STA302 - Lecture 9

Cedric Beaulac

June 11, 2019

Introduction

Today's plan

- Today is the last lecture of the semester! We will :
  - Do a quick review of our model, overfitting and large date set issue
  - Introduce Principal Component Analysis
  - Introduce Lasso and Ridge regression
- Ridge and Lasso and modern techniques, but not too modern! They are adapted to new challenges but made their proof. They are reliable but really advanced topics. I do not expect you to become Redge and Lasso experts but rather understand the motiviation for those techniques and learn enough about them to get you started.

Review

Review of the model

- We have established a basic model that allow p predictors x_j to influence our prediction for a given response y.
- $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, where $\mathbf{X}$ is $n \times (p+1)$ matrix of observed predictor values, $\mathbf{y}$ a $n \times 1$ vector of response values, β the $(p_1) \times 1$ vector of parameters and $e \sim MVN(0, \sigma^2 I)$.
- We estimate β with $\hat{\beta}_{MLE} = (\mathbf{X}^T\mathbf{X})^1\mathbf{X}^T\mathbf{y}$.
- It implies $\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$.

Review of large data set issues : large p

- **It reduces interpretation.** Occam's Razor states that among several plausible explanations for a phenomenon, the simplest is best. Even though Occam's razor's principal is debatable, everyone agrees that it si easier to explain a simpler model and that in orderto get the big picture, we are willing to sacrifice small details.
- **It increases the variance of the estimates.** Having more parameters to estimate increases the variace of the estimators. This is reflected in $s^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-(p+1))$.

Review of large data set issues : large p

- **It is more prone to overfitting.** The more parameter, the more complexe the model can be which increases the chances that we fit *too much* the data set to the detriment of generalization abilities.
- **It increases the chances of collinearity issues.** Collinearity is caused by having too many variables poviding similar information. The more predictors you have in the model the higher the chances are that some provide similar information.

Review of overfitting

- We say a model overfits if it has too good performances on the training set to the detriment of test set performances.
- This is truly undesireable because we often have to deal with a small data set but the purpopse of a model is to understand a more global relationship.
- We wish to understand this relationship to solve new problems, for prediction purposes and for insightful decision making.
- A model that overfits can not help us with these problems.

Principal Component Analysis (PCA)

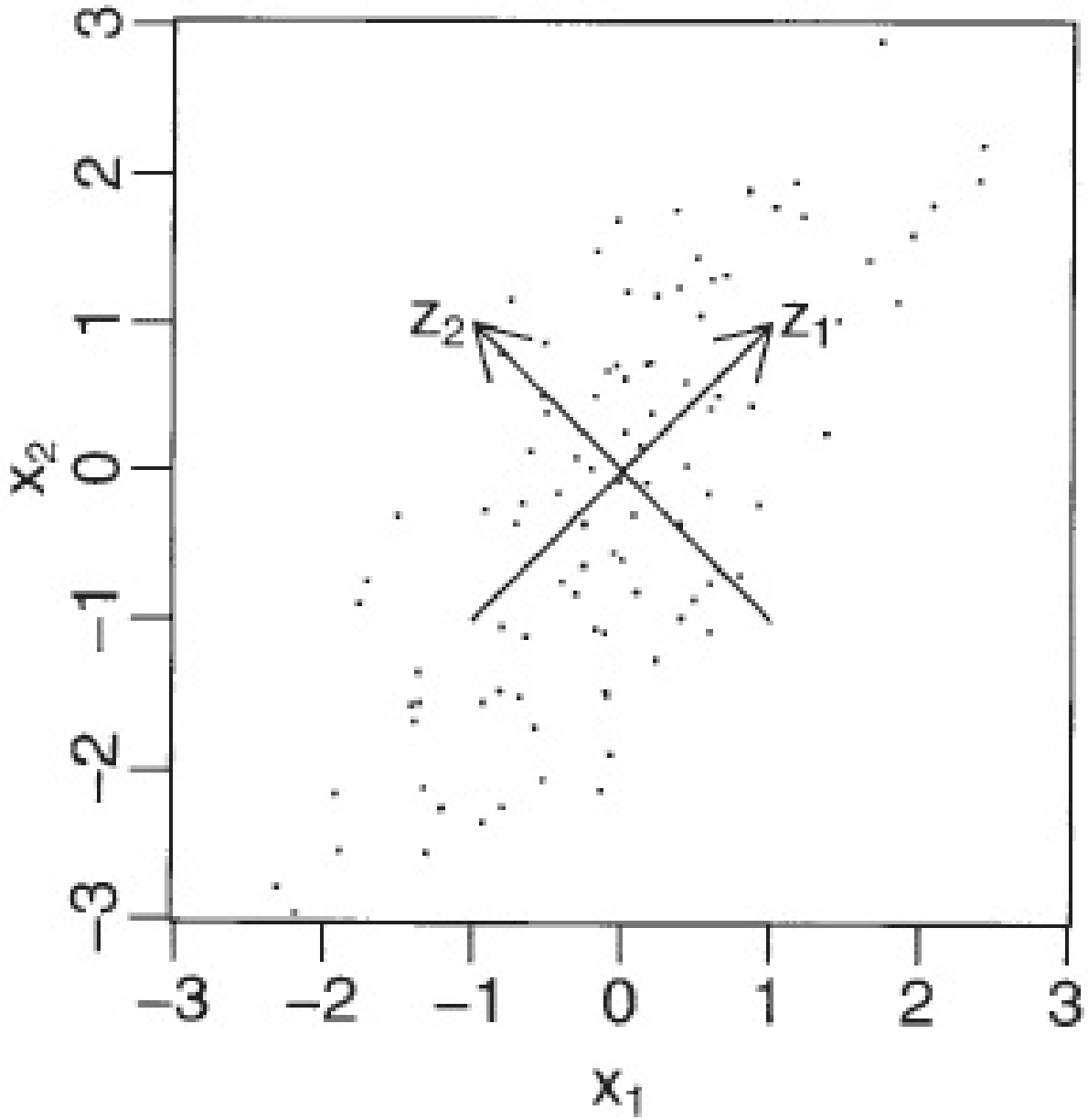
Principal Component Analysis (PCA) : Motivation

- Last week we introduce basic techniques to reduces the number of parameters in a fitted model.
- All of them rellied on taking some variables out of the model.
- Principal components are a reparametrization of the current system in order to create uncorrolated predictors.
- More precisely, we project the predictors onto an orthogonal space.
- Most times we try to project the predictors on a lower dimensionnal space that preserves as much variability as possible.

Principal Component Analysis (PCA) : Motivation

- This way we completely solve the collinearity problems of any predictor set.
- We also ensure our variables are rich in information and thus we can obtain a model with a reduce number of variable to keep the variance low and chances of overfitting.
- It solves 3 out of the 4 problems related to large p we mentionned. But it does make interepretation even harder.
- This transformation is extremely simple and fast.
- It has MANY other purposes from unsupervised learning to data compression.
- I don't want you to become PCA expert but rather to understand the concept and hear about it once before getting into it next year.

Principal Component Analysis (PCA) : Intuition

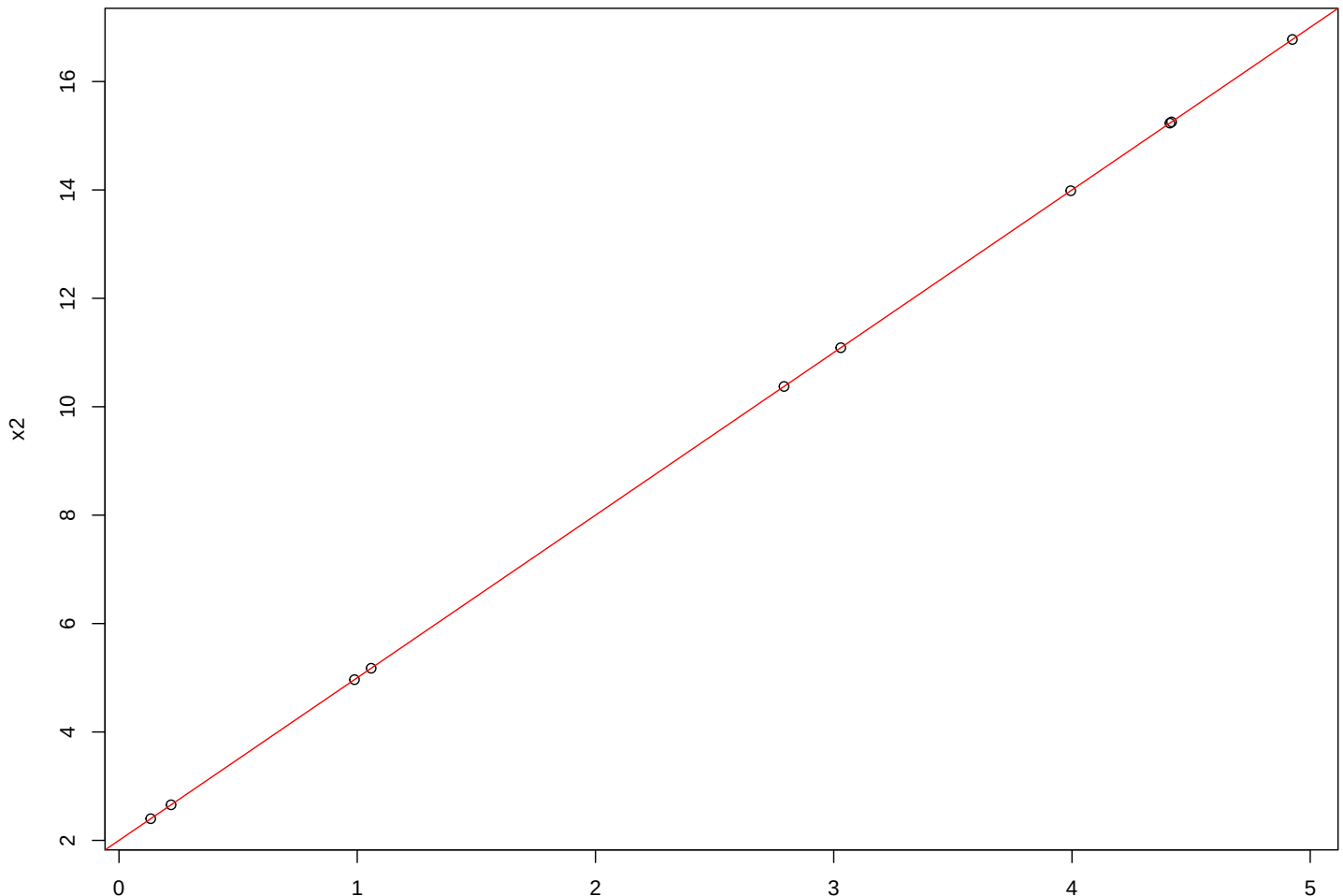


Principal Component Analysis (PCA) : A matrix algebra problem

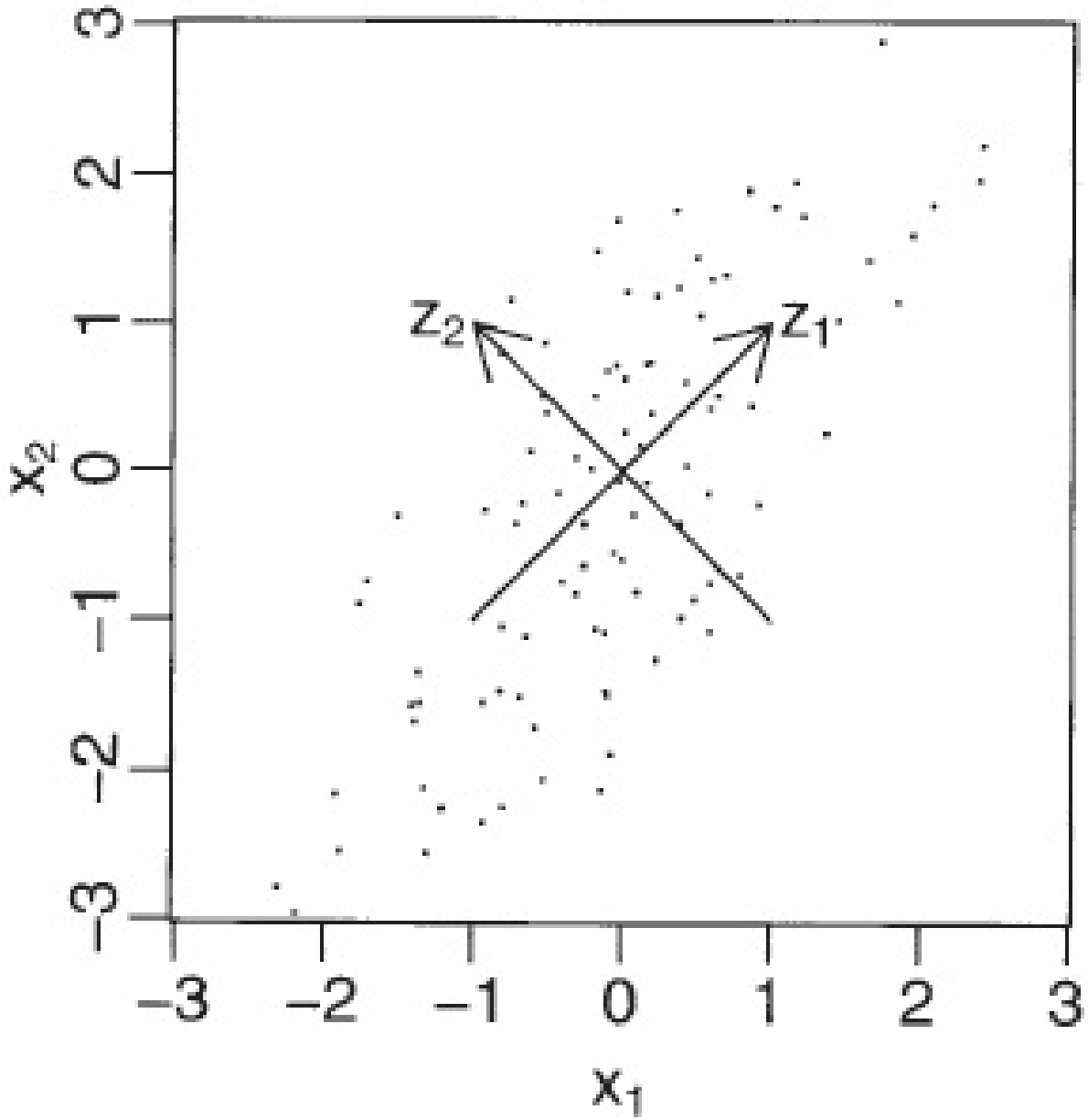
- There exist infinitely many 2-dimensional orthogonal space we could project the predictors on.
- If we were to reduce the dimensionnality (to one) ideally we would like to do a projection that keeps observations as distinguishable as possible.
- We say we want to preserve the variability in the data and the way to proceed is to project the predictors onto the axis that maximize the variance of the new vectors.

Principal Component Analysis (PCA) : Intuition

- If the predictors are perfectly corrolated it is easy to perceived hwo we can reduce the dimensionnality without loosing information.



Principal Component Analysis (PCA) : Intuition



Principal Component Analysis (PCA) : A matrix algebra problem

- Defining $\mathbf{S}$ as the observed covariance matrix :

$$\mathbf{S} = \begin{bmatrix} \sum_{i=1}^{n}(x_{i,1} - \bar{x}_1)^2 & \dots & \sum_{i=1}^{n}(x_{i,1} - \bar{x}_1)(x_{i,p} - \bar{x}_p) \\ . & \dots & . \\ \sum_{i=1}^{n}(x_{i,p} - \bar{x}_p)(x_{i,1} - \bar{x}_1) & \dots & \sum_{i=1}^{n}(x_{i,p} - \bar{x}_p)^2 \end{bmatrix}$$

- Notice that this is $\mathbf{X}^T\mathbf{X}$ after we've removed predictors means.

Principal Component Analysis (PCA) : A matrix algebra problem

- With $\mathbf{z} = \mathbf{X}\mathbf{u}$ where $\mathbf{Z}$ is our lower dimension space $m(1) < p$ and $\mathbf{U}_{p\times 1}$ is the projection vector.
- The variance of the projected observation $Var(\mathbf{z}) = \mathbf{u}^T\mathbf{S}\mathbf{u}$.
- We want $\mathbf{u}$ to be a directioon in the original predictor space, so let's define $\mathbf{u}$ as a vector of norm 1 : $\mathbf{u}^t\mathbf{u} = 1$.
- This leads to a maximization problem with a constraint that we take care of using a Lagrange multiplier :

$$\mathbf{u}^T\mathbf{S}\mathbf{u} + \lambda(1 - \mathbf{u}^t\mathbf{u})$$

Principal Component Analysis (PCA) : A matrix algebra problem

- Once we optimize we obtain : $\mathbf{Su} = \lambda \mathbf{u}$.
- This implies λ is a eigenvalue of $\mathbf{S}$ and $\mathbf{u}$ an eigenvector of $\mathbf{S}$.
- Finally if we left-multiply by $\mathbf{u}^T$ we get $\mathbf{u}^T\mathbf{Su} = \lambda$ (because $\mathbf{u}^t\mathbf{u} = 1$) and thus λ is the variance of the projected data.
- To maximize the variance, we select $\mathbf{u}$ as the eigenvector associated with the largest eigenvalue.

Principal Component Analysis (PCA)

- To generalize this process, suppose we have p predictors. We can project the predictor matrix $\mathbf{X}$ on a lower dimension orthogonal space $\mathbf{Z}$ of size $m < p$ using a projection matrix $\mathbf{U}_{p \times m}$.
- $\mathbf{Z}_{n \times m} = \mathbf{X}_{n \times p} \mathbf{U}_{p \times m}$
- The matrix consist of the eigenvectors associated with the m largest eigenvalues of the data correlation matrix $\mathbf{S}$.
- Then we can fit a linear model using $\mathbf{Z}$: $\mathbf{y} = \mathbf{Z}\beta + \mathbf{e}$.
- We have lower number of predictors and they are all uncorrolated.

Principal Component Analysis (PCA)

- On the flip side, all of the predictors are extremely hard to interpret, for example $z_i = u_{i,1}x_1 + u_{i,2}x_2 + ... + u_{i,p}x_p$.
- This is an extremely short introduction to PCA who deserves a couple of full lectures, because it is amazing. (Least sqaure reconstruction error solution as well!)
- I simply want you to undertsand what's good and bad about PCA.
- These tools will be defined more rigorously in latter courses.

Ridge and Lasso Regression

Ridge regression : Motivation

- We could prevent overfitting by controlling the size of the parameters.
- $\bar{y}$ does not *overfit*. If $\beta_i = 0$ for $i \in \{1, ..., p\}$, then $\bar{y} = \beta_0$.
- Intuitively, if the β's are small they can only affect the y so much and it minimizes the chances of overfitting.
- Techniques that reduce the size of the parameters are known as shrinkage techniques and ridge regression is one of those technique.

Ridge regression

- To establish what the ridge regression is, we have to go back to lecture 3 and re-defined the regression problem as a prediction error minimization problem.
- Recall that $\widehat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ is the vector of predicted value of the response for the predictor matrix $\mathbf{X}$.
- The observed error is $\widehat{\mathbf{e}} = (\mathbf{y} - \mathbf{X}\hat{\beta})$.
- The sum of square error is $(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$.
- One way to established $\hat{\beta}$ is to find the parameters that minimizes the square errors.

Ridge regression

- Our parameter are the solution of minimization of :

$$(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$$

- One way to force small $\hat{\beta}$ would be to also minimize $\sum_{i=1}^{n} \beta_i^2 = \hat{\beta}^T\hat{\beta}$.

Ridge regression

- Thus we can establish $\hat{\beta}_{\text{ridge}}$ as the solution of the minimization of :

$$(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda\hat{\beta}^T\hat{\beta},$$

where λ is an hyper-parameter controlling the penalty on $\sum_{i=1}^{n} \beta_i^2$.

- If $\lambda = 0$ then we have our regular $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^1\mathbf{X}^T\mathbf{y}$ and as $\lambda \to \infty$ then all of β's goes to 0 and we have a model where we predict using $\bar{y}$.

Ridge regression

- Let's minimize :

$$(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda\hat{\beta}^T\hat{\beta}$$

$$\frac{d}{d\hat{\beta}}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda\hat{\beta}^T\hat{\beta} = 0$$

$$-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\beta} + 2\lambda\hat{\beta} = 0$$

$$2\mathbf{X}^T\mathbf{y} = 2\mathbf{X}^T\mathbf{X}\hat{\beta} + 2\lambda\hat{\beta}$$

$$\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X} + \lambda I)\hat{\beta}$$

$$(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y} = \hat{\beta}$$

Ridge regression

$$\Rightarrow \hat{\beta}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

- We have a closed-form (analytic) solution due to the convexity of the penalty term which is easy to implement.
- It protects against overfitting by shrinking the reducing the effect of the parameters.
- It also prevent collinearity issues since $\mathbf{X}^T\mathbf{X} + \lambda I$ is always invertible and λI ensure the determinant is not too small. This was the motivation of the Ridge creators.
- So we love Ridge.

Ridge regression : A constrained optimization problem

- The ridge regression problem can be expressed as a constrained optimization problem. In fact :

$$\hat{\beta}_{\text{ridge}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} (y_i - \underbrace{(\beta_0 + (\sum_{j=1}^{p} \beta_j x_{i,j}))}_{\hat{y}_i})^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t$$

- Just know that there exist a one-to-one correspondence between λ and t, and that both of the formulations lead to the same solution.

Ridge regression

- *But how do we decide on the penalty term λ ?* Great question!
- Here's how : We establish a huge list of possible λs. Then we try them all and we select the λ that produces the best results on the test set.
- *But I though we shouldn't use the test set for training ?* That's true, let's partition the data differently.
- We will now partition the data set into three pieces : A training set, a validation set and a test set.
- The training set is used to established the estimates $\hat{\beta}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$, the validation set is used to select λ and we can finally use new observations (test set) to test our model.

Ridge regression

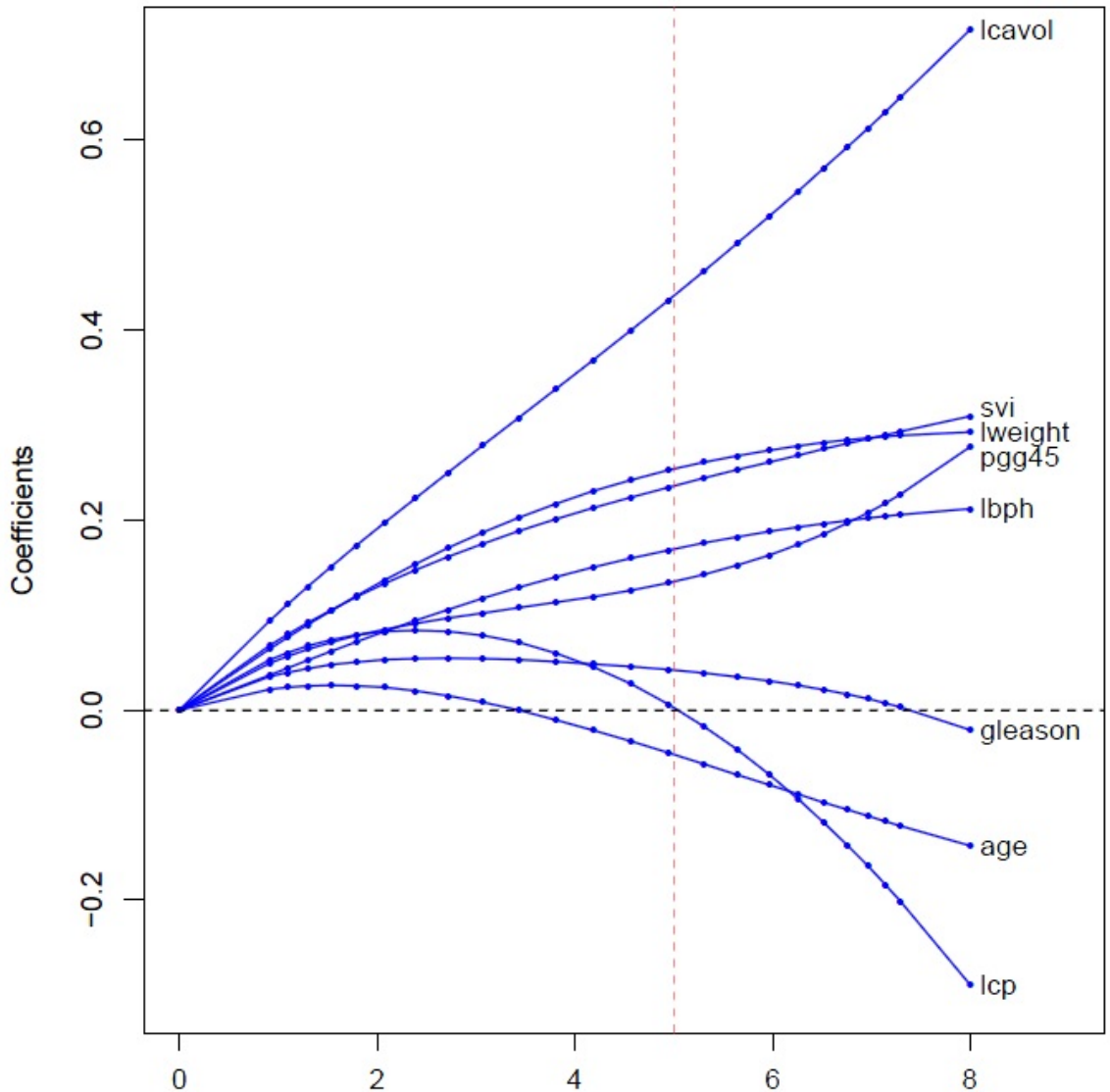


Figure 4: As t increases the parameters converges to MSE estimates.

Ridge regression : Conclusion

- A simple modification to the objective function (function we optimize) leads the Ridge regression estimates.
- The modification is in fact a penalty on the sum of squared parameters value.
- It shrinks the parameters to lower values.
- The result is a model that naturally prevents overfitting and collinearity problem.
- The model does NOT reduce the number of parameters, thus does not improve interpretability.

Lasso regression

Lasso regression : Motivation

- Can we actually enforce *sparsity* using a penalty on the parameters.
- When we talk about sparsity we actually refer the ability to fit model that needs a reduced amount of parameters.
- Lasso is a simple modification to the the penalty term that has huge implication.
- It leads to a model that naturally eliminate and shrink parameters effectively solving all of the problems related to large number of parameters.
- The new penalty being non-convex, the problem is now much more complicated to solve and we cannot compute the exact solution.

Lasso regression

- Once again we cast our penalty term as part of the optimization problem.
- This time we want to minimizes :

$$\sum_{i=1}^{n}(y_i - \underbrace{(\beta_0 + (\sum_{j=1}^{p}\beta_j x_{i,j}))}_{\hat{y}_i})^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

$$= (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|$$

- Which is exactly like Ridge but now we use the $\mathcal{L}_1$ norm instead of $\mathcal{L}_2$ for the penatly.

Ridge regression : A constrained optimization problem

- Once again, this problem can be expressed as a constrained optimization problem. In fact :

$$\hat{\beta}_{\text{lasso}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} (y_i - \underbrace{(\beta_0 + (\sum_{j=1}^{p} \beta_j x_{i,j}))}_{\hat{y}_i})^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t$$

- Once again there exist a one-to-one correspondence between λ and t, and that both of the formulations lead to the same solution.

Lasso regression : The problem

- This new penalty term makes the solutions nonlinear and there is no closed form espression for it as in ridge regression.
- Computing the lasso solution is a quadratic programming problem.
- BUT efficient algorithms are available for computing the entire path of solutions as λ varies with the same computational cost as for ridge regression (Thank you Tibshirani and Hastie!)

Lasso regression : The amazing result

- Because of the nature of the constraint, making t small will actually cause some parameters to be exactly zero.
- Thus the lasso does a kind of continuous subset selection naturally.
- When comparing this to Ridge, as t gets smaller and smaller all the parameter shrink in parallele while in Lasso some goes to 0 before others.

Lasso regression : The amazing result

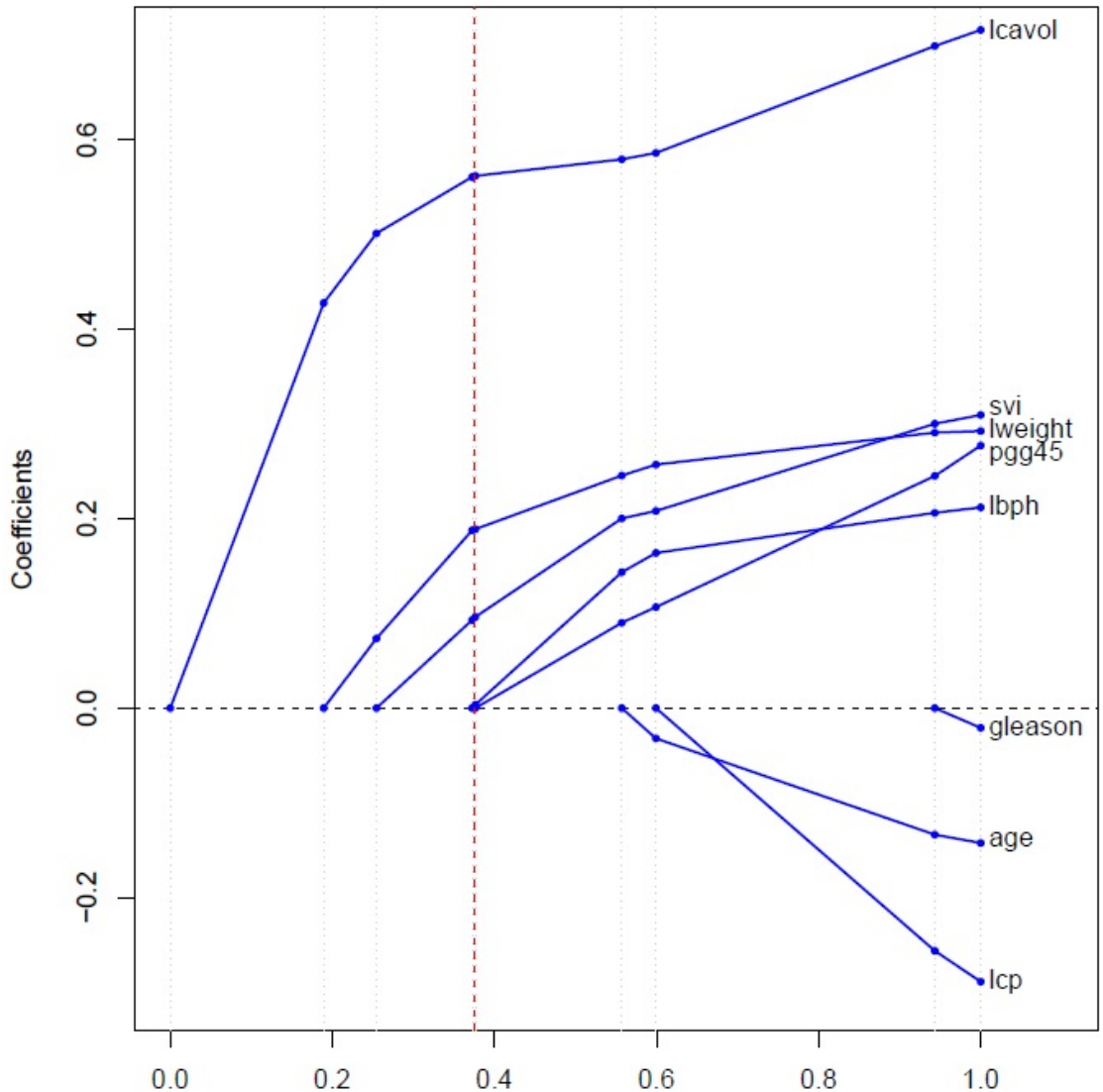
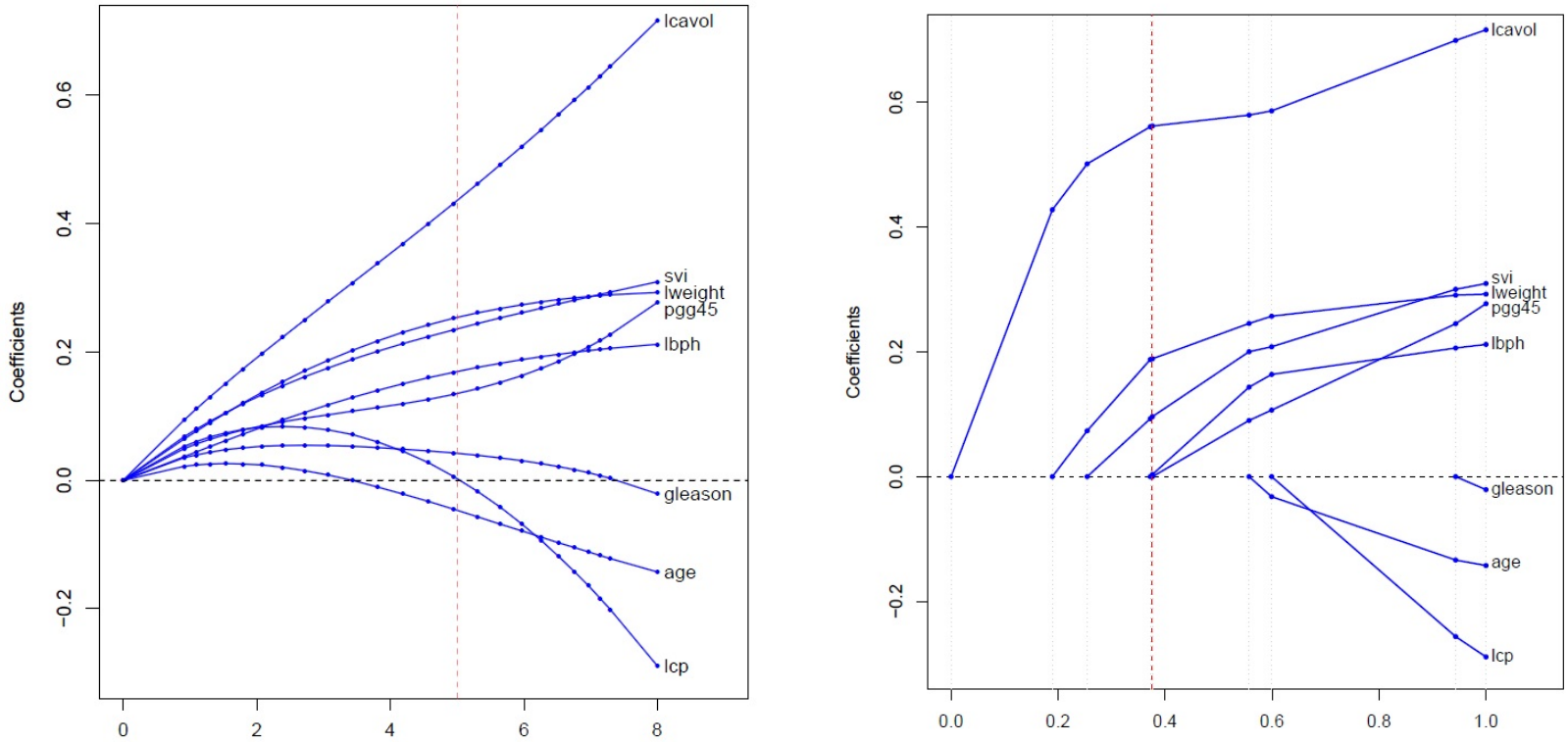


Figure 5: As the t increases the parameters converges to MSE estimates.

Lasso vs Ridge



(a) Ridge **(b)** Lasso

Figure 6: Difference in Ridge and Lasso of parameter values as t decreases

Lasso vs Ridge : but why ?

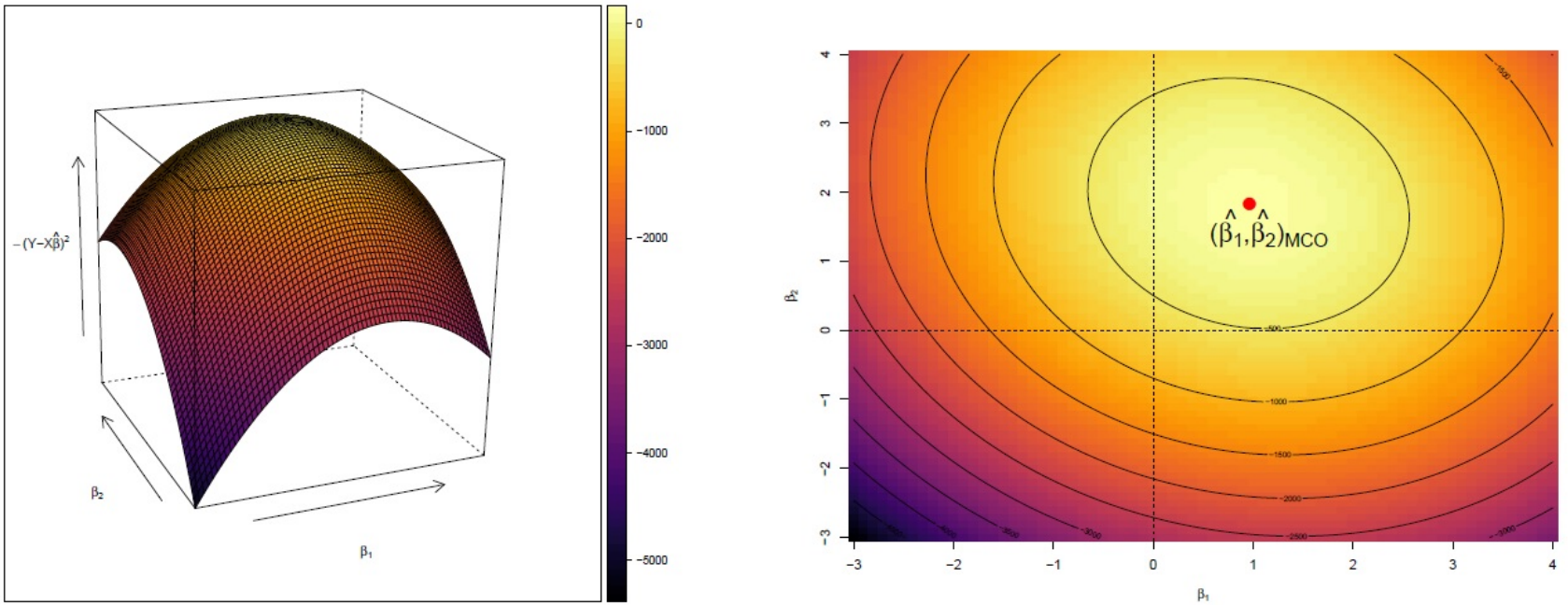


Figure 7: Optimization surface. (Credit to Sahir Rai Bhatnagar)

Lasso vs Ridge : but why ?

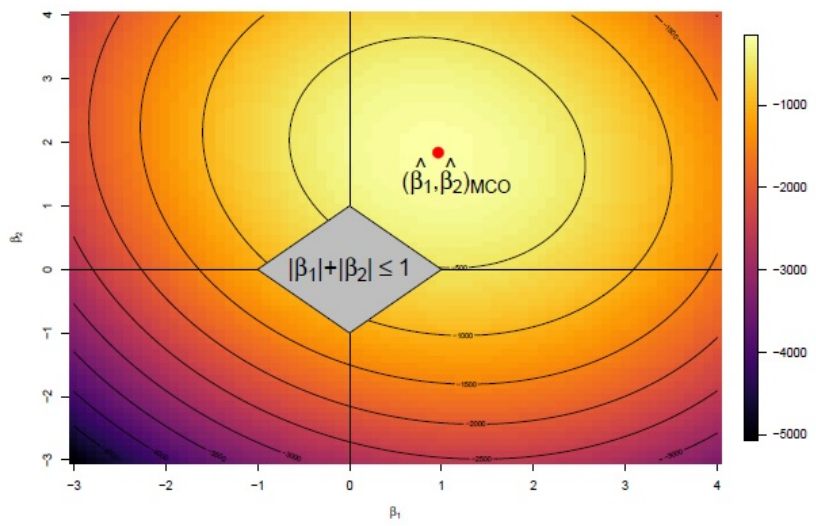


Fig. 1: `lasso`

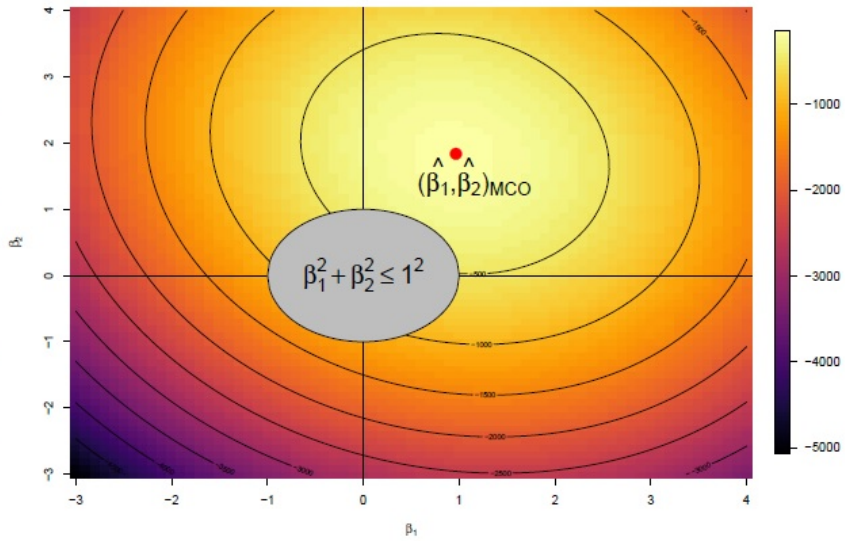


Fig. 2: `ridge`

Figure 8: Different constraints lead to different contours shape. (Credit to Sahir Rai Bhatnagar)

Lasso vs Ridge : but why ?

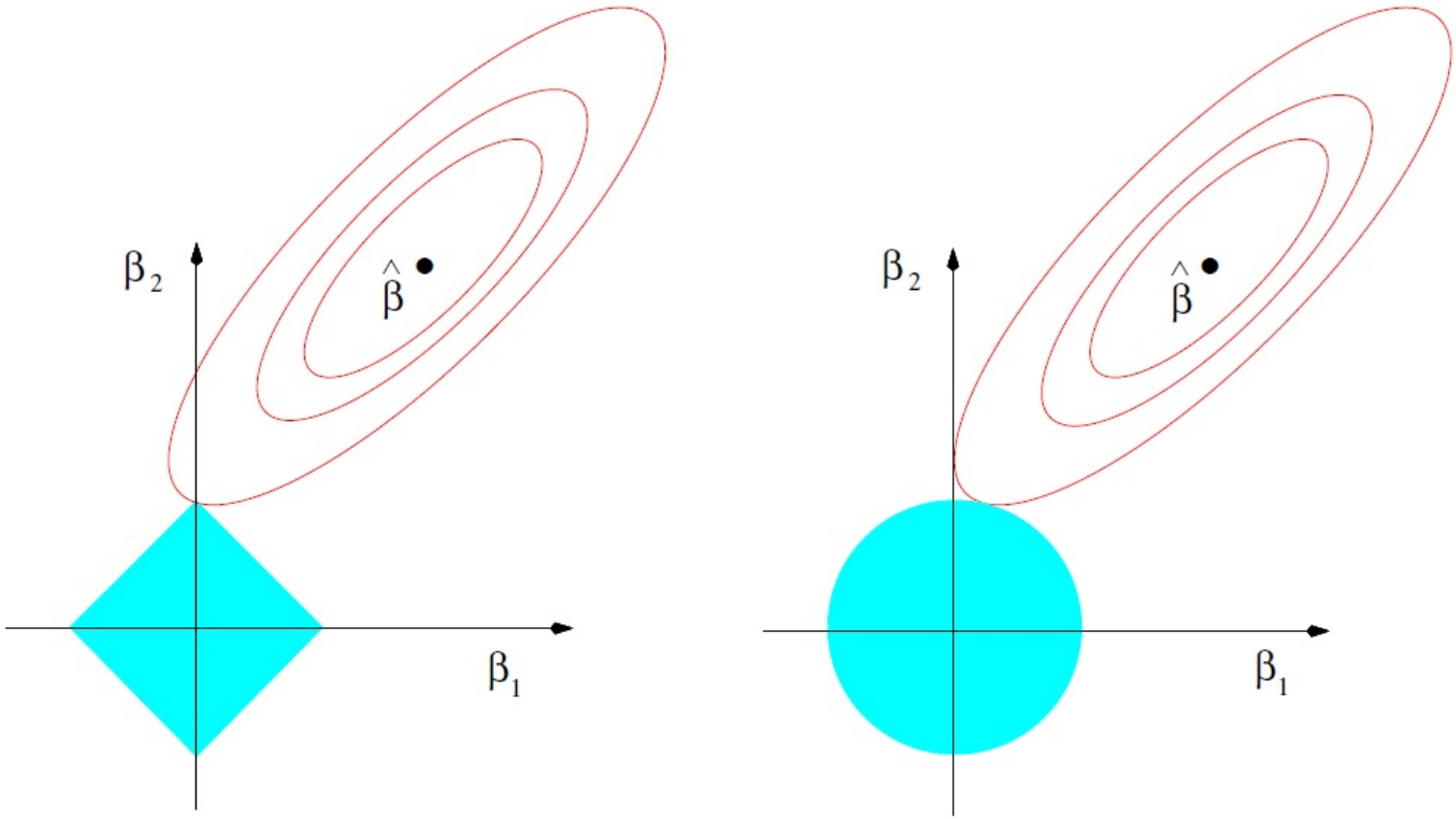


Figure 9: Different constraints lead to different contours shape.

Lasso vs Ridge : Now I'm just stealing pictures from books

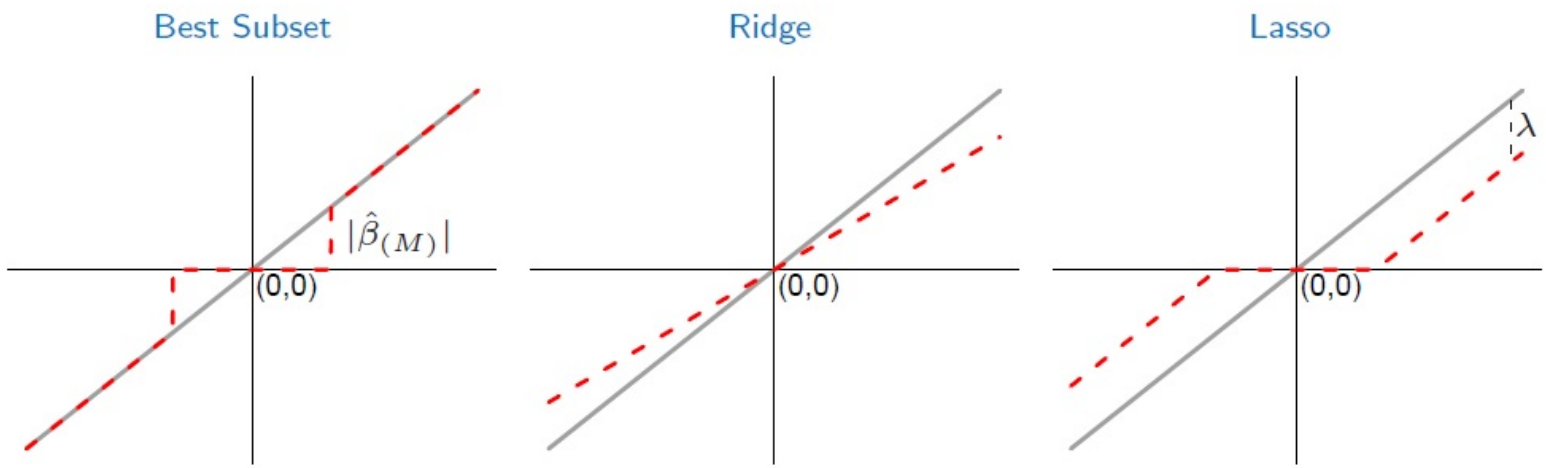


Figure 10: Different shinkage effect

Elastic net

- Ridge and Lasso regression are extremly similar in nature, they are both trying to find the least square minimizers while respecting a constraint.
- The contraint can be expressed as $\sum_{j=1}^{p} |\beta_j|^q$ where $q = 1$ for Lasso and $q = 2$ for Ridge.
- *What if we were to try other values of q ?* Well again it would change the contours of the constraint surface.

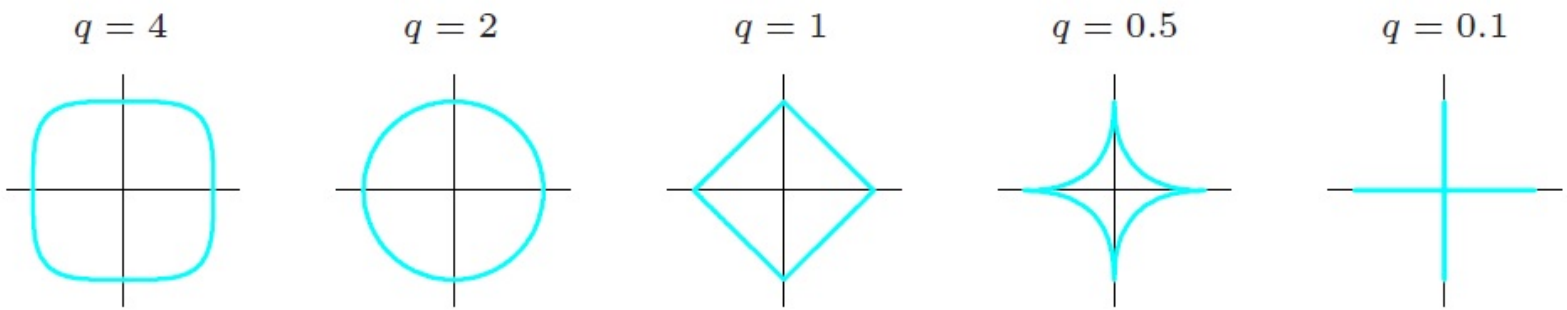


Figure 11: Different values of q leads to difference contours.

Elastic net

- Of course, we are not going to code the otpimal and efficient algorithms to get these solutions.
- The GLMnet and elastic net packages in R(Trevor Hastie,Hui Zou and Rob Tibshrianni) are freely available and let you fit the solution for various q and t, use cross-validation to get optimal value of these hyper-parameters and more.
- We'll use it in the R-lab.

Variable selection : Post-selection inference

- Post-selection inference is still a problem when using these techniques.
- In fact there is not even a distribution in the model yet.
- But Rob Tibshrianni worked on Post-selection inference from 2010 to 2015 and came up with way to compute the distribution of the parameter conditional on the shrinkage method used.
- selectiveInference package in R (available since 2017) allow to do valid inference for parameters that were selected using Elastic net.

Conclusion

- PCA is simple to implement and solve most of the issues related with large parameters but also has other uses outside linear regression.
- Ridge and Lasso are modifications to the typical optimization problem for Linear Models.
- It leads to techniques that reduces the chances of overfitting and ensure low collinearity.
- Lasso also forces sparsity which in fact seamlessly do variable selection.
- A solution to the post-seletion inference was proposed for the variable selection technique.
- Most of these progress are due to Trevor Hastie and Rob Tibshrianni.
- These are advanced techniques, I just hope you got curious and when they'll be introduced again in fourth year, you'll be more confortable thant other students.

Practice Problems

- WAY TOO HARD PROBLEMS : The elements of statistical learning ch.3 : 5,6 and 7.

External Sources

- The elements of statistical learning ch.3 Click here

Conclusion

- It's been a pleasure being your course instructor.
- Good luck with the final!