# Bayesian Lasso Critique

Shawn Unger
Michal Malyska
Alin Morariu

Tuesday 26th March, 2019

**Abstract**

This paper serves as a critical analysis of the Bayesian Lasso proposed by Trevor Park and George Casella in 2008. We examine some of the theoretical shortcomings of the concept which give rise to high variance parameter estimates. The Bayesian approach is then tested on simulated data sets and we try to impose a hard thresholding rule in order to achieve an optimal level of shrinkage (only include relevant covariates in the model). All simulations were completed in Stan.

# Contents

# 1 Introduction

Introduced in 1996 by Robert Tibshirani, the Lasso (Least Absolute Shrinkage and Selection Operator) is an extension on the classical linear regression model which includes a penalization term. It was motivated by the desire to prioritize simplicity and interpretability of the model while maintaining a similar or better prediction accuracy when compared to its classical cousin. The additional term in the loss function (shown in 1) effectively "shrinks" regression parameters (pushes them towards zero).

$$L(x, y) = (y - X\boldsymbol{\beta})^T(y - X\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

The Bayesian Lasso, proposed by Park and Casella in 2008 establishes the parameter estimation under a Bayesian framework by implementing a hierarchical structure which ultimately places independent Laplace priors on the regression parameters. The paper provides interval estimates for the Lasso parameter via both empirical Bayes, by Marginal Maximum Likelihood, and a diffuse hyperprior ([2]).

The purpose of this paper is to investigate the performance of the above mentioned methods through a simulation based approach. We provide a brief analysis and a theoretical exploration of the priors, and it's connection to the traditional Lasso. We continue by testing the model's performance on several simulated data sets (from various distributions) and comparing it to the traditional Lasso. Lastly, using the simulated data sets which have a known number of regression parameters plus additional noise covariates, we attempt to improve performance by adjusting the hard thresholding by which we zero out the $\beta_j$'s. In other words, the key question being answered is under what conditions does the Bayesian Lasso perform just as well or better when compared to conventional methods.

# 2 Methodology

## 2.1 Theoretical Consideration - Critique of Park and Casella

### 2.1.1 Connection to the Traditional Lasso

A key aspect in defining the Bayesian Lasso is ensuring consistency between it and the least squares Lasso. In order to make this connection clear (and later highlight a shortcoming of the Bayesian Lasso), it is important to understand the geometry of problem (for illustration purposes, discussion of the geometry will be limited to $\mathbb{R}^2$ but the concept can be easily generalized). Because of the normal distribution the independent random variable in a regression follows, one can draw it's contours on the plane with the mean of the distribution (equivalently, the centre) being $\hat{\beta}_{OLS}$. The Lasso introduces a penalization term which states that $||\boldsymbol{\beta}||_1 < c$ for a predefined constant $c$. This region forms a diamond shape

around the origin and the solution to the Lasso problem is simply the first intersection of contours of the normal distribution and set diamond (see 1).

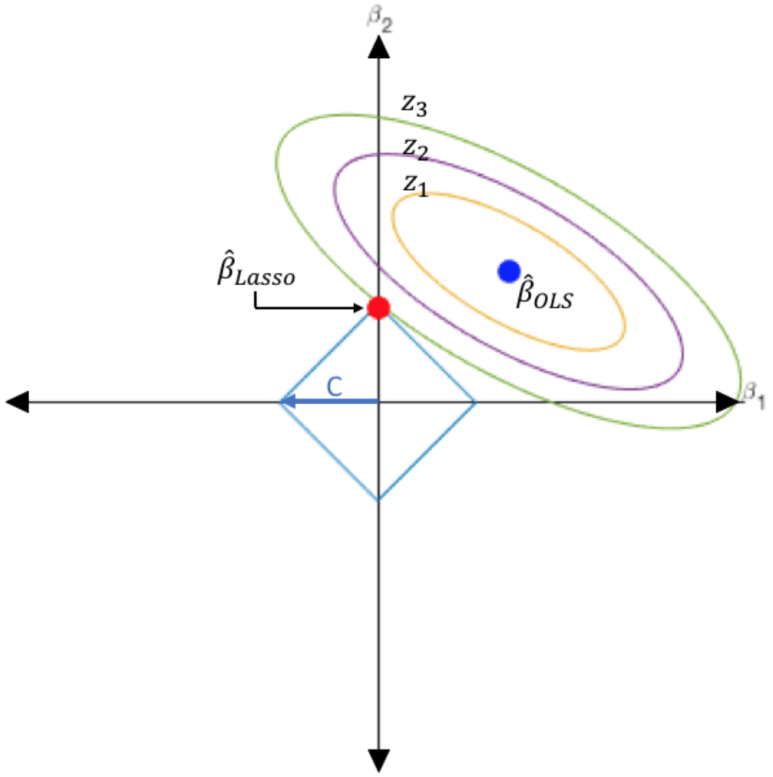


Figure 1: Geometric interpretation of traditional least squares Lasso regression parameter estimation [1]

The essence of this is captured in the Laplace priors placed on the independent, identically distributed (iid) $\beta_j$. It can shown that the least square solution pictured above is analogous to the Maximum A Posteriori (MAP) estimator of the Bayesian Lasso (proof detailed in Appendix A). More notably, this approach provides us with more than one point estimate because we have a full distribution for the values of $\boldsymbol{\beta}$ where we can take other measures of centrality for our point estimates (eg. mean or median of the posterior). This is *precisely* the problem which arises when using the Bayesian Lasso. We explore this further in the next section.

### 2.1.2 Parameter Estimation from a Posterior Distribution

Once the posterior distribution of the parameters is established, fitting the model is simply a simulation exercise. Using a sampling technique such as the Markov Chain Monte Carlo (MCMC) or a Gibbs sampler (a wider variety exists and can be implemented), we generate a large sample from which we derive our point estimate and credible intervals for the parameters. That is all to say that the quality of your estimate depends on the sample you have drawn (specifically, how representative the sample is of the true posterior). Unlike the least squares approach where you have a deterministic system ($\boldsymbol{\beta}$ is found by matrix multiplication and is entirely determined by the data), the Bayesian Lasso determines the values of the parameters stochastically via simulation. The data is used as a mean of refining the posterior distribution in order to find the most representative measure possible, but estimates are nevertheless varying. The implication is that we can never be sure that any estimate under the Bayesian Lasso will indeed be the red point shown in Figure 1. But instead is representative of some region around that point with the equality only holding as the sample size approaches infinity. The variability of the estimates can result in some unwanted and unidentifiable consequences; namely in that the results can over-shrink or under-shrink $\beta_j$. Over-shrinkage can occur in the case where the posterior places too much weight near zero while under-shrinkage in the case where too little weight is placed near zero.

### 2.1.3 Selection of Shrinkage Rate - $\lambda$

A very important consideration in the Lasso is the value which the shrinkage rate $\lambda$ in equation 1 takes. Typically, this value is chosen via cross-validation, generalized cross-validation, and ideas based on Stein's unbiased risk estimate ([2]). In the Bayesian approach, this procedure is captured by the Laplace priors placed on the regression coefficients. As shown in Appendix A, the change of variables $\lambda = \frac{\sigma^2}{\tau}$, the shrinkage hyperparameter of the Lasso regression is inversely proportional to the scale parameter of the Laplace distribution. The implication here is that the wider we set our prior (large $\tau$), the smaller the resulting shrinkage rate which is indicative of the our prior knowledge of the prior. In cases where a diffuse prior is required, minimal shrinkage occurs meaning that the Lasso will have limited advantages over regular regression.

An alternative to allowing $\lambda$ to be determined by the spread of the Laplace priors is to estimate compute its maximum marginal likelihood with the use of the EM algorithm. As before, the shrinkage factor will be inversely proportional to the spread in the priors but this time it'll be with respect to the expectations due to the empirical nature of our estimator. Park and Casella provide a closed form estimator for the shrinkage factor in equation 2 with the starting point in equation 3 depending on the ordinary least squares regression parameters.

$$\lambda^{(k)} = \sqrt{\frac{2p}{\sum_{j=1}^{p} \mathbb{E}_{\lambda^{(k-1)}}(\tau_j^2|y)}} = \sqrt{\frac{2}{\mathbb{E}_{\lambda^{k-1}}(\tau^2|y)}} \Bigg|_{\tau_1=\ldots=\tau_j=\tau} \tag{2}$$

$$\lambda^{(0)} = \frac{p\sqrt{\hat{\sigma}^2_{LS}}}{\sum_{j=1}^{p} |\beta_j|} \tag{3}$$

Using the maximum marginal likelihood on $\lambda$ means that we are relying on the data to tell us how much shrinking we should do. Typical data set contain no information on which covariates are important so allowing it to determine the shrinkage rate does not provide any meaningful insight. Furthermore, the EM algorithm will not converge to the to MLE, but drift around the true value (have convergence as $n \to \infty$) with the drift becoming smaller with additional iterations (i.e. we can get a good *enough* estimate that will likely not shrink an additional $\beta_j$ under the zeroing threshold.

Alternative methods such as using a sensible (should only allow for positive values of $\lambda$) hyperprior are also explored. In this case, the shrinkage parameter is estimated using a sampling technique which gives the best fit to the data. However, the additional level in our model has a downfall. At each level, we have compounding variance which is unfortunately not identifiable. That is to say we are increasing the amount of variability in our prediction but have no insight into whether it is caused by the data or our limited knowledge of the model parameters. Nevertheless, this method is preferred to the maximum marginal likelihood method approach since we can produce a credible interval for the parameter which allows us to test various values and compare the respective performances.

## 2.2 Working with Sparsity

### 2.2.1 Defining Sparsity

Lasso regression is particularly helpful in eliminating covariates which do not provide additional information to the model. As such, the goal is always to have a model which is sparse in it's $\beta_j$'s. When defining sparsity, it is clear that it must be contingent on the dimension of the vector we are working with. Below are three definitions which we will use to later enforce sparsity on the model along with a brief justification.

- Less than square root of total entries are non-zero (strong sparsity)

$$\#\beta_j \neq 0 \leq \sqrt{d} \tag{4}$$

- Less than half of total entries are non-zero (weak sparsity)

$$\#\beta_j \neq 0 \leq \frac{d}{2} \tag{5}$$

- Less than a constant number of entries are non-zero (this choice will be particularly useful when exploring problems in which we have a significant amount of prior knowledge on or want/need to ensure the model contains very few parameters)

$$\#\beta_j \neq 0 \leq c \tag{6}$$

### 2.2.2 Enforcing Sparsity

Since the posterior distribution of regression coefficient does not lead to them being exactly equal to 0 to improve the performance of the Bayesian Lasso, we are proposing a hard thresholding scheme for zeroing out the regression coefficients by allowing the cut off point, $\epsilon$, to be a function of the number of parameters in the data set and the number of points in the data set. The goal is to have an adaptive way of selecting the cutoff in order to improve the performance of model (more on this in section 2.3). Using the following way of setting a $\beta_j$ to 0, we investigate possible dependencies of the threshold on the number of parameters in and the number of data points. We also showcase that there seems to be very little dependence of the value of $\epsilon$ on the smallest real regression coefficient $\beta_{j_{min}}$ that would not normally be known.

- Linear scaling with dimension of model parameter vector

$$\frac{d\beta_i}{||\boldsymbol{\beta}||_2} < \epsilon \rightarrow \beta_i \mapsto 0$$

- non-linear (expected to perform better since it has higher resolution for low values as the dimension of the model increases)

$$\sqrt[d]{\frac{\beta_i}{||\boldsymbol{\beta}||_2}} < \epsilon \rightarrow \beta_i \mapsto 0$$

For both of these measures, we aim to find the threshold that results in achieving one of the three definitions of sparsity discussed above. More over, we would like to include only the true/relevant $\beta_j$'s in our model. To that extent we will introduce a new variable in our simulation: the ok

## 2.3 Performance Evaluation Measures

In order to evaluate the performance of our models and the various proposed adjustments we will use the L norm to and compare our methods to the least square method estimates and the true values (known from simulation). They are defined as follows

- Unnormalized L norm

$$M_{L,p}(\beta, \hat{\beta}) = \left( \sum_{i=1}^{p} |\beta_i - \hat{\beta}_i|^L \right)^{\frac{1}{L}} \tag{7}$$

- Normalized L norm

$$M_{L,p}(\beta, \hat{\beta}) = \left( \sum_{i=1}^{p} \left| \frac{\beta_i - \hat{\beta}_i}{\beta_i} \right|^L \right)^{\frac{1}{L}} \tag{8}$$

- Maximal normalized deviance

$$M_{L,p}(\beta, \hat{\beta}) = \max_{i \in \{1,\ldots,p\}} \left| \frac{\beta_i - \hat{\beta}_i}{\beta_i} \right|^L \tag{9}$$

The set of various deviance measurements allows us to analyze the performance of the Bayesian Lassos using different assumptions. Furthermore, by considering the deviance as a function of dimension, we can have a more in-depth look a how the error term changes with the change of dimensions used in the simulations. Nevertheless, the goal of the first Deviance measurement defined, the normalized L norm (7), allows us to look at the overall deviance between the estimated beta values versus their true value. The next deviance measurement, the normalized L norm (8), allows normalizing by the true value of the beta. This primarily assists with considering the deviation of the estimation but controlling for the bias of having larger true beta values on the model miss-specification. Finally, the last measure, maximal normalized deviance (9), is a measure that allows considering the maximal miss specified beta value when controlling for the beta value. This will allow us to consider the potential of least difference between true and estimated beta values when using the model.

# 3 Results

Due to a limited compute we were only able to perform 128 simulations. We fixed the Laplace prior spread parameter that determines the shrinkage rate, at 10. Thus $\lambda = 0.1$ for all runs. The simulation was planned to include a range of values, however as is, we ran significantly past our limit (26 hours of computing). The current simulation go over a range of values for:

- Number of real regression coefficients (m) - 1, 5, 10, 50
- Number of noise regression coefficients (k) - 1,10,100
- Values of the real regression coefficients (b) - 0.1,0.5,1,5
- Dataset Size (n) - 10,100,1000

From there we have produced a number of plots of the different definitions of epsilon for different sparsity regimes and made some simple observations.

## 3.1 Plots of cutoff $\epsilon$

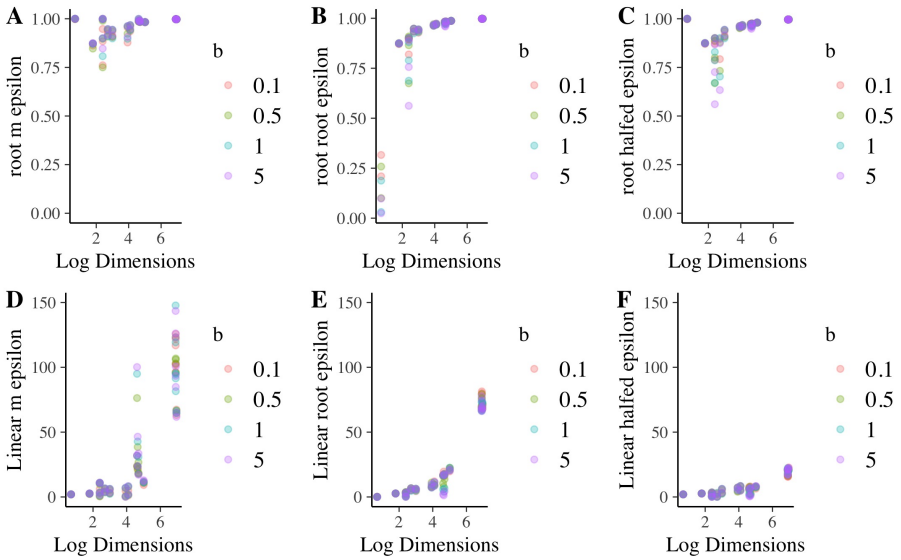


Figure 2: Log D vs Epsilon values

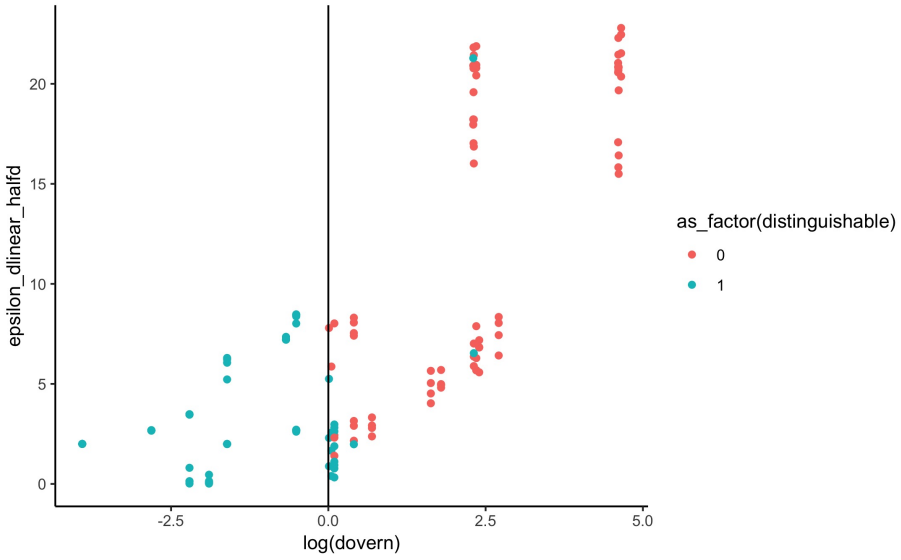


Figure 3: D over N with Epsilon Linear in half of D

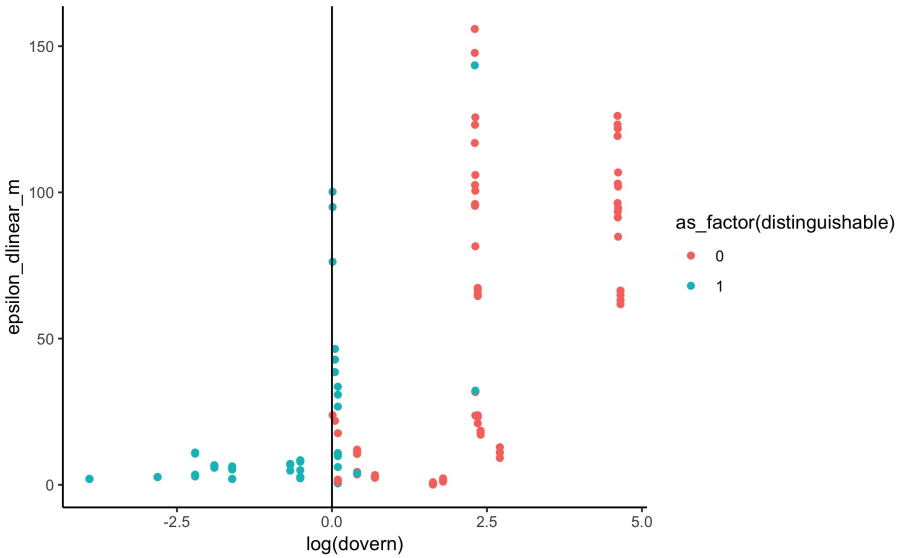


Figure 4: D over N with Epsilon Linear in M

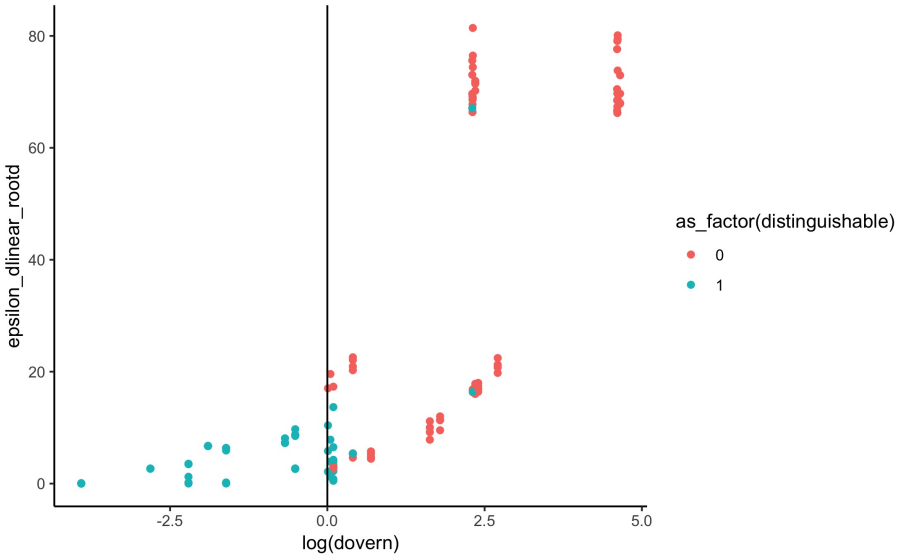


Figure 5: D over N with Epsilon Linear in Root D

The first thing that sticks out was that it was always possible to find a hard threshold that separates all the real regression coefficients from the noise coefficients whenever the number of data points was larger than the number of coefficients. Of course, this requires prior knowledge of the number of real coefficients, however, it means that at least the magnitudes are well ordered.

We continue by analyzing the results attained form plotting the log Dimensions verse various measures of deviance in 6. The first thing that is easily noticeable and expected, is that the deviance increases with dimensions whether it is calculated using the normalized or unnormalzied deviance. The more interesting result, is the maximal deviance also increases with dimensions, however, it seems to hover around the same values. Meaning that though the maximal deviance does increase with dimensions, it does not increase substantially. Nevertheless, we then can use 7 to analyze the relationship between the normalized L norm deviations, as dimension increases, i.e. dividing deviation by the dimensions of the L norm deviations. When analyzing these figures it seems that rather then a noticeable large difference in deviation with increasing with dimensions, there is a more subtle difference even though one can still expect a large deviation with larger dimensions.

### 3.2 Posterior Distributions of Regression Parameters

A visual analysis of the posterior distributions on each of the $\beta_J$

## 4 Discussion and Conclusion

TO DO...

# Appendices

## A Connection between Bayesian and Least Squares Lasso

We place a Laplace prior (centered around 0) on each regression parameter.

$$p(\beta_j|\tau) = \frac{1}{2\tau}\exp\left(-\frac{|\beta_j|}{\tau}\right) \tag{10}$$

Since the priors are iid, we can easily get the joint prior on $\boldsymbol{\beta}$.

$$p(\boldsymbol{\beta}|\tau) \propto \exp\left(-\frac{1}{2\tau}\sum_{j=1}^{p}|\beta_j|\right) \tag{11}$$

Additionally, we are working under the assumption that our data follows a normal distribution.

$$y \sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n)$$

With the accompanying likelihood

$$f(y|X, \boldsymbol{\beta}, \sigma^2) = (\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}(y - XB)^T(y - XB)\right) \tag{12}$$

Multiplying A and 12, along with some simplifications yields a term in the exponential which is similar to the least square Lasso.

$$f(y|X, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}|\tau) \propto (\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}(y - XB)^T(y - XB)\right)\exp\left(-\frac{1}{2\tau}\sum_{j=1}^{p}|\beta_j|\right)$$

$$\propto \exp\left(-\frac{1}{2\sigma^2}(y - XB)^T(y - XB) - \frac{1}{2\tau}\sum_{j=1}^{p}|\beta_j|\right)$$

Applying the $-2\log(\cdot)$ transform to the above

$$-2\log(f(y|X, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}|\tau)) \propto \frac{1}{\sigma^2}\left[(y - XB)^T(y - XB) + \frac{\sigma^2}{\tau}\sum_{j=1}^{p}|\beta_j|\right]$$

Performing a small change of variables where $\lambda = \frac{\sigma^2}{\tau}$ we get the form

$$-2\log(f(y|X, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}|\tau)) \propto (y - XB)^T(y - XB) + \lambda\sum_{j=1}^{p}|\beta_j| \tag{13}$$

Notice that is simply the loss function of the least squares Lasso regression. More importantly, the MAP under the Bayesian approach is analogous to the solution of constraint optimization of the loss function of the least squares Lasso regression. This proves that the mode of the posterior distribution on $\boldsymbol{\beta}$ is identical to $\hat{\boldsymbol{\beta}}_{Lasso}$.

# B Visualization of Simulation Results

More to come later...

## B.1 Plots of Posterior on $\beta$

## B.2 Other

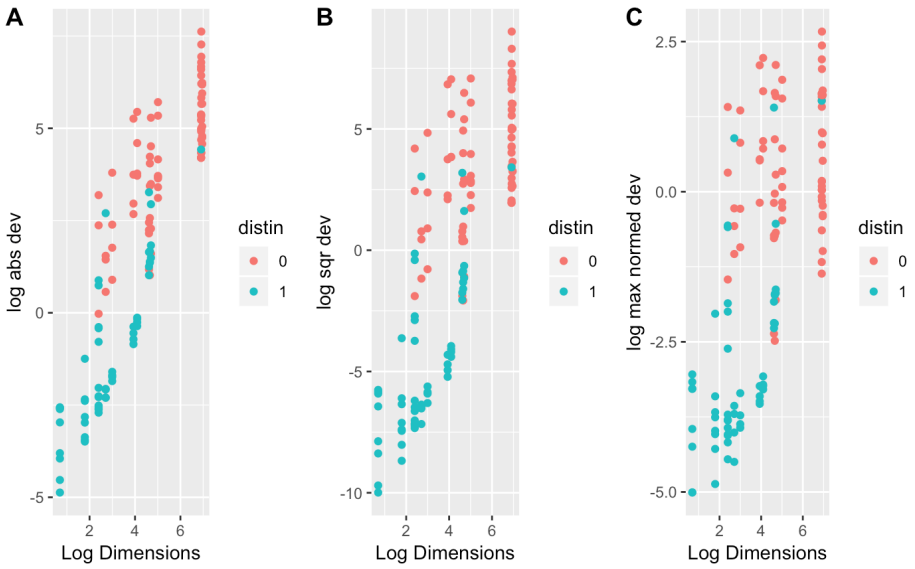


Figure 6: Log D vs Measures of Deviance

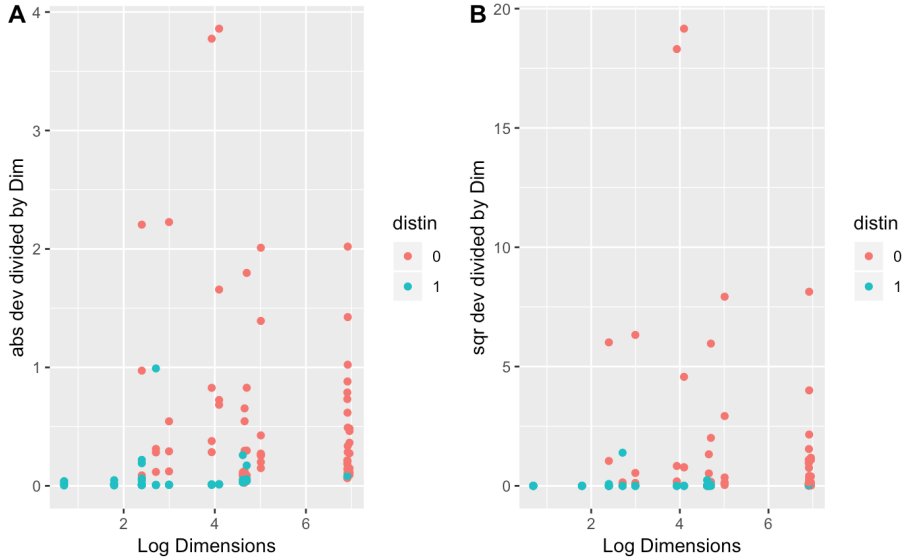


Figure 7: Log D vs Measures of Deviance divided by Dimension

# References


[1] David Gold. *Dealing With Multicollinearity: A Brief Overview and Introduction to Tolerant Methods*. [Online; accessed March 21, 2019]. 2017. URL: https://waterprogramming.wordpress.com/2017/02/22/dealing-with-multicollinearity-a-brief-overview-and-introduction-to-tolerant-methods/.

[2] Trevor Park and George Casella. "The Bayesian Lasso". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 681–686. DOI: 10.1198/016214508000000337. eprint: https://doi.org/10.1198/016214508000000337. URL: https://doi.org/10.1198/016214508000000337.