Prediction with Many Regressor and Big Data

Prf. José Fajardo
FGV/EBAPE

Outline

1. What is “Big Data”?
2. Prediction with many predictors: the MSPE, OLS, and the principle of shrinkage
3. Ridge regression
4. The Lasso
5. Principal components
6. Application to prediction of school test scores
7. Summary

1. What is “Big Data”?

“Big Data” means many things:

- Data sets with many observations (millions)
- Data sets with many variables (thousands, or more)
- Data sets with nonstandard data types, like text, voice, or images

1. What is "Big Data"? (2 of 3)

"Big Data" also has many different applications:

- Prediction using many predictors
  - Given your browsing history, what products are you most likely to shop for now?
  - Given your loan application profile, how likely are you to repay a bank loan?
- Prediction using highly nonlinear models (for which you need many observations)
- Recognition problems, like facial and voice recognition

1. What is "Big Data"? (3 of 3)

"Big Data" has different jargon, which makes it seem very different than statistics and econometrics…

- "Machine learning:" when a computer (machine) uses a large data set to learn (e.g., about your online shopping preferences)

But at its core, machine learning builds on familiar tools of prediction.

- We focus on one of the major big data applications, prediction with many predictors. We treat this as a regression problem, but with many predictors we need new methods that go beyond OLS.
- For prediction, we do not need – and typically will not have – causal coefficients.

Challenges for Management/Regulation of ML in Financial Services

Algorithms have demonstrable errors

Engineers build black-box algorithms, but are not trained to evaluate

Need "best practices" to analyze the black box

- Credit Scoring Example
  - **Instability** of joint distribution of outcomes, novel features
  - **Poor performance** when extrapolating
  - **Manipulation** of novel features
  - **Discrimination** and **Fairness**
  - Ever-changing **adverse selection** problem as competing firms change models, marketing strategies
  - When are results more or less **reliable**?
- **Equilibrium** effects
  - Agents using ML interact
  - Collusion (airline prices)
  - Instability (financial market crashes, correlated mistakes across firms)
  - Google maps examples
- Need models of individual behavior and eqm selection to study eqm changes
  - Why existing AI/ML is a long way from solving "harder" problems

ML and Econometrics

Causal inference vs. Supervised ML

- Supervised learning:
  - Can evaluate in test set in model-free way
    - MSE: $\sum(Y_i - \hat{\mu}(X_i))^2$
- Causal inference
  - Objective: unbiased/consistent parameter estimation
  - Parameters of interest not observed in test set
  - Can estimate objective (MSE of parameter), but requires maintained assumptions, often not model-free

    Infeasible MSE: $\sum\left(\theta_i - \hat{\theta}(X_i)\right)^2$
  - Tune for counterfactuals: distinct from tuning for fit, also different counterfactuals select different models
  - Theoretical assumptions, domain knowledge
  - Sampling variation matters even in large data sets
    - Statistical theory and inference play important roles

2. Prediction with many predictors: the *MSPE*, OLS, and the principle of shrinkage

The many-predictor problem:

- The goal is to provide a good prediction of some outcome variable *Y* given a large number of *X* 's, when the number of *X* 's (*k*) is large relative to the number of observations (*n*) – in fact, maybe $k > n$!
- The goal is good out-of-sample prediction.
  - The **estimation sample** is the *n* observations used to estimate the prediction model
  - The prediction is made using the estimated model, for an **out-of-sample (OOS)** observation – an observation *not* in the estimation sample.

The Predictive Regression Model

The **standardized predictive regression model** is the linear model, with the exception that all the *X* 's are normalized (standardized) to have a mean of zero and a standard deviation of one, and *Y* is deviated from its mean:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i \qquad (14.2)$$

- We assume $E(Y \mid X) = \beta_1 X_1 + .. + \beta_k X_k$ so $E(u_i \mid X_i) = 0$
- Because all the variables, including *Y*, are deviated from their means, the intercept is zero – so is omitted from (14.2).
- As usual, (14.2) allows for linearities in *X* by letting some of the *X* 's be squares, cubes, logs, interactions, etc.
- Throughout this slides, we use standardized *X* 's, demeaned *Y* 's, and the standardized predictive regression model (14.2).

The Mean Squared Prediction Error

The Mean Squared Prediction Error (MSPE) is the expected value of the squared error made by predicting Y for an observation not in the estimation data set:

$$MSPE = E\left[Y^{OOS} - \hat{Y}(X^{OOS})\right]^2$$

where:

- Y is the variable to be predicted
- X denotes the k variables used to make the prediction, (X^{OOS}, Y^{OOS}) are the values of X and Y in the out-of-sample data set.
- The prediction $\hat{Y}(X^{OOS})$ uses a model estimated using the estimation data set, evaluated at X^{OOS}.

- The MSPE measures the expected quality of the prediction made for an out-of-sample observation.

The First Least Squares Assumption for Prediction

- For prediction, it does not matter whether model coefficients have a causal interpretation – so we do not need the first least squares assumption for causal inference, which is defined in terms of a causal effect.
  - In the predictive model (14.2), β is defined to be the coefficient in the (linear) conditional expectation, $E(Y|X)$.
- But it does matter that data for which we will be making the prediction is similar to the data used to estimate the model:

  The first least squares assumption for prediction:
  (X^{OOS}, Y^{OOS}) are drawn from the same distribution as the estimation sample, (X_i, Y_i), $i = 1,\ldots, n$.

The Oracle Prediction

The **oracle prediction** is the best-possible prediction – the prediction that minimizes the MSPE – if you knew the joint distribution of Y and X.

The oracle prediction is the conditional expectation of Y given X, $E(Y^{OOS}|X = X^{OOS})$

- Suppose not. Then the forecast error could be predicted using X^{OOS} - but if so, the forecast couldn't have been the best possible, because it could be improved using the predicted error.

The MSPE for the standardized predictive regression model

OOS value of *Y*: $Y^{OOS} = \beta_1 X_1^{OOS} + ... + \beta_k X_k^{OOS} + u^{OOS}$

Prediction: $\hat{Y}^{OOS} = \hat{\beta}_1 X_1^{OOS} + ... + \hat{\beta}_k X_k^{OOS}$

Prediction error: $Y^{OOS} - \hat{Y}^{OOS} = (\beta_1 - \hat{\beta}_1) X_1^{OOS} + ... + (\beta_k - \hat{\beta}_k) X_k^{OOS} + u^{OOS}$

Let $E(u^{OOS})^2 = \sigma_u^2$. Then the MSPE for the standardized predictive regression model is,

$$MSPE = \sigma_u^2 + E\left[(\hat{\beta}_1 - \beta_1) X_1^{OOS} + ... + (\hat{\beta}_k - \beta_k) X_k^{OOS}\right]^2$$

- The term σ_u^2 is the MSPE of the oracle forecast – can't be beat!
- The second term arises because the β's aren't known, so must be estimated using the estimation sample.

The MSPE of OLS

- Suppose the β's are estimated by OLS. In the standardized predictive regression model, the MSPE of OLS is approximately,

$$MSPE_{OLS} \cong \left(1 + \frac{k}{n}\right)\sigma_u^2$$

  - This approximation holds if *u* is homoskedastic and *k*/*n* is small.
- For a given *n*, the MSPE of OLS increases linearly with the number of predictors. A big problem with many predictors!
- Is there an estimator for which the MSPE increases more slowly than OLS, as more predictors are added?
  - We need such an estimator for hundreds or thousands of predictors

The Principle of Shrinkage (1 of 2)

In the 1950s, the statisticians figured out that you could reduce the MSPE, relative to OLS, by allowing the estimator to be biased in the right way.

- When the *X*'s are uncorrelated, these estimators are biased towards zero – or "shrunk" towards zero – and have the form,

$$\hat{\beta}^{JS} = c\hat{\beta}$$

  where $0 < c < 1$ and "JS" stands for James-Stein.
- But how could introducing bias possibly help???

The Principle of Shrinkage (2 of 2)

- The James-Stein shrinkage estimator:

$$\hat{\beta}^{JS} = c\hat{\beta}$$

where $0 < c < 1$.

- As c gets smaller:
  - The squared bias of the estimator increases,
  - But the variance decreases.
  - This produces a bias-variance tradeoff. If k is large, the benefit of smaller variance can beat out the cost of larger bias, for the right choice of c – thus reducing the MSPE.
- **The estimators all have a shrinkage interpretation**

Estimating the MSPE

The MSPE is a bit tricky to estimate – it isn't just the regression SER, it is for an out-of-sample, not in-sample, observation.

Split-sample estimation of the MSPE

- This method simulates the out-of-sample prediction exercise – but using only the estimation sample (which is all you have!):
  1. Estimate the model using half the estimation sample.
  2. Use the estimated model to predict Y for the other half of the data – called the "reserve" or "test" sample – and calculate the prediction error.
  3. Estimate the MSPE using the prediction errors for the test sample:

$$MSPE_{split-sample} = \frac{1}{n_{test}} \sum_{\substack{\text{observations in} \\ \text{test sample}}} \left(Y_i - \hat{Y}_i\right)^2$$

Estimating the *MSPE* by *m*-fold cross-validation

- The split-sample estimate typically overstates the MSPE because the model is estimated on only 50% of the data.
- This problem is reduced by using m-fold cross validation.
- ***m*-fold cross-validation**, for the case $m = 10$:
  1. Estimate the model on 90% of the data and use it to predict the remaining 10%
  2. Repeat this on the remaining 9 possible subsamples (so there is no overlap on the test samples).
  3. Estimate the MSPE using the full set of out-of-sample predictions

3. Ridge Regression

The **ridge regression estimator** shrinks the estimate towards zero by penalizing large **squared values** of the coefficients.

The ridge regression estimator minimizes the penalized sum of squared residuals,

$$S^{Ridge}(b;\lambda_{Ridge}) = \sum_{i=1}^{n}\left(Y_i - b_1 X_{1i} - \ldots - b_k X_{ki}\right)^2 + \lambda_{Ridge}\sum_{j=1}^{k} b_j^2$$

where $\lambda_{Ridge}\sum_{j=1}^{k} b_j^2$ is a "penalty term."

- If the regressors are uncorrelated,

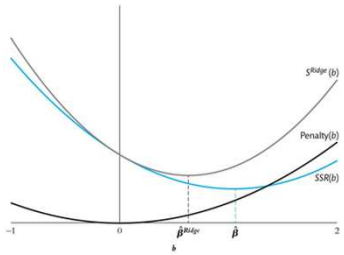
$$\hat{\beta}_j^{Ridge} = \left(\frac{1}{1+\lambda_{Ridge}/\sum_{i=1}^{n} X_i^2}\right)\hat{\beta}_j$$

so the ridge regressor has the James-Stein form, $\hat{\beta}^{JS} = c\hat{\beta}$

Ridge Regression in a Picture

The ridge regression penalty term penalizes the sum of squared residuals for large values of β, as shown here for $k = 1$:

- The value of the ridge objective function, $S^{ridge}(b)$, is the sum of squared residuals plus a penalty which is a quadratic in b.
- Thus, the penalized sum of squared residuals is minimized at a smaller value of b than is the unpenalized SSR.



Choosing the Ridge Regression penalty factor λ_{Ridge}

The ridge regression estimator has an additional parameter, λ_{Ridge}:

$$S^{Ridge}(b;\lambda_{Ridge}) = \sum_{i=1}^{n}\left(Y_i - b_1 X_{1i} - \ldots - b_k X_{ki}\right)^2 + \lambda_{Ridge}\sum_{j=1}^{k} b_j^2$$

- It would seem natural to choose λ_{Ridge} by minimizing both b and λ_{Ridge} – but doing so would simply choose $\lambda_{Ridge} = 0$, which would just get you back to OLS!
- Instead, λ_{Ridge} can be chosen by minimizing the m-fold cross-validated estimate of the *MSPE*.
  - Choose some value of λ_{Ridge}, and estimate the MSPE by m-fold cross-validation
  - Repeat for many values of λ_{Ridge}, and choose the one that yields the lowest MSPE.

Empirical Example: Predicting School-level test scores

Data set: a school-level version of the California elementary district data set, augmented with additional variables describing school, student, and district characteristics

The full data set has 3932 observations. Half of those (1966) are used now – the remaining 1966 are reserved for an out-of-sample comparison of the ridge v. other prediction methods, done later.

The data set has 817 predictors…

Predicting School-level test scores

Variables in the 817-predictor school test score data set

Main variables (38)	
Fraction of students eligible for free or reduced-price lunch	Ethnicity variables (8): fraction of students who are American Indian, Asian, Black, Filipino, Hispanic, Hawaiian, two or more, none reported
Fraction of students eligible for free lunch	Number of teachers
Fraction of English learners	Fraction of first-year teachers
Teachers' average years of experience	Fraction of second-year teachers
Instructional expenditures per student	Part-time ratio (number of teachers divided by teacher full-time equivalents)
Median income of the local population	Per-student expenditure by category, district level (7)
Student–teacher ratio	Per-student expenditure by type, district level (5)
Number of enrolled students	Per-student revenues by revenue source, district level (4)
Fraction of English-language proficient students	
Ethnic diversity index	
+ Squares of main variables (38)	
+ Cubes of main variables (38)	
+ All interactions of main variables ($38 \times 37/2 = 703$)	
Total number of predictors = $k = 38 + 38 + 38 + 703 = 817$	

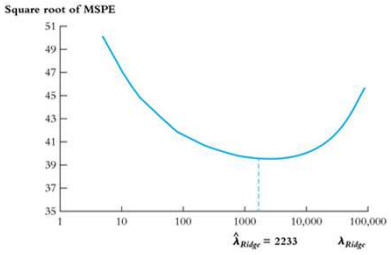
Predicting School-level test scores: Ridge Regression

λ_{Ridge} is estimated by minimizing the 10-fold cross-validated MSPE

The resulting estimate of the shrinkage parameter is 39.5

Root MSPE's:
OLS: 78.2
Ridge: 39.5

$k = 817$, $n = 1966$



Ridge results cuts the square root of the MSPE in half, compared to OLS!

4. The Lasso

The **Lasso estimator** shrinks the estimate towards zero by penalizing large **absolute values** of the coefficients.

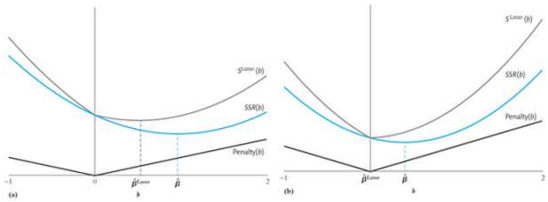
The Lasso regression estimator minimizes the penalized sum of squared residuals,

$$S^{Lasso}(b;\lambda_{Lasso}) = \sum_{i=1}^{n}(Y_i - b_1X_{1i} - \ldots - b_kX_{ki})^2 + \lambda_{Lasso}\sum_{j=1}^{k}|b_j|$$

where $\lambda_{Lasso}\sum_{j=1}^{k}|b_j|$ is the "penalty term."

- This looks a lot like ridge estimation – but it turns out to have very different properties…

Lasso in Pictures



When the OLS estimator is large, the Lasso estimator shrinks it slightly towards zero – less than ridge…

…but when the OLS estimator is small, the Lasso estimator shrinks it **all the way** to zero, so that the Lasso estimator is **exactly** zero.

Thus, the Lasso estimator sets some – many – of the *β*'s *exactly* to 0

More on Lasso (1 of 2)

Lasso sets some – many – of the *β*'s *exactly* to 0

- This property gives the Lasso its name: the **L**east **A**bsolute **S**election and **S**hrinkage **O**perator. Selection, because it selects a subset of the predictors to use for prediction – and drops the rest.
- This feature means that Lasso can work especially well when in reality many of the predictors are irrelevant.
- Models in which most of the *true β*'s are zero – that is, in which E(*Y*|*X*) only depends on just a few *X*'s – are called **sparse**.
- Lasso produces sparse models, and works well when the population model is in fact sparse.

More on Lasso (2 of 2)

- Lasso has another unusual property: the estimated model, and selected variables, depends on how the variables are specified.
- For example, if model A uses the dummy variables Freshman, Sophomore, Junior (and omits senior since they are deviated from their means); and model B uses Freshman, ≤Sophomore, and ≤Junior, then Lasso will in general give different predictions for models A and B, although OLS (and ridge) will give the same predictions.
- Technically, Lasso predictions are not invariant to linear transformations of the regressors

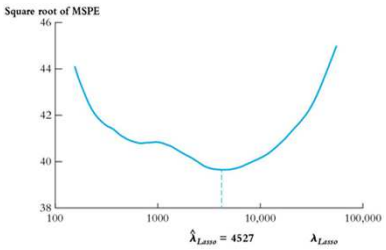
Predicting School-level test scores: Lasso

λ_{Lasso} is estimated by minimizing the 10-fold cross-validated MSPE

The resulting estimate of the shrinkage parameter is 4527

Root MSPE's:
OLS: 78.2
Lasso: 39.7

$k = 817$, $n = 1966$



The Lasso estimator retains only 56 of the 817 predictors.
Like ridge, Lasso cuts the square root of the MSPE in half, compared to OLS!

5. Principal Components

- Ridge and Lasso reduce the MSPE by shrinking (biasing) the estimated coefficients to zero – and in the case of Lasso, by eliminating many of the regressors entirely.
- Instead, **Principal components regression** collapses the very many predictors into a much smaller number ($p << k$) of linear combinations of the predictors
- These the linear combinations – called the **principal components of *X*** – are computed so that they capture as much of the variation in the original *X*'s as possible.
- Because the number *p* of principal components is small, OLS can be used, with the principal components as (new) regressors.

Principal Components in Pictures, $k = 2$

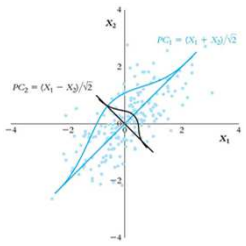
Suppose you have 2 X's, and you want to choose a linear combination of those Xs (say, $aX_1 + bX_2$) that captures as much of the variation of the X's as possible in a single summary variable. What values of a and b would you use?

The Principal Components solution is to choose a and b to solve,

$$\max \text{var}(aX_1 + bX_2), \text{ subject to } a^2 + b^2 = 1$$

For 2 X's that are positively correlated, the resulting choices of a and b are $a = b = 1/\sqrt{2}$

This is shown in the figure -->



Principal Components, $k > 2$

For $k > 2$ X's, the principal components are the linear combinations of the X's that have the greatest variance *and* that are uncorrelated with the previous principal components.

So, the jth principal component PC_j, solves,

$$\max \text{var}\left(\sum_{i=1}^{k} a_{ji} X_i\right), \text{ subject to } \sum_{i=1}^{k} a_{ji}^2$$

and subject PC_j to being uncorrelated with $PC_1, \ldots, PC_{j-1}$.

The first *p* principal components are the linear combinations of *X* that capture as much of the variation in *X* as possible.

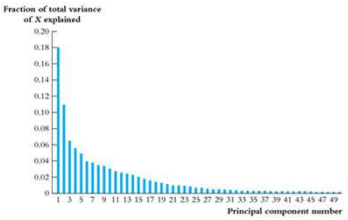
Principal Components as Data Compression

- Principal components can be thought of as a data compression tool, so that the compressed data have fewer regressors with as little information loss as possible.
- Data compression is used all the time to reduce very large data sets to smaller ones. A familiar example is image compression, where the goal is to retain as many of the features of the image (photograph) as possible, while reducing the file size.
- In fact, many data compression algorithms build on or are cousins of principal components analysis.

How many Principal Components? (1 of 2)

One way to choose p is to plot the increase in the average R^2 resulting from adding the pth principal components to a regression of X on $PC_1,\ldots,PC_{p-1}$.

This plot is known as a **scree plot**. Here is the scree plot for the school test score data set □



- The first principal component explains 18% of the variation in the 817 X's!
- The first 10 PC's explain 63% of the variation in the 817 X's!
- Still, it is rather hard to know where to draw the line…

How many Principal Components? (2 of 2)

The scree plot is informative (you should look at it) but doesn't provide a simple rule for choosing p.

- The number of principal components p is like the ridge and Lasso penalty factors λ_{Ridge} and λ_{Lasso} - all are additional parameters needed to implement the procedure.
- Like λ_{Ridge} and λ_{Lasso}, p can be estimated by minimizing the m-fold cross validated estimate of the MSPE.
  - For a given value of p, the principal components forecast is obtained by regressing Y on $PC_1,\ldots,PC_{p-1}$ using the estimation sample, then using that model to predict in the test sample

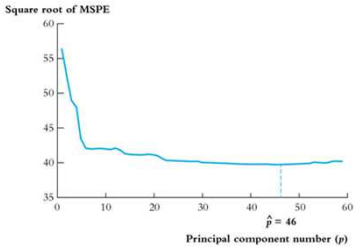
Predicting School-level test scores: Principal Components

k = 817, n = 1966

p is estimated by minimizing the 10-fold cross-validated MSPE

The resulting estimate is 46

Root MSPE's:
OLS: 78.2
Principal Components: 39.7



- Principal Components collapses the 817 predictors to 46.
- Like ridge and Lasso, PC cuts the square root of the MSPE in half, compared to OLS!

6. Application to School Test Scores

Data set

- Half the observations (1966) used for model estimation including estimation of λ_{Ridge}, λ_{Lasso}, and p.
- The other half is reserved for an out-of-sample test, comparing the various forecasts
- Three sets of predictors are used:
  - **Small** (k = 4): Student-teacher ratio, median local income, teacher's average years of experience, instructional expenditures per student
  - **Large** (k = 817): The regressors used up to now
  - **Very large** (k = 2065): Additional school and demographic variables, squares and cubes, and interactions.
    - For the large data set, $k > n$!!

Out-of-sample performance of predictive models for School Test Scores (1 of 5)

Predictor Set	OLS	Ridge Regression	Lasso	Principal Components
Small (k = 4)				
Estimated λ or p	–	–	–	–
In-sample root MSPE	53.6	–	–	–
Out-of-sample root MSPE	52.9			
Large (k = 817)				
Estimated λ or p	–	2233	4527	46
In-sample root MSPE	78.2	39.5	39.7	39.7
Out-of-sample root MSPE	64.4	38.9	39.1	39.5
Very large (k = 2065)				
Estimated λ or p	–	3362	4221	69
In-sample root MSPE	–	39.2	39.2	39.6
Out-of-sample root MSPE	–	39.0	39.1	39.6

1. OLS gets worse with more predictors – and you can't even run OLS when $k > n$

Out-of-sample performance of predictive models for School Test Scores (2 of 5)

Predictor Set	OLS	Ridge Regression	Lasso	Principal Components
Small (k = 4)				
Estimated λ or p	–			
In-sample root MSPE	53.6			
Out-of-sample root MSPE	52.9			
Large (k = 817)				
Estimated λ or p	–	2233	4527	46
In-sample root MSPE	78.2	39.5	39.7	39.7
Out-of-sample root MSPE	64.4	38.9	39.1	39.5
Very large (k = 2065)				
Estimated λ or p	–	3362	4221	69
In-sample root MSPE	–	39.2	39.2	39.6
Out-of-sample root MSPE	–	39.0	39.1	39.6

2. The cross-validated MSPE, computed with the estimation sample, is a good estimate of the out-of-sample MSPE

Out-of-sample performance of predictive models for School Test Scores (3 of 5)

Predictor Set	OLS	Ridge Regression	Lasso	Principal Components
Small (k = 4)				
Estimated λ or p	–			
In-sample root MSPE	53.6			
Out-of-sample root MSPE	52.9			
Large (k = 817)				
Estimated λ or p	–	2233	4527	46
In-sample root MSPE	78.2	39.5	39.7	39.7
Out-of-sample root MSPE	64.4	38.9	39.1	39.5
Very large (k = 2065)				
Estimated λ or p	–	3362	4221	69
In-sample root MSPE	–	39.2	39.2	39.6
Out-of-sample root MSPE	–	39.0	39.1	39.6

3. Lasso, Ridge, and PC all provide big improvements over OLS

Out-of-sample performance of predictive models for School Test Scores (4 of 5)

Predictor Set	OLS	Ridge Regression	Lasso	Principal Components
Small (k = 4)				
Estimated λ or p	–			
In-sample root MSPE	53.6			
Out-of-sample root MSPE	52.9			
Large (k = 817)				
Estimated λ or p	–	2233	4527	46
In-sample root MSPE	78.2	39.5	39.7	39.7
Out-of-sample root MSPE	64.4	38.9	39.1	39.5
Very large (k = 2065)				
Estimated λ or p	–	3362	4221	69
In-sample root MSPE	–	39.2	39.2	39.6
Out-of-sample root MSPE	–	39.0	39.1	39.6

4. For these data, Ridge, Lasso, and PC have very similar out-of-sample MSPEs – however this will not be true in general.
- For these data, Ridge has a very slight edge

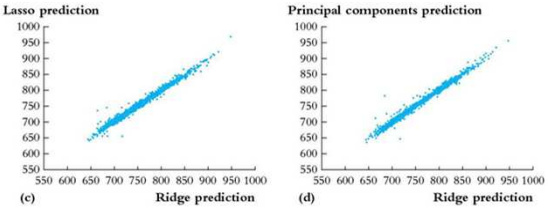
Out-of-sample performance of predictive models for School Test Scores (5 of 5)

Predictor Set	OLS	Ridge Regression	Lasso	Principal Components
Small (k = 4)				
Estimated λ or p	–			
In-sample root MSPE	53.6			
Out-of-sample root MSPE	52.9			
Large (k = 817)				
Estimated λ or p	–	2233	4527	46
In-sample root MSPE	78.2	39.5	39.7	39.7
Out-of-sample root MSPE	64.4	38.9	39.1	39.5
Very large (k = 2065)				
Estimated λ or p	–	3362	4221	69
In-sample root MSPE	–	39.2	39.2	39.6
Out-of-sample root MSPE	–	39.0	39.1	39.6

5. For these data, there isn't much gain to using the very large data set, however this will not be true in general.

Out-of-sample performance of predictive models for School Test Scores

For these data, the predictions made by Ridge, Lasso, and principal components are similar to each other (as shown in these scatterplots), however they are quite different from the (worse) OLS predictions.



7. Summary (1 of 2)

- With many predictors, OLS will produce poor out-of-sample predictions.
- By introducing the right type of bias – shrinkage towards zero – the variance of the prediction can be reduced by enough to offset the bias and result in smaller MSPE.
- Ridge and Lasso reduce the MSPE by shrinking (biasing) the estimated coefficients to zero – and in the case of Lasso, by eliminating many of the regressors entirely.
- Principal components collapses X into fewer uncorrelated linear combinations that capture as much of the variation of the X's as possible. Predictions are then made using the OLS regression of Y on the principal components.

Summary (2 of 2)

- All three methods require an additional parameter: λ_{Ridge} for Ridge, λ_{Lasso} for Lasso, and p for principal components. This parameter can be estimated by minimizing the m-fold cross-validated estimate of the MSPE.
- The different methods have strengths in different situations, and which works best in a given application is an empirical question.