High Dimensional Data

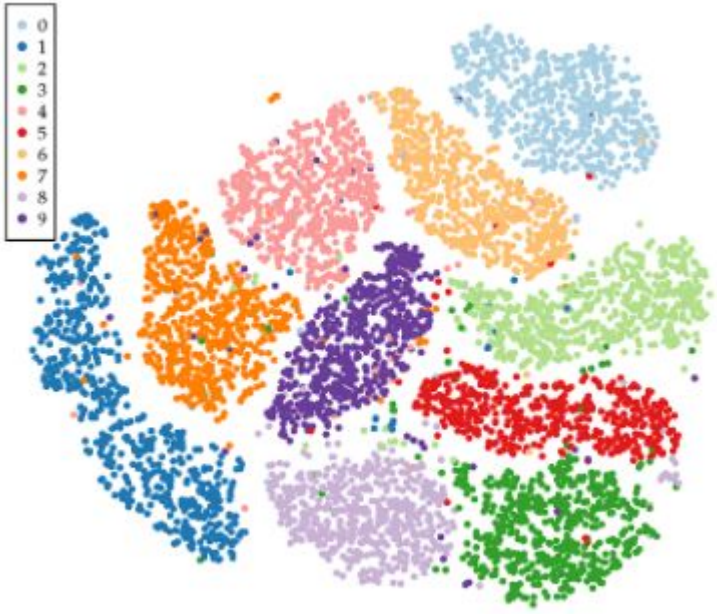
Section 1-Team 11
Andrew Burnick, Killian McKee, Pinki Nathani, and Mark Zhang

Agenda

- Introduction to high dimensional data (H.D.D.)
- Review ridge regression and lasso
- Challenges of working in high dimensions
- Regression in high dimensions
- Interpreting high dimensional results
- Takeaways
- Activity

What is High Dimensional Data?

- Traditional statistics:
  - Low dimensional data i.e. high sample sizes, small number of predictors
  - Limited # of predictors
  - Models reliant on large sample sets to be accurate
- High dimensional data: data with more predictors than observations
  - n<p
  - Technology makes modeling these possible



Importance & Examples of H.D.D.

- Importance: modeling scenarios w/ expensive or limited samples, many predictors
  - EX1. Predicting blood pressure with millions of individual DNA mutations
  - EX2. Early stage pharmaceutical trials (N<=10)
  - EX3. Search histories for internet users

Ridge & Lasso Review (Shrinkage)

Ridge

- Type of regularization
  - Alternative to subsetting
- Keeps all model features
- Reduces magnitude of coefficients
- Seeks to minimize equation below
  - Left side= RSS error
  - Right side= penalty for adding parameters
  - Lambda dictates penalty size
    - Big lambda, coefficients approach 0

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\sum_{j}^{m}\beta_j^2$$

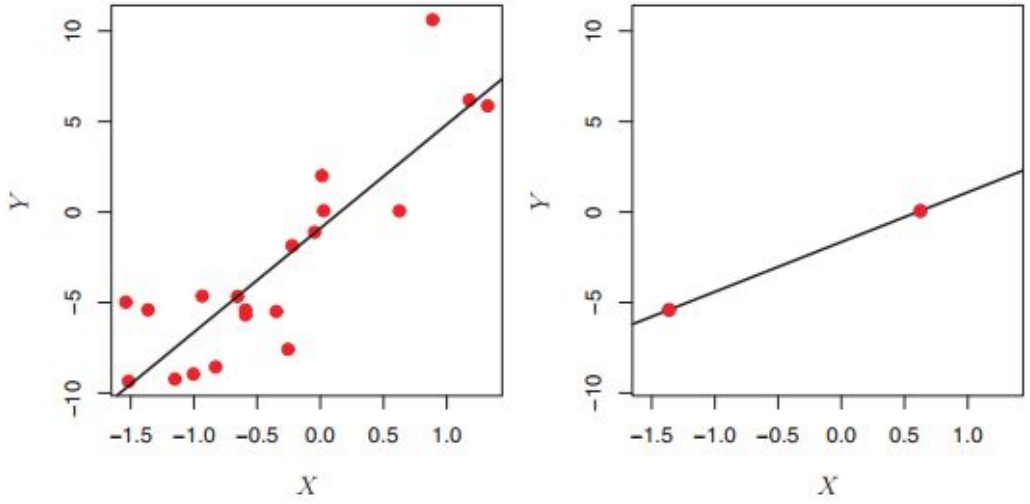
Lasso

- Another type of regularization
- Shrinks parameters AND performs variable selection
  - Can 0 out coefficients
- Seeks to minimize equation below
  - Left side = RSS error
  - Right side = penalty for adding parameters, but can 0 coefficients
  - Lambda dictates penalty

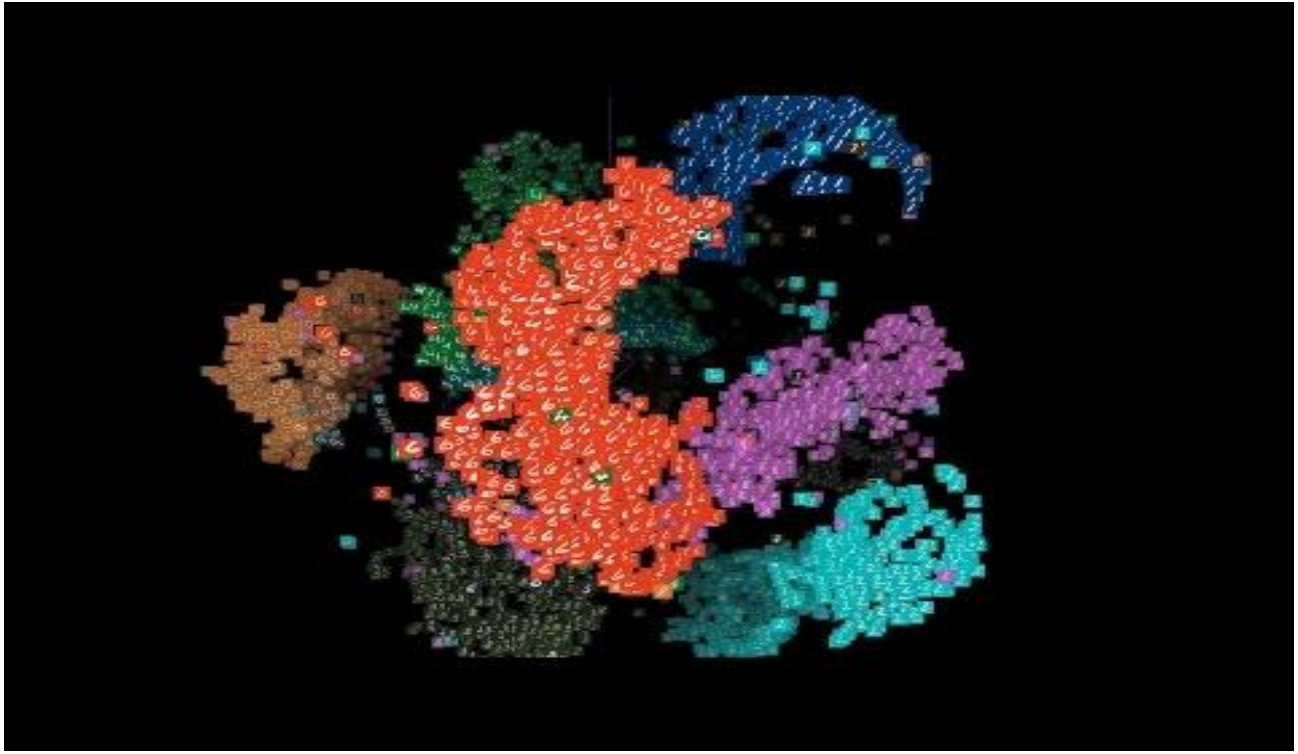
$$RSS+\lambda\sum_{j=1}^{p}|\beta_j|$$

What Goes Wrong in High Dimensions?

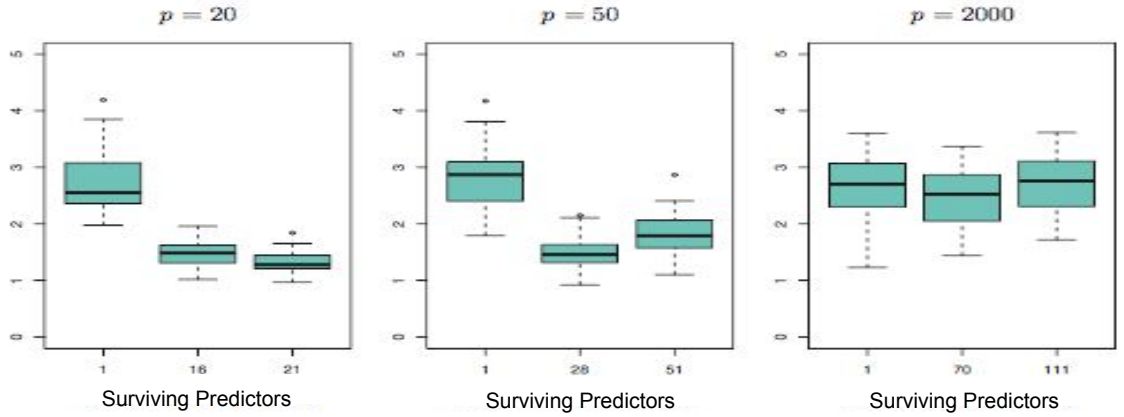
- Traditional regression/classification methods are infeasible in high dimensional settings
  - Least Square, Logistic Regression, LDA etc.
  - Models will show a perfect fit regardless of whether or not there truly is a relationship between features and the response variable.



Regression in Higher Dimensions: Video



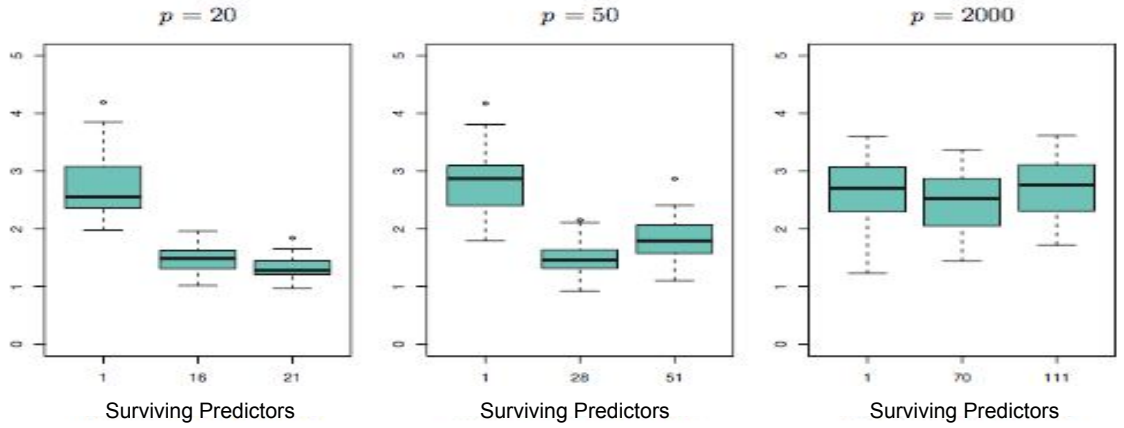
Regression in Higher Dimensions: Example



Y axis=MSE
X Axis=# of surviving predictors
Chosen Model=Lowest MSE
Based on 20 truly correlated features

- Stepwise selection, ridge regression, the lasso, and principal components regression are useful for performing high dimensional regressions.
  - These methods avoid overfitting by using a more flexible approach than least squares
- The above diagram illustrates a simulated lasso regression in a high dimensional setting
  - There are p = 20, 50, or 2,000 features of which 20 are truly associated with the outcome--Lasso must sort through the cases where p>truly correlated variables

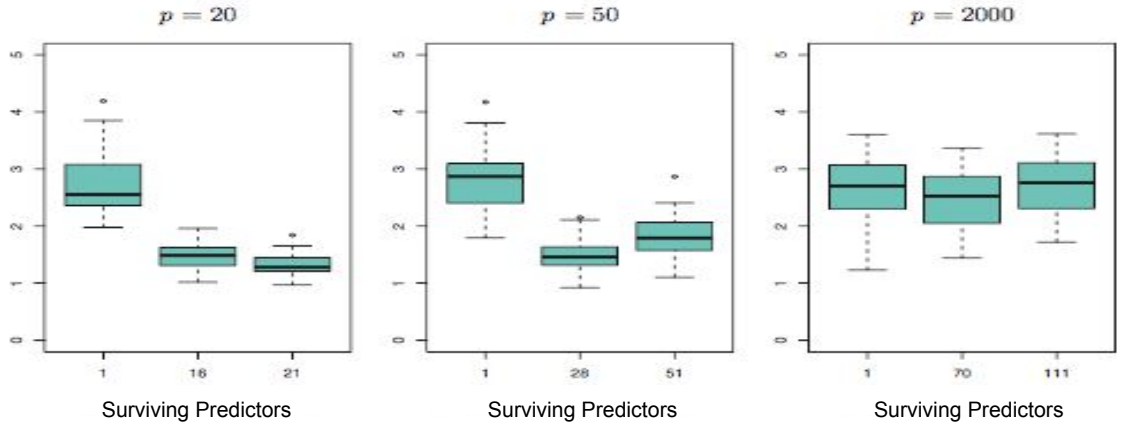
Regression in Higher Dimensions: Example



Y axis=MSE
X Axis=# of surviving predictors
Chosen Model=Lowest MSE
Based on 20 truly correlated features

- Lasso with n = 100 observations and three different p values (Left, Center, Right), the # of features applied to each of the three cases.
- Of p features, ~ 20 are associated with the response variable. The boxplots show test MSEs stemming from 3 different values of the tuning parameter λ. Lambda increases rapidly to reduce noise when p> 20
- Degrees of freedom are reported on the X-axis; for the lasso, this is the number of estimated non-zero coefficients that the Lasso finds for each MSE reported.

Regression in Higher Dimensions: Example

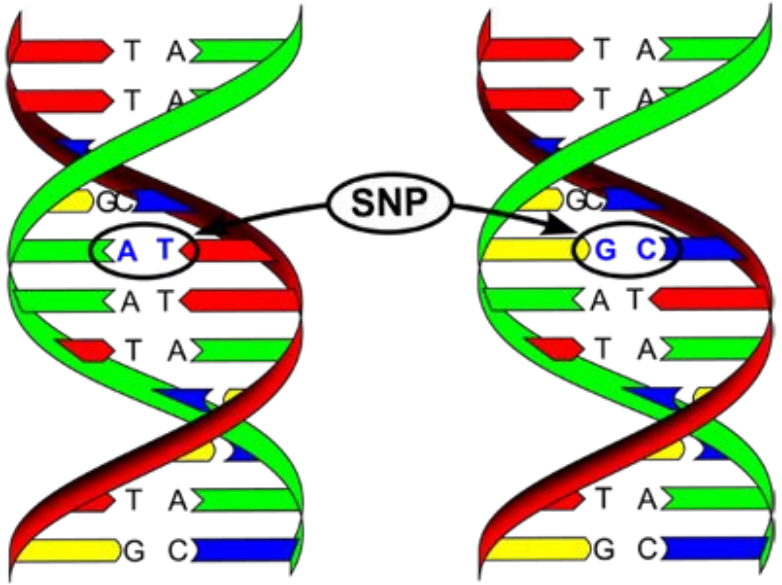


Y axis=MSE
X Axis=# of surviving predictors
Chosen Model=Lowest MSE
Based on 20 truly correlated features

- **When p = 20**, the lowest test MSE was obtained with the smallest shrinkage since p=20. Lambda is smallest here since the best model (lowest Test error) only includes the truly correlated features.
- **When p = 50**, the lowest test MSE was achieved with a substantial amount of regularization--Lasso reduced features from 50 to 28 with a slightly higher MSE than before
- **When p = 2,000** the Lasso regularized down to 111 predictors (about 20 actually relevant). The model has substantial shrinkage and the highest Lambda of all 3 scenarios

Interpreting High Dimensional Results

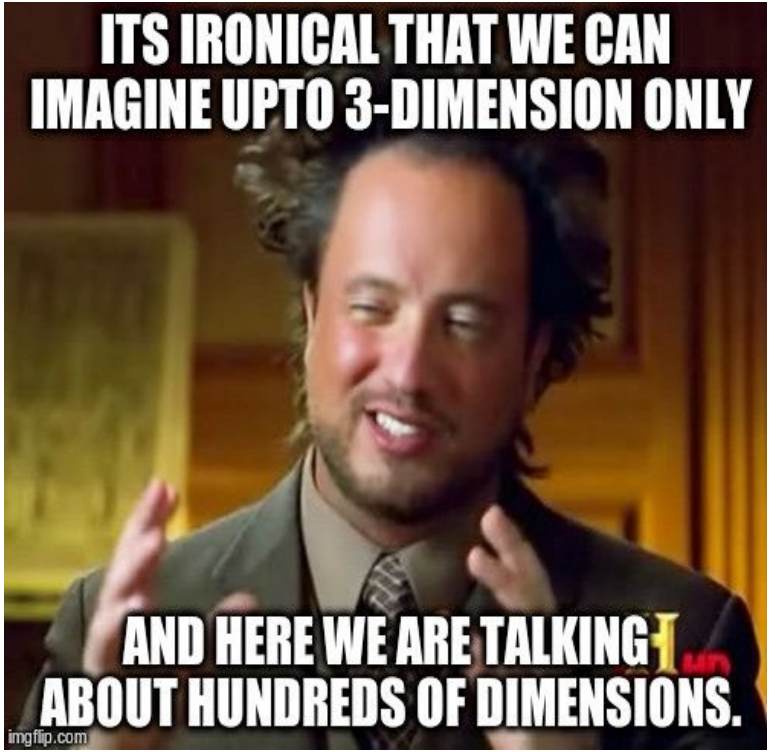
- Must consider extreme potential for multicollinearity
- Revisiting blood pressure example
  - Forward stepwise identifies 17 genetic predictors (SNPs) for a given blood pressure are these the only 17 capable of accurate prediction?
    - No, we have identified 1 of many possible models
    - Another independent data set could reveal 20 completely different, equally accurate genetic predictors for blood pressure.



High Dimensional Data : Conclusions

5 Key Takeaways

1. Subsetting/regularization/shrinkage plays a key role in high-dimensional problems
   a. Use forward step, lasso, ridge, PCA
2. Appropriate tuning parameter selection is crucial for good predictive performance
3. The test error tends to increase as the dimensionality of the problem (i.e. the number of features or predictors) increases, unless the additional features are truly associated with the response.
4. Similarly, adding more predictors improves fit, but increases bias (recall the bias/variance trade off)
5. Multicollinearity needs to be considered



Activity

1. Room will be split into teams
2. Fill out the quiz
3. The 1st place team gets to pick out a snack 1st, 2nd goes next, etc.

Activity Answers

1. What could happen if you run a least squares in R when p > n?
   (D. R would only use as many features as the number of observations)
2. If we keep adding features until p = n, what could happen to the test MSE?
   (C, as more features are added into the model, the flexibility will go up, but the test MSE will also go up)
3. In a setting of p >= n, what causes the R-Squared term to be 1?
   (C, because the residual would be 0)
4. What is multicollinearity?
   (High correlation between 3 or more predictor variables)
5. Give a business example of a high dimensional modeling situation we haven't yet discussed.
   (Tracking customer clicks through a website to predict purchases)
6. List three ways we can reduce dimensionality in a high dimensional setting
   (Ridge, Lasso, PCA)
7. What is the primary difference between ridge and lasso?
   (Lasso can 0 out coefficients while ridge is only capable of bringing them near 0)
8. Describe two challenges associated with handling high dimensional data as opposed to a low dimensional setting.
   (Extreme potential of multicollinearity problem, difficulty of dimension reduction when p is very large)