NYC DATA SCIENCE
ACADEMY

The Curse of Dimensionality

Data Science Bootcamp

Outline

- Part 1: Taking a New Perspective
- Part 2: Dimension Reduction
- Part 3: Vectors of Highest Variance
- Part 4: The PCA Procedure
- Part 5: Alternatives to PCA: Ridge & Lasso Regression
- Part 6: Cross-Validation
- Part 7: Review

PART 1

Taking a New Perspective

Taking a New Perspective

“Perspective is everything when you are experiencing the challenges of life.”

-Joni Eareckson Tada

Taking a New Perspective

A riddle...

Taking a New Perspective

How many didn’t?

Taking a New Perspective

Let’s try that again...

Taking a New Perspective

There are 30 cows in a field,

and 20 ate chickens.

How many didn't?

PART 2

Dimension Reduction

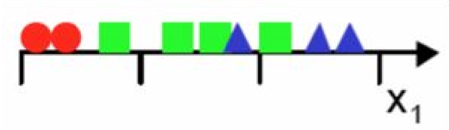
When can High Dimensionality be Adverse?

- Sparsity becomes exponentially worse as the dimensionality of our data increases.

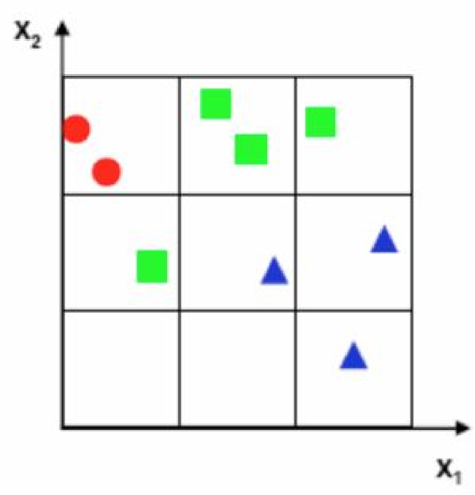
- Given a number of observations, additional dimensions spread the points out further and further from each other.

- There tends to be insufficient repetition in various regions of the high-dimensional space. Less repetition makes inference more difficult:
  - Are the results replicable?
  - What about regions that don't have any observations at all?

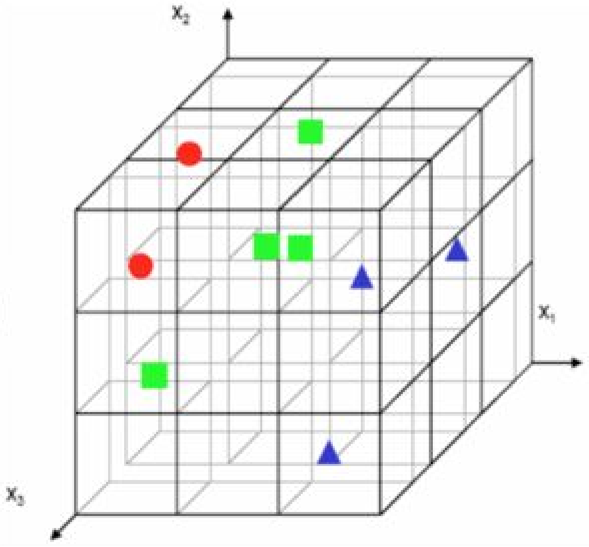
When can High Dimensionality be Adverse?



9 observations
3 sections



9 observations
9 sections



9 observations
27 sections

When can High Dimensionality be Adverse?

- Collecting data is expensive, both monetarily and temporally.
  - You might be working on a budget; the collection of additional variables may not be necessary and could hinder the return on your investment.

- There is too much complexity with higher-order data.
  - Often we not only seek the most accurate solution, but also one that is simple and interpretable.

- We may have redundancy in our measured dimensions.
  - While all the variables in our dataset might be different from one another, the information they contain as a group may overlap.

When can High Dimensionality be Adverse?

- Consider measuring the following on all students at a university:
  - Hours of sleep.
  - Hours of partying.
  - Hours in the library.
  - Number of tests taken.
  - Number of enrolled classes.
  - Number of meals eaten.
  - Amount of physical activity.
  - Amount of printer usage.

- While all these variables measure different things, they might be interrelated; is there a common factor that may help inform measurements on all of these variables? GPA?

Don’t Just Throw Away Data!

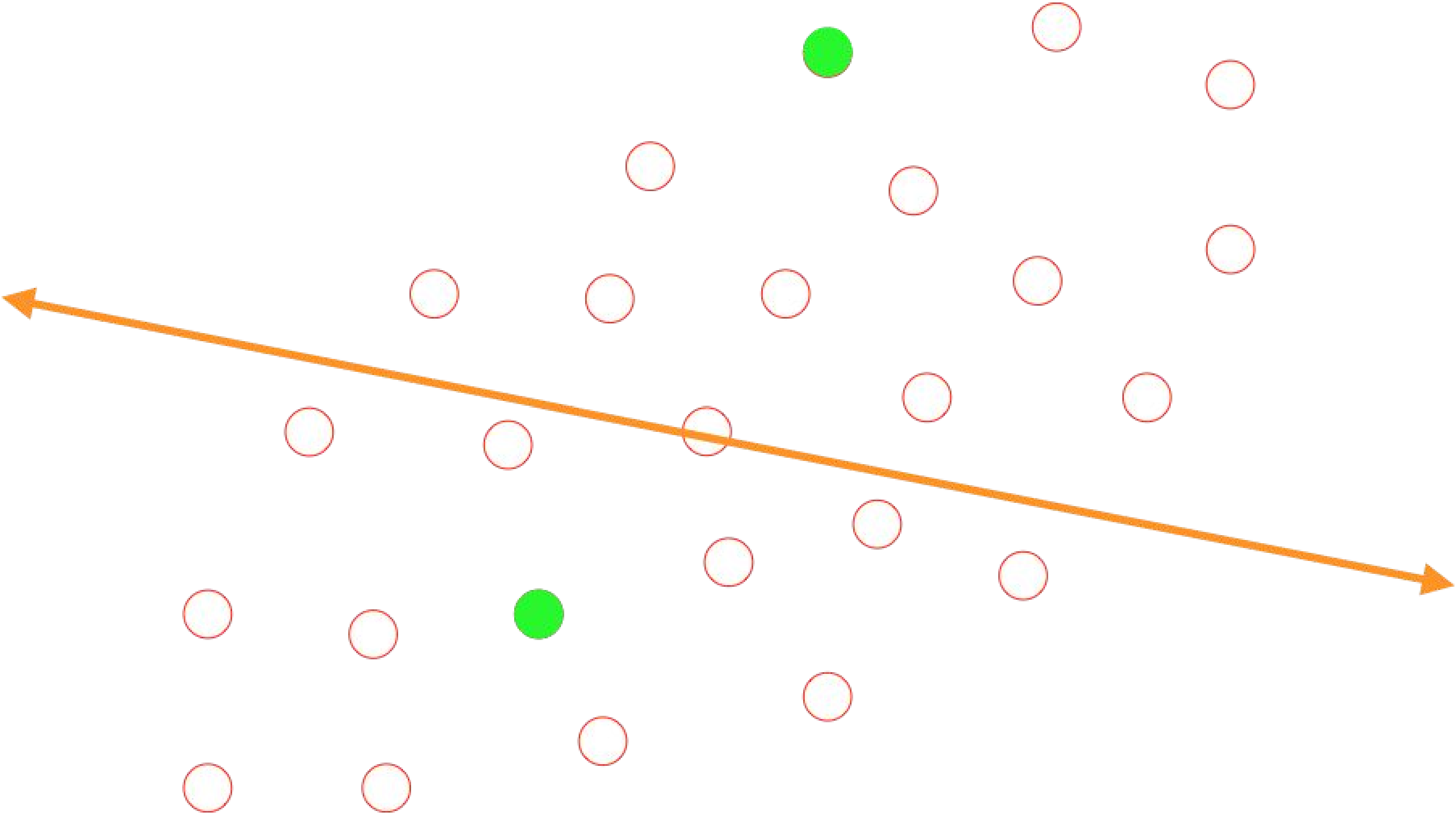
- Remember, data is a commodity. It is important to not frivolously disregard variables or observations without careful consideration.
  - Instead, be careful and selective in the dimensions we choose to analyze.

- Reducing the dimensionality of our data can often inform interpretability and statistical inference in the long run, but we can’t avoid losing some information.
  - Try to preserve as much structure from the original data as possible.

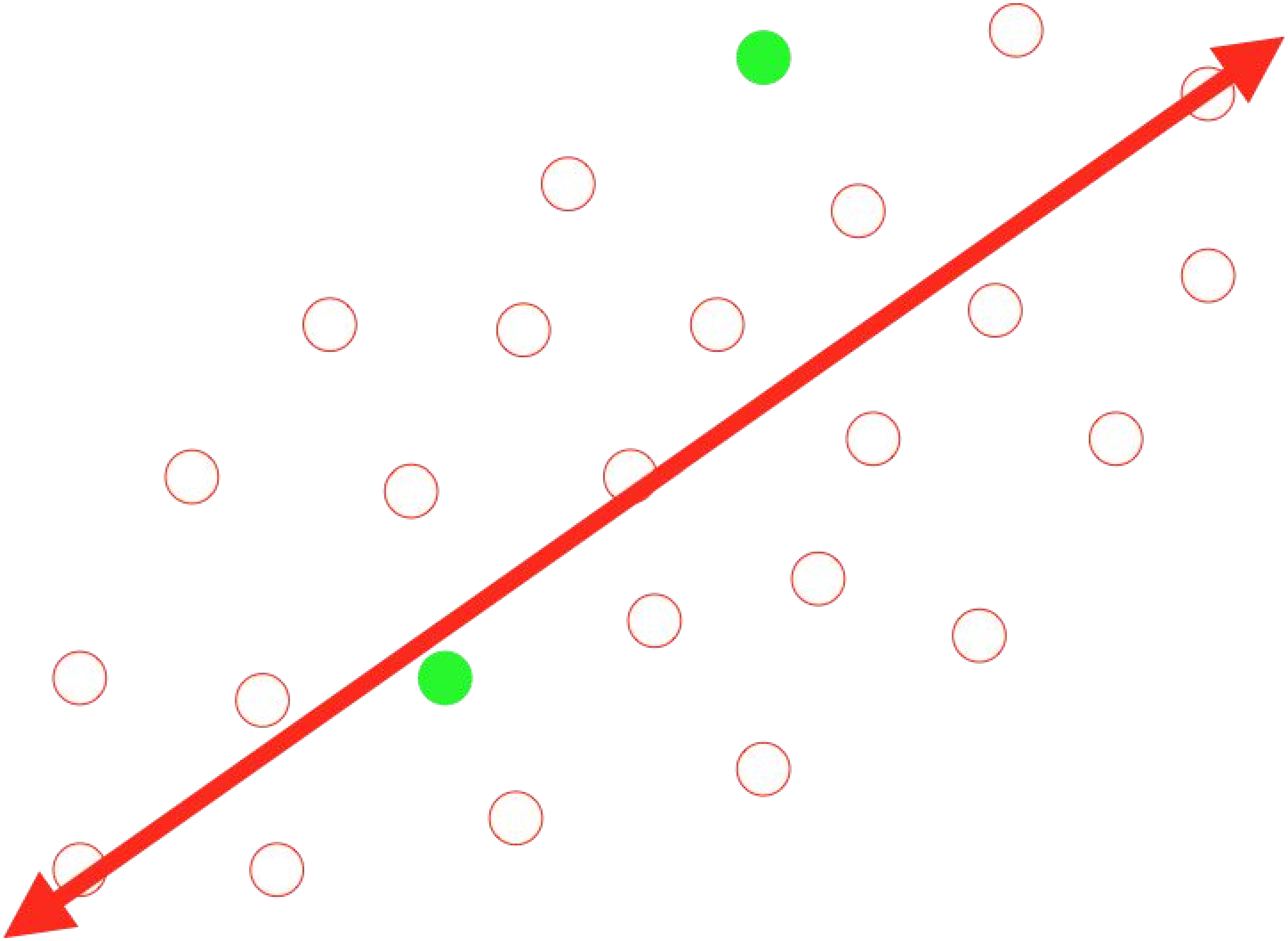
PART 3

Vectors of Highest Variance

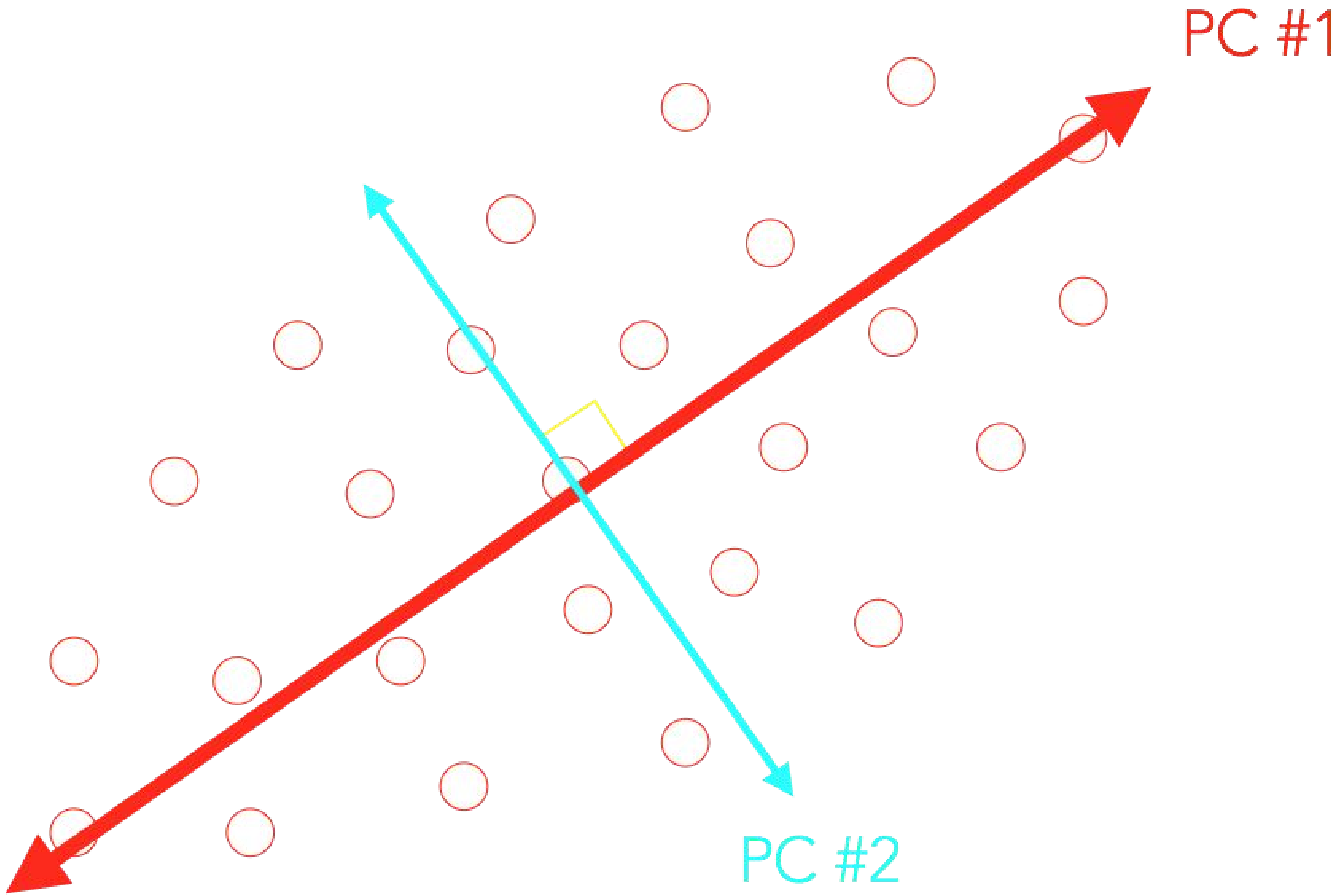
Vectors of Highest Variance



Vectors of Highest Variance



Vectors of Highest Variance



PART 4

The PCA Procedure

PCA Mathematically

- An overview of the PCA procedure mathematically:

- Center the data at 0 by subtracting off the mean from each variable:
  - Pragmatically, this allows the future mathematical processes to be easier.
  - Conceptually, PCA is modeling the variances of the data -- the mean doesn’t matter as much. We can always add the mean back in later if we desire to do a bit of back-construction.

$$x'_{i,j} = x_{i,j} - \mu_j$$

PCA Mathematically

- Compute the covariance matrix Σ:
  - Observe a unique property of convergence.

$$\Sigma v$$

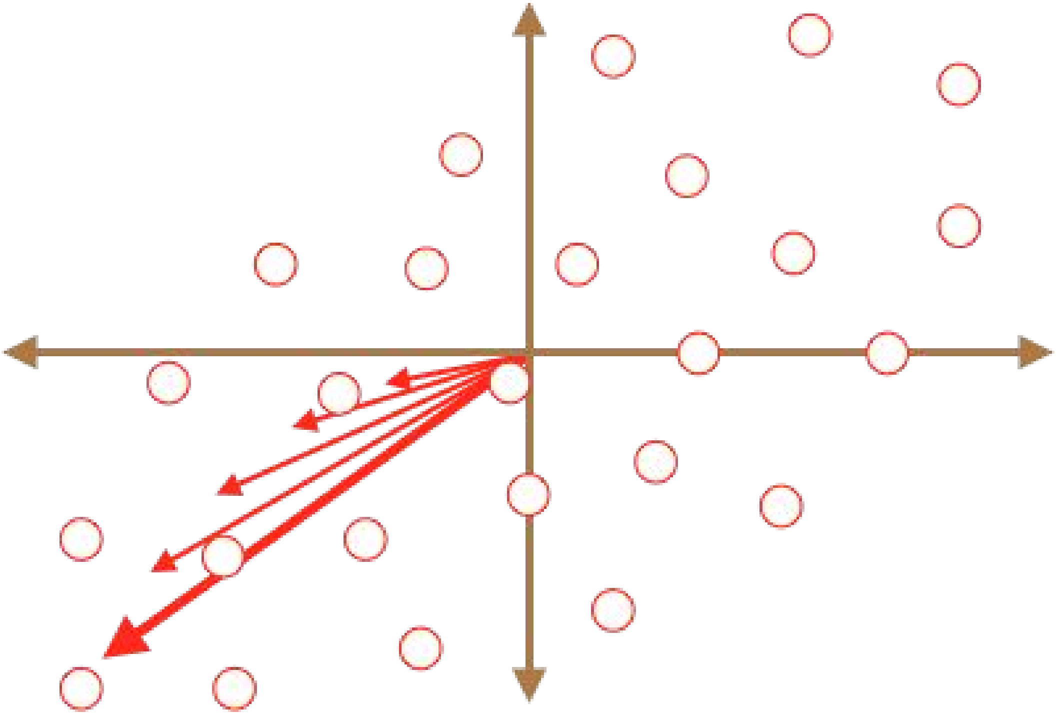
$$\Sigma(\Sigma v)$$

$$\Sigma(\Sigma(\Sigma v))$$

$$\Sigma \ldots \Sigma(\Sigma(\Sigma v)) \approx e$$

PCA Mathematically

- Compute the covariance matrix Σ:
  - Observe a unique property of convergence.



PCA Mathematically

- Find the eigenvectors e of Σ:
  - Solve the equation:

$$det(\Sigma - \lambda I) = 0$$

  - Compute the eigenvectors by finding the solutions to:

$$\Sigma e = \lambda e$$

  - The principal components are the eigenvectors e.
  - The eigenvectors are ordered by the magnitude of the corresponding eigenvalues λ.

PCA Mathematically

- Determine how many principal components to use:
  - Strike a balance between the total amount of variance that is captured by the principal components and the number of principal components selected.
  - Use the first k principal components.

- Project the original data onto the chosen k principal components.

The Result of PCA

- What do we get?
  - Transformed data that straddles only k carefully selected dimensions that preserve as much original structure as possible.

- Same data, new perspective.

Other Properties of PCA

- The following results are useful properties that can be proved using calculus and linear algebra (omitted for brevity).

- The eigenvectors of Σ yield orthogonal directions of greatest variability (principal components).

- The eigenvalues λ correspond to the magnitude of variance along the principal components.

PART 5

Alternatives to PCA: Ridge & Lasso Regression

Alternatives to Principal Component Analysis

- While principal component analysis is relatively straightforward and computationally inexpensive, its main drawback is interpretability.
  - What do the new dimensions mean?
  - Is this subjective?

- What if we had an alternative to the PCA procedure that allowed us to retain the original dimension measurements while also reducing the overall dimensionality? This can be attained by performing:
  - Subset selection
  - Shrinkage/regularization

Subset Selection

- In subset selection, we identify a subset of p predictors that we believe are related to the response. We then fit a model using least squares regression on the reduced set of variables.

- Subset selection can be performed in a myriad of ways:
  - Best subset selection
  - Forward stepwise regression
  - Backward stepwise regression
  - Both stepwise regression

- Select a model based on selected criteria:
  - AIC
  - BIC
  - $R^2_{\text{Adj.}}$

Shrinkage/Regularization

- In shrinkage/regularization, we fit a model involving all predictors; however, the estimated coefficients are shrunken towards 0 relative to the least squares estimates. As a result:
  - Estimate variance is reduced.
  - Variable selection can be performed.

- In order to regularize the coefficient estimates, we must further constrain the original least squares minimization problem. How do we do this?
  - Ridge regression
  - Lasso regression

Multiple Linear Regression: Mathematically

- Recall that in multiple linear regression we wish to quantify the relationship between X and Y as follows:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

- Our original task was to find the estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ that reduce the sum of the squared vertical distances from the observations to the regression surface (i.e., the RSS) as much as possible.
  - Solved the minimization problem using basic calculus and linear algebra.

$$RSS = \sum_{i=1}^{n} e_i^2$$

Ridge Regression: Mathematically

- Ridge regression is an extension of the minimization problem posed by multiple linear regression; rather than attempting to simply reduce the RSS, ridge regression attempts to minimize the following:

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

- **NB:** Here, λ is a tuning parameter that essentially determines how the ridge regression will operate.

Ridge Regression: Mathematically

- As in the multiple linear regression setting, ridge regression attempts to find coefficient estimates that render the RSS as small as possible, thus fitting the data well.

- Additionally, there is an added shrinkage penalty (the extra term containing the tuning parameter λ).
  - Notice, the shrinkage penalty is simply the sum of the squared coefficient estimates. This penalty will be small when the estimates are close to 0. Therefore, it has the effect of shrinking the coefficient estimates as a group!

- The value of λ determines the relative impact of the RSS and shrinkage penalty terms on the resulting coefficient estimates.
  - A small λ penalizes the RSS more than the shrinkage penalty.
  - A large λ penalizes the shrinkage penalty more than the RSS.

Ridge Regression: Mathematically

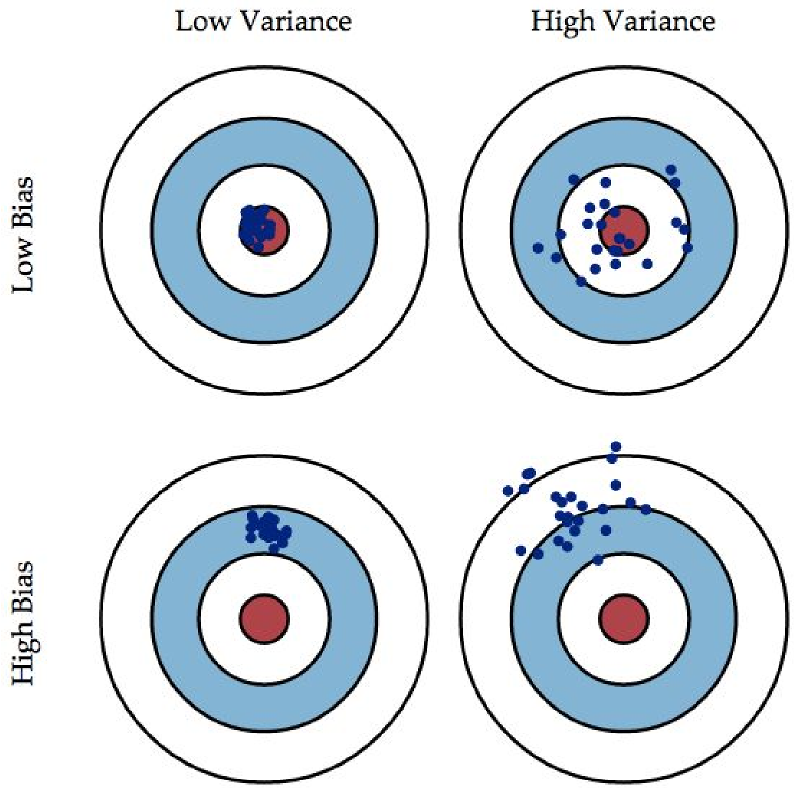
- Recall that the standard least squares coefficient estimates are scale equivariant: if we multiply a predictor variable by a constant, the corresponding least squares coefficient estimate will be scaled down by the same constant.
  - Regardless of how a predictor variable is scaled, the resulting product with the corresponding estimated coefficient will remain the same.

- In ridge regression, this is not the case. Coefficient estimates can change dramatically when multiplying a given predictor by a constant due to the shrinkage penalty; it depends on the sum of the squared coefficients!

- To avoid the issue of overvaluing or undervaluing certain predictor variables simply based on their magnitudes, we must standardize the variables prior to performing ridge regression.

Ridge Regression: Mathematically

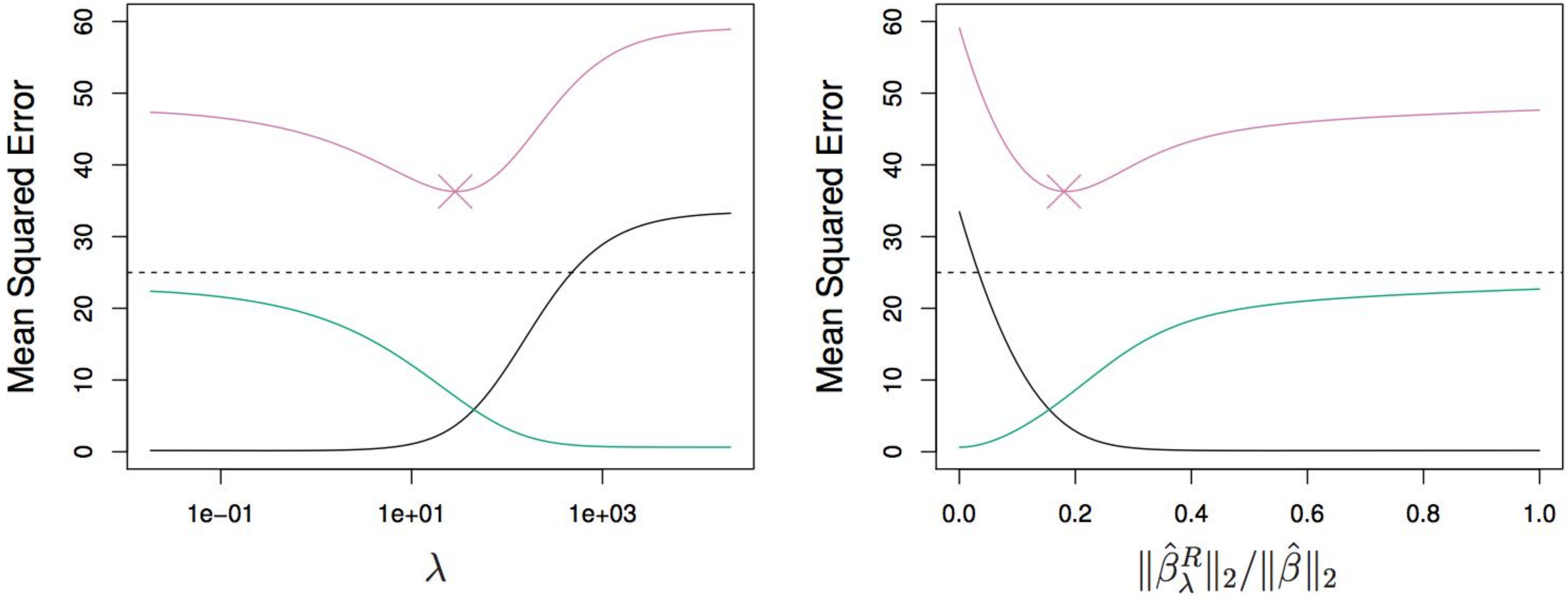
- Ideally, in any scenario, we hope to uncover a model that has:
  - Low bias (i.e., high accuracy)
  - Low variance (i.e., high precision)



Image

Ridge Regression: Mathematically

- By shrinking the coefficient estimates towards 0 by increasing λ, we see that:
  - The bias (black) increases slightly but remains relatively small.
  - The variance (green) reduces substantially.
  - The mean squared error (red) of the predictions drops.



Lasso Regression: Mathematically

- The main disadvantage of ridge regression is that, while parameter estimates are shrunken, they only asymptotically approach 0 as we increase the value of λ.
  - Thus, the resulting model still includes estimates for all parameters.

- Lasso regression is another extension of the minimization problem posed by multiple linear regression; rather than attempting to simply reduce the RSS, lasso regression attempts to minimize the following:

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

- **NB:** The only difference between the lasso and ridge regressions is that lasso implements the l_1 penalty (norm) rather than the l_2 penalty.

Lasso Regression: Mathematically

- While both ridge and lasso regression have the effect of shrinking coefficient estimates towards 0, the lasso necessarily forces some coefficient estimates to be exactly 0 (when λ is sufficiently large).

- Lasso regression has the added advantage of essentially performing variable selection, yielding models that are both accurate and parsimonious.

- Once again, the value of λ determines the relative impact of the RSS and shrinkage penalty terms on the resulting coefficient estimates.
  - In both ridge and lasso regression, it is important to select an appropriate value of λ by means of cross-validation.

Lasso VS Ridge

- Why is it the case that lasso regression results in coefficient estimates that are exactly 0? Why does ridge regression only end up shrinking the coefficients?

- We can restate the optimization problems of lasso and ridge regression, respectively, as the following Lagrangian multiplier scenarios:

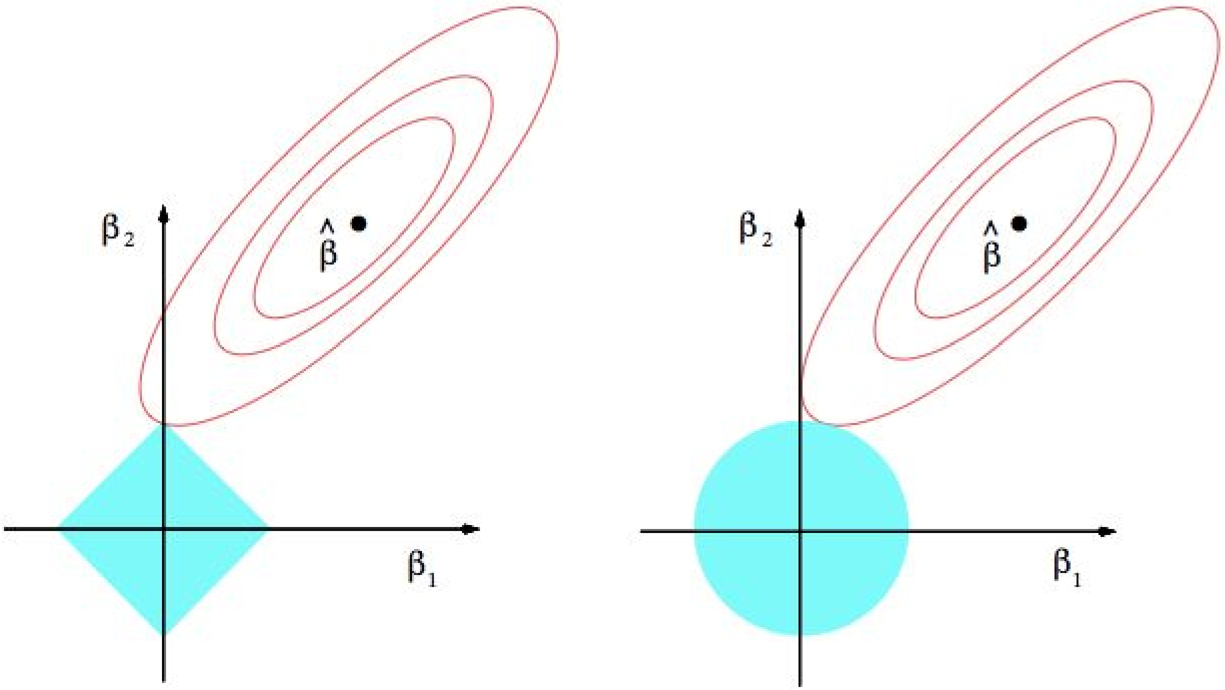
$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

Lasso VS Ridge: Visually

- The red ellipses represent the contours of the least squares error function.
- The blue regions represent the constrained regions for lasso and ridge, respectively.



Image

Selecting the Tuning Parameter: λ

- How do we choose the best value of λ?

- Let's take a look at the method of cross-validation.

PART 6

Cross-Validation

The Training & Test Sets

- When conducting cross-validation, we generally split the observations in our data into different (non-overlapping) sections:
  - The training set
  - The test set

- We use the observations in the training set to fit an initial model, but then can calculate the error in different ways:
  - The training error is calculated by applying the fitted model to the observations used to create the initial model (the training set).
  - The test error is calculated by applying the fitted model to the observations not used to create the initial model (the test set); this is like assessing the predictions of new observations.

- Why does the training error underestimate the test error?

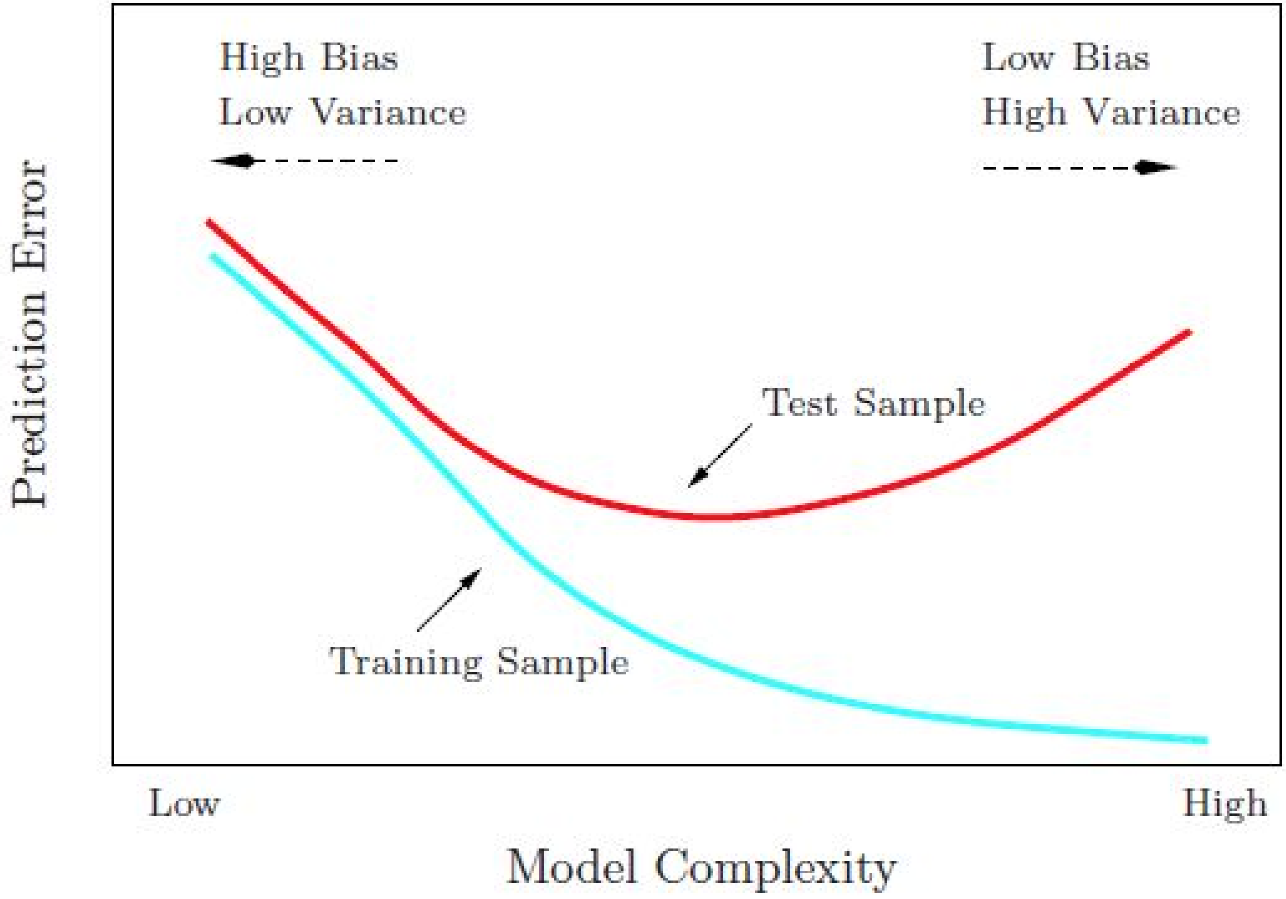
The Bias-Variance Tradeoff

- Recall that:
  - Bias is how far off on the average the model is from the truth.
  - Variance is how much the estimate varies about its average.

- With low model complexity:
  - Bias is high because predictions are more likely to stray from the truth with an inflexible model.
  - Variance is low because there are only few parameters being fit.

- With high model complexity:
  - Bias is low because the model can adapt to more subtleties in the data.
  - Variance is high because we have more parameters to estimate from the same amount of data.

The Bias-Variance Tradeoff



K-Fold Cross-Validation

- The results of K-fold cross-validation can help determine the best model at hand by estimating the test error among ultimate models.

- The process boils down to the following steps:
  - ➢ Divide your data into K (relatively) equal parts.
  - ➢ Leave out one of the K parts (call it part k), and put it to the side.
  - ➢ Fit the model to the remaining (K - 1) parts all together as your training set.
  - ➢ Use part k as your test set to estimate the prediction error.
  - ➢ Repeat this process K times, once each for the different splits of your data.

K-Fold Cross-Validation

- The test error can be estimated from the results of this process by essentially computing a weighted average of the K folds as follows:

$$CV_K = \sum_{k=1}^{K} \frac{n_k}{N} MSE_k$$

- Here:
  - K is the number of groups (folds).
  - n_k is the number of observations in fold k out of the total N observations.
  - MSE$_k$ is the mean squared error obtained by using fold k as the test set, and the remaining data as the training set.

- **NB:** For classification problems, the MSE is simply replaced by the error rate.

Yet Again, How do we Choose K?

- There is yet again another bias-variance tradeoff!

- Consider the extreme case where we choose K to be equal to the total number of observations (leave-one-out cross-validation). LOOCV will yield:
  - Nearly unbiased estimates of the test error since the training sets will be extremely similar to the overall dataset.
  - High estimate variances because we are in effect averaging the outputs of n fitted models that all have been created from extremely similar datasets; the results are highly correlated with one another.
    - In other words, the mean of highly correlated quantities has higher variance than the mean of uncorrelated quantities.

Yet Again, How do we Choose K?

- So what do we do?

- 5- or 10-fold cross-validation is typically used because these values have been empirically shown to yield estimates of the test error rate that tend to neither suffer from extreme bias nor high variance.

Selecting the Tuning Parameter: λ

- In order to best implement the ridge and lasso regression methods, we need to have a way of determining the best value of λ (or, equivalently, the constraint s in the Lagrange multiplier formulation of the problem).

- As mentioned earlier, cross-validation helps us check by iterating across a slew of λ values and computing the cross-validation error rate for each.
  - Split the data into training and test sets (10-fold CV).
  - Select the λ for which the cross-validation error is the smallest.

- Lastly, refit the model using all available observations, this time with the best selected value of the tuning parameter.

PART 7

Review

Review

- Part 1: Taking a New Perspective

- Part 2: Dimension Reduction
  - Why can High Dimensionality be Adverse?
  - Don’t Just Throw Away Data!

- Part 3: Vectors of Highest Variance

- Part 4: The PCA Procedure
  - PCA Mathematically
  - The Result of PCA
  - Other Properties of PCA

- Part 5: Alternatives to PCA: Ridge & Lasso Regression
  - Alternatives to PCA
  - Subset Selection
  - Shrinkage/Regularization
  - Mathematically
    - Multiple Linear Regression
    - Ridge Regression
    - Lasso Regression
  - Lasso VS Ridge
    - Visually

- Part 6: Cross-Validation
  - The Training & Test Sets
  - The Bias-Variance Tradeoff
  - K-Fold Cross-Validation
  - Yet Again, How do we Choose K?
  - Selecting the Tuning Parameter: λ