymbol{\beta}^*\|_1$ and does not take into account the sparsity $\bar{s}$. We can improve the bound in Equation (8) in the case where $\|\boldsymbol{\beta}^*\|_1$ is large, giving

$$\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_\lambda^{\text{LASSO}} - \boldsymbol{\beta}^*)\|_2^2 \leq 2n\lambda \min\left\{\|\boldsymbol{\beta}^*\|_1, \|(\widehat{\boldsymbol{\beta}}_\lambda^{\text{LASSO}} - \boldsymbol{\beta}^*)_{J_0}\|_1\right\} \tag{9}$$

where $J_0 \subset \{1, \ldots, p\}$ is the smallest set such that $\phi(|J_0|) > 0$.

The bounds presented in Equations (6), (8) and (9) are only useful for sufficiently large tuning parameters, such that the set $\mathcal{T}$ has high probability. However, it has been shown that high correlations lead to higher probability of $\mathcal{T}$, thus allowing for the choice of smaller tuning parameters λ. Hence, using $\lambda \propto \sigma\sqrt{n^{-1}\log(p)}$ may not suitable in the highly correlated setting (Hebiri and Lederer, 2013), as it will over-penalize parameter values even though the set $\mathcal{T}$ already has high probability.

In Section 3 we explore these result through our experiments that show in highly correlated settings, the optimal tuning parameter is smaller than what the rule of thumb would suggest.

3 Experiments and Results

The main focus of this work was to explore the behaviour of the prediction error and the value of the optimal tuning parameter for varying correlations in the design matrix, but the code was extended to vary other parameters and check the results align with intuition.

3.1 Experiment design

To generate synthetic datasets $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$ we followed the approach of Algorithm 1 in (Hebiri and Lederer, 2013) to attempt to replicate results and allow comparison of exploratory analysis to previous work. Namely, we generated $\boldsymbol{y} \in \mathbb{R}^n$ by Equation (1), with $\boldsymbol{\beta}_j^* = \mathbb{1}\{j \leq s\}$ and design matrix $\boldsymbol{X} = (\boldsymbol{x}_i)_{i=1}^n$ (effectively stacking n horizontal vectors $\boldsymbol{x}_i : \boldsymbol{x}_i \overset{iid}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma})$, $\boldsymbol{x}_i \in \mathbb{R}^p$, $\boldsymbol{\Sigma}$ has diagonal entries of 1 and off diagonal entries given by ρ).

Where we studied the effects of correlation with 'extraneous' or 'impertinent'[2] variables on the regression we followed the approach of (Hebiri and Lederer, 2013) Algorithm 2 to generate them. For each column $\boldsymbol{X}^{(j)}$ of the initial design matrix we added $p - 1$ columns sampled according

[1] Heuristically, this can be seen as follows. If $\boldsymbol{X}$ is correlated, then only a small number of its columns are linearly independent and thus it has low rank and a kernel of large dimension. Therefore, it's likely that when minimizing over Δ, we will be able to find a vector that's "close" to the kernel subspace leading to $\phi(\bar{s}) \approx 0$

[2] 'extraneous' here meaning they are not used in the calculation of the value of $\boldsymbol{y}$ in Equation (1), effectively representing an analyst being unsure of which parameters to regress on and as such adding in more and more of them.

to $\boldsymbol{X}^{(j)} + \eta \boldsymbol{N}$, where η is a correlation parameter and $\boldsymbol{N}$ is a vector sampled from a standard multivariate normal distribution - giving a resulting design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p^2}$.

To summarize, we design the regression task such that we can vary the following parameters:

- n - number of observations.
- p - number of parameters in 'original' design matrix.
- s - number of non-zero parameters in $\boldsymbol{\beta}$, i.e. the sparsity.
- σ - the standard error of the sampling noise.
- ρ - the (uniform) covariance of elements across a row.
- η - the standard error of the added term when generating 'correlated' extraneous variables. When $\eta = 0$, or not specified, no extension of the design matrix occurred.

Then, further following the method of (Hebiri and Lederer, 2013), Least Angle Regression (LARS) was used to fit the LASSO to provide an estimate $\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{LASSO}}$ across a range of λ values and for each of the $\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{LASSO}}$ values the squared error $\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{LASSO}} - \boldsymbol{\beta}^*)\|_2^2$ was calculated as a metric for prediction error (PE). The λ that gave the least prediction error, $\lambda_{\min}$, for a given configuration of the parameters was recorded along with all other λ and their respective PE. The experiment was then repeated with the given parameters 1000 times (generating new $\boldsymbol{X}, \boldsymbol{y}$ each time) and any averages of values described were taken over those iterations, e.g. $\overline{\lambda_{\min}}$ is the mean of the optimal λ values obtained and $\overline{PE}$ is the mean of the prediction errors for a given λ.

We show our results in the following subsections and then proceed to analyse them in the subsequent section. Wherever the 'Control' is mentioned it refers to design matrix according to $(n = 20, p = 40, s = 4, \sigma = 1, \rho = 0, \eta = 0)$.

3.2 Replication of (Hebiri and Lederer (2013), Table 1)

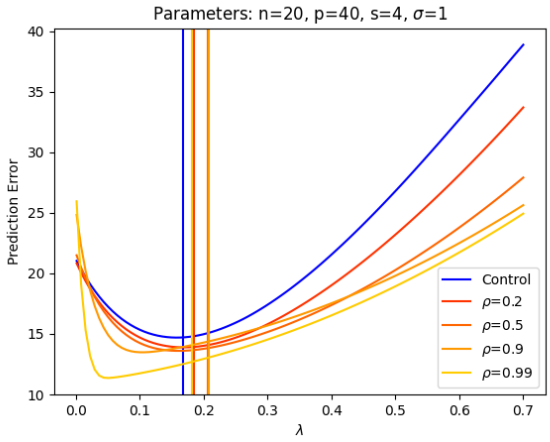
We attempted to replicate the results from (Hebiri and Lederer, 2013), however, we were not able to. One minor difference which should be noted is that their objective function does not have the scaling of $(2n)^{-1}$, which results in a different range of lambda values, but even taking this into account replication was not possible. This led us to expand our results to include the medians of the data, which qualitatively follows their results more closely. We explain our reasoning for this in the discussion section. The full results can be seen in Table 1.

n	p	s	σ	ρ	$M(\lambda_{\min})$	$\overline{\lambda_{\min}}$	$M(PE_{\min})$	$\overline{PE_{\min}}$
20	40	4	1	0.99	0.0451	0.2029	3.2315	7.8624
				0.9	0.0973	0.21792	7.5889	10.8736
				0	0.1494	0.1657	13.4606	13.8090
50	40	4	1	0.99	0.033	0.1541	3.9025	9.5894
				0.9	0.0692	0.1663	10.4720	13.8749
				0	0.1374	0.1394	13.8854	14.5964
20	400	4	1	0.99	0.0612	0.1974	3.2173	8.4664
				0.9	0.1294	0.2296	9.7687	13.7952
				0	0.1976	0.2064	15.9592	16.4541
20	40	10	1	0.99	0.0371	0.2451	15.9592	47.4660
				0.9	0.0572	0.2426	18.6319	42.9660
				0	0.0692	0.0786	20.4649	21.1991
20	40	4	3	0.99	0.1815	0.2412	5.7365	16.3672
				0.9	0.4303	0.4638	21.4506	28.4905
				0	0.9077	0.6698	51.7368	59.9019

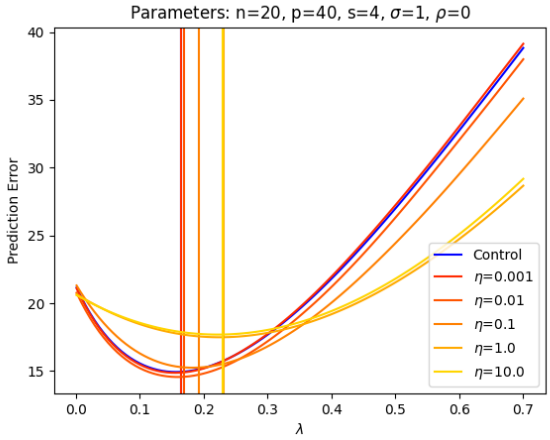
Table 1: Experimental results for different values of the parameters, where $M(\cdot)$ denotes the median. It is important to note that columns 6 and 8 were obtained in a different run of experiments to columns 7 and 9. However, the procedure and methods where identical for both cases.

3.3 Exploratory varying of design matrix structure

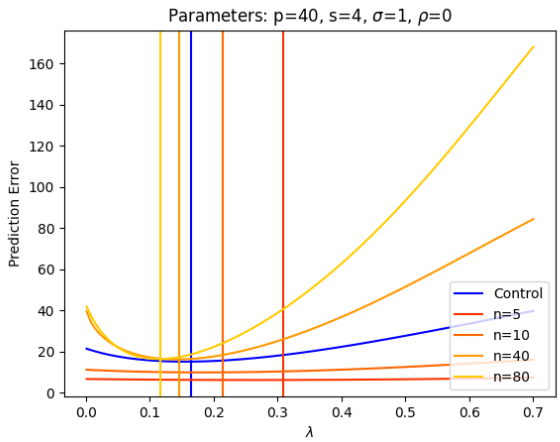
We vary the values of ρ, η, n, p, s and σ to examine the performance of the LASSO and effect on the optimal tuning parameter. Our results are presented in Figure 1.



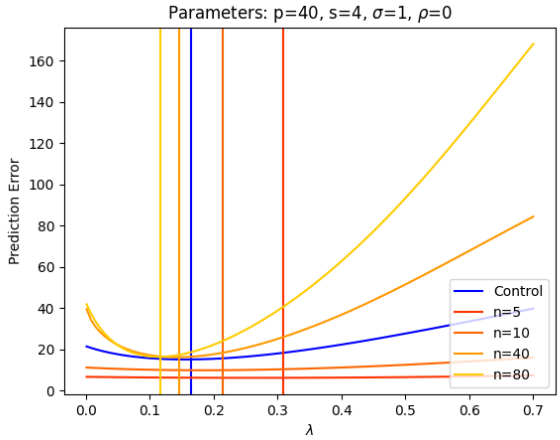
(a) ρ - Varying ρ represents how much correlation is there among our real predictors, as well as correlation with *some* extraneous parameters.



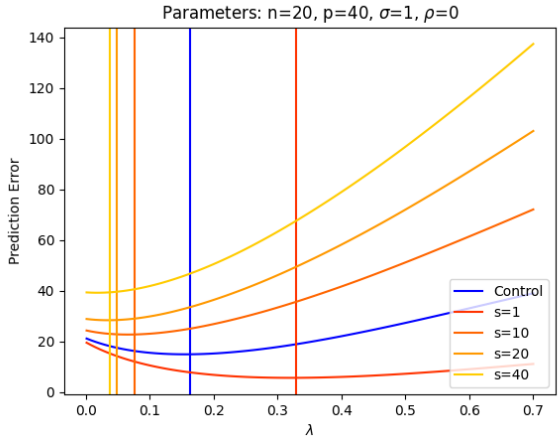
(b) η - Varying η changes how correlated the extra extraneous parameters we append. Implicitly this also controls correlation of the extra extraneous variables with η small representing large correlation and vice versa.



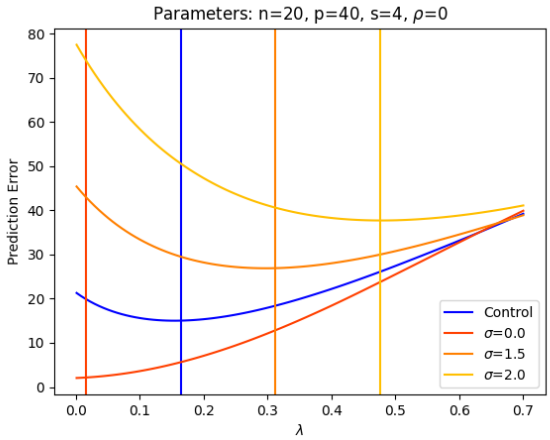
(c) n - Varying n amounts to varying the degree to which the problem is under or over specified relative to the number of parameters.



(d) p -Varying p amounts to varying both the sparsity and the degree to which the problem is under or over specified relative to the number of parameters - e.g. is there too many or too few equations for the variables we have.



(e) s - Varying s directly represents varying the sparsity of the data generating process.



(f) σ - Varying σ directly represents varying the size of the sampling noise.

Figure 1: Variation of $\overline{\lambda_{min}}$ (vertical lines) and $\overline{PE}$ (curves) with various design matrix parameters

4 Discussion

The results above can be broadly split into two sections worthy of discussion: i) the analysis of $\overline{\lambda_{\min}}$ with ρ - where there were some surprising results, and ii) the analysis with other parameters - where results followed prior research and/or intuition.

4.1 The variation of $\overline{\lambda_{\min}}$ with ρ

When beginning to experiment on this project we initially encountered surprising results. We noticed that increasing correlation, ρ, did not lead to a smaller average optimal parameter, $\overline{\lambda_{\min}}$. This contradicted the theoretical and experimental results from (Hebiri and Lederer, 2013) (as seen in Table 1). To investigate this further, we approximated the distribution of $\lambda_{\min}$ by plotting histograms of our results (see Figure 2). We discovered that as the correlation increased, the mode of the distribution shifted towards zero as expected, however, the tail of the distribution began growing, leading to the observation of extreme values which had a significant effect on the mean. As such, we decided to explore the results using an alternate summary statistic: the median. The change of summary statistic allowed us to recover the expected behaviour from $\lambda_{\min}$; a graphical representation of this phenomenon can be seen in Figure 3. We cannot be sure as to why (Hebiri and Lederer, 2013) did not come across this problem, however, we would have liked to explore this further if given more time.

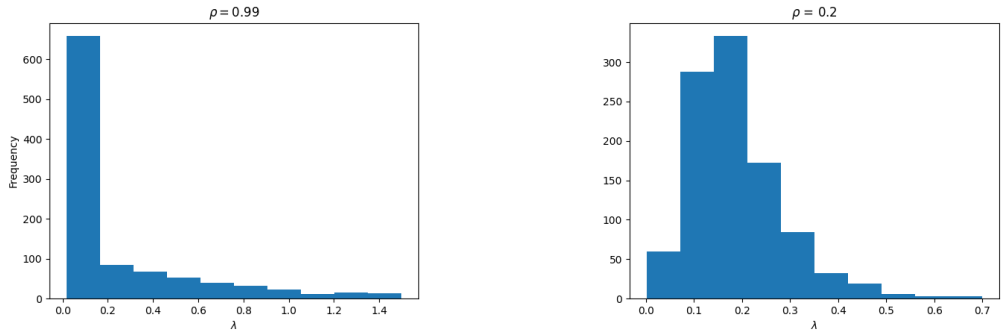


Figure 2: Histograms to estimate the distributions of $\lambda_{\min}$. We vary ρ and fix the rest of the parameters to the control case. We can see that as we increase the correlation, the mode shifts to the left but the tail of the distribution gets heavier.

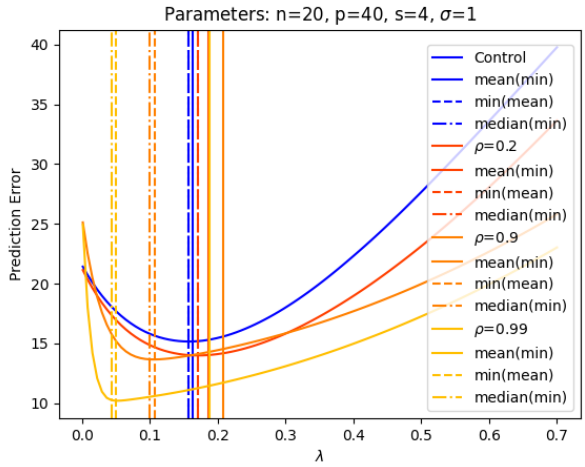


Figure 3: Graphical representation of different summary statistics. We plot the mean of the optimal lambdas (label given by mean(min)) and their median (label given by median(min)), for different correlation values. We can see how the means do not decrease with correlation. We further include the minimizer of the average prediction error, which we can see also follows the expected behaviour closely (we could not, however, make sense of this value statistically).

4.2 The variation of $\overline{\lambda_{\min}}$ with other parameters

Figure 1b shows that increasing η, i.e. decreasing correlation between the original columns and those added in Algorithm 2, increases the optimal value of λ. Furthermore, there appears to be a floor/ceiling on the optimal λ that likely corresponds to using the traditional optimal λ for the original (p columns) and extended (p^2 columns) matrices respectively. This indicates that we do indeed recover traditional recommendations of λ when there is no correlation and that for higher correlation we tend towards a λ that is linked to the effective rank of the design matrix.

Figure 1c shows the intuitive results that as n increases $\lambda_{\min}$ decreases - this corresponds to needing less regularisation as we get more equations for the given number of variables. In the regime where $n \geq p$, we could also just use OLS.

Figure 1d shows the result that as p increases $\lambda_{\min}$ increases. Qualitatively increasing p with fixed s and n not only makes the problem more sparse but also increases the ratio p/n resulting in a design matrix with a larger gap between the row and column rank values compared to the original problem. Thus a larger regularization parameters λ is needed to control $\|\boldsymbol{\beta}\|_1$. In fact, looking back at the theoretical results of Section 2, we expect the optimal tuning parameter to scale as $\sqrt{\log(p)}$. This is consistent with what we observe in Figure 1d, in which the gap between the optimal λ's is larger between $p = 10$ and $p = 50$ (increase proportional to $\sqrt{\log(5)}$) than between $p = 50$ and $p = 100$ (increase proportional to $\sqrt{\log(2)}$).

Figure 1e shows that higher sparsity values correspond to smaller optimal tuning parameters. This is a reasonable observation, because the LASSO is well-suited to dealing with sparse problems. A higher value of the tuning parameter λ imposes a higher penalty on the l_1 norm, thus favouring solutions $\widehat{\boldsymbol{\beta}}_{\lambda}^{\text{LASSO}}$ that are more sparse. Hence, when $\boldsymbol{\beta}^*$ is not sparse (e.g. $p = 30$ or $p = 40$), the optimal tuning parameter will be smaller as to allow for more non-zero entries.

Figure 1f shows that increasing σ corresponds to larger optimal tuning parameters. This is reasonable, because we expect that obtaining a good prediction while increasing the observation noise, would require higher penalisation of the model parameters (so that the model fits the data rather than the observation noise).

5 Conclusion

To conclude, we have replicated the qualitative results of (Hebiri and Lederer, 2013) that for greater parameter correlation we should expect lower values of optimal λ and verified intuition and existing recommendations about how λ choice should vary with various other changes to the design matrix structure.

However, we could not replicate the quantitative portion of the paper; both having different scaling (likely from different optimisation functions in the LARS packages used) and observing a phenomenon that resulted in drift in $\overline{\lambda_{\min}}$ such that we needed to analyse the median of $\lambda_{\min}$ to get qualitative agreement.

Further work could include higher resolution study of variance of optimal λs with various parameters to attempt to derive some more accurate rules of thumb for optimal tuning parameter selection. More interestingly, further study could also be done to quantitatively investigate and understand why the ℓ_2 prediction error seems to flatten in the region of $\lambda_{\min}$ for high correlations thus causing the heavy tails observed in Figure 2.

Code

`R` and `Python` implementations can be found at `https://github.com/mkomod/lasso-bounds`. Noting the `R` implementation uses Iterated Weighted Least Squares over LARS when fitting the LASSO but similar observations were found.

References

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Stat.*, 37(4): 1705–1732, 2009. ISSN 00905364.

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1(0):169–194, 2007.

A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552–581, 2017.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity*. Chapman and Hall/CRC, May 2015.

M. Hebiri and J. Lederer. How correlations influence lasso prediction. *IEEE Trans. Inf. Theory*, 59(3):1846–1854, 2013.

I. M. Johnstone and D. M. Titterington. Statistical challenges of high-dimensional data. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, 367(1906):4237–4253, 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, 58(1):267–288, 1996.

Y. Wu and L. Wang. A survey of tuning parameter selection for high-dimensional regression. *Annu. Rev. Stat. Its Appl.*, 7:209–226, 2020.