BIG DATA and MACHINE LEARNING in Econometrics

Anna Simoni[2]

[2]CREST, CNRS - ENSAE and École Polytechnique

Outline

Economics and Big Data

- Economists make use of newly available large-scale administrative data or private sector data (often obtained through collaborations with private firms), giving rise to new opportunities and challenges.

- Administrative data : universal or near-universal population coverage.

- These data opportunities also raise some important challenges : (1) developing methods for researchers to access and explore data in ways that respect privacy and confidentiality concerns ; (2) developing the appropriate data management and programming capabilities ; (3) designing approaches to summarize, describe, and analyze large-scale and relatively unstructured data sets.

Economics and Big Data

- Several features differentiate modern data sets from data used in earlier research :
  1. data are now often available in real time ;
  2. that data are available on previously unmeasured activities (personal communications, social networks, search and information gathering, and geolocation data) ;
  3. data come with less structure.

- Figuring out how to organize and reduce the dimensionality of largescale, unstructured data is becoming a crucial challenge in empirical economic research.

- Economic models are useful for analyzing big data sets (*e.g.* the design of online advertising auctions and exchanges).

Public sector data : Administrative records

- In the course of administering the tax system, social programs, and regulation, the government collects highly detailed data on individuals and corporations, education, social insurance, and local government spending.

- Administrative data offer several advantages over traditional survey data : high data quality (no missingness), long-term panel structure. Examples :
  - tax data allow for the creation of relatively homogeneous time series spanning many decades (high incomes no longer under-reported), see Piketty and Saez ;
  - useful in documenting regional disparities in economic mobility and health care spending, in identifying the sizable differences in wages and productivity across otherwise similar firms.

  $\Rightarrow$ have helped to guide policy discussions.

- Administrative data also have a high value if used for causal inference and policy evaluation.

Private sector data : Collection and collaborations. I

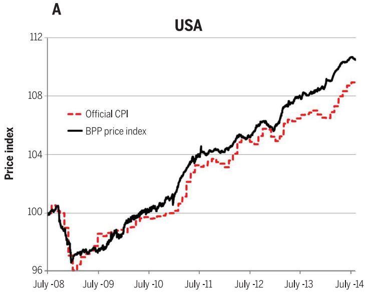
- Vast amount of information collected by Internet companies such as Google, Amazon, and Facebook, . . .

- but also firms in every sector of the economy routinely collect and aggregate data on their customers and their internal businesses. Examples : banks, credit card companies, and insurers ; retailers such as Walmart, Monoprix, . . . ; private companies that specialize in data aggregation (*e.g.* credit bureaus or marketing companies).

- One potential application of private sector data is to create statistics on aggregate economic activity that can be used : 1) to track the economy or as inputs to other research, 2) for understanding firm investment decisions and macroeconomic activity.

- A second application of private data is to allow researchers to look inside specific firms or markets to study employee or consumer behavior or the operation of different industries.

Private sector data : the example of The Billion Prices Project



Private sector data : Collection and collaborations. II

- Relative to administrative data, company data have some important differences :
  - Sampling usually is not representative.
  - Data collection emphasizes recency and relevance for business use.
  - Private entities are not bound by some of the bureaucratic constraints that limit public agencies (more detailed data, the computing resources can be more powerful, and private companies can have more flexibility to run experiments).
- These highly granular data are often used to find targeted variation that plausibly allows for causal estimates (*e.g.* estimates of the effects of sales tax collection, pricing changes, . . .).
- Large-scale granular data can also be particularly useful for assessing the robustness of identifying assumptions.

Outline

Statistical learning and Big Data

- Statistical learning : set of tools for modeling and understanding complex datasets.
- It is an area in statistics and involves parallel developments in computer science and machine learning.
- The field encompasses many methods such as the lasso and sparse regression, classification and regression trees, and boosting and support vector machines.
- With the explosion of “Big Data” problems, statistical learning has become a very hot field in many scientific areas as well as marketing, finance, and other business disciplines.
- In this class : we analyze the particular features of “Big Data” in economics and econometrics $\Rightarrow$ detection of relationships in the data vs. prediction (machine learning) and summarization (data mining).

Definitions : data sets

- Larger data become more and more available.
- n= number of observations ; p= number of variables
- "Classical statistics/ econometrics" : big n, small p (*tall* data) ; $\Rightarrow$ standard theory, computational demanding
- "High-dimensional data" or "Big Data" : $n \ll p$ (small n, big p ; *fat* data) ; $\Rightarrow$ non-standard theory, computational demanding
- Conventional statistical and econometric techniques (*e.g.* regression) often work well but there are issues unique to big datasets that may require different tools :
  - more powerful data manipulation tools,
  - variable selection tools,
  - tools to model more flexible relationships than simple linear models.

Definitions : The statistical prediction problem

- Inputs X= measured or present variables. Synonyms : predictors, features or independent variables

- These inputs have some influence on one or more outputs.

- Y= output variable, also called response or dependent variable or outcome variables. It can be *quantitative* or *qualitative*.

- We wish to predict Y based on X :

$$Y = f(X) + \epsilon$$

- $f(\cdot)$ unknown function, $X = (X_1, \dots, X_p)$, p predictor variables, ϵ= random error term.

- We have a training set of data, in which we observe the outcome and feature measurements for a set of objects (such as people).

- We look for a "good" prediction of Y given a new value of X ("good" means it minimizes some loss function associated with new out-of-sample observations of X).

Definitions : supervised and unsupervised learning

- Supervised Learning : Presence of the outcome variable to guide the learning process.

  Goal : e.g. to use the inputs to predict the values of the outputs.

  Methods : regression methods (linear, lasso, ridge, etc.), bagging, trees, random forests, ensemble learning, . . .

- Unsupervised Learning : only features are observed, no measurements of the outcome variable.

  Goal : describe how the data are organized or clustered.

  Methods : Association Rules, PCA, cluster analysis

Definitions : regression vs. classification

The distinction in output type Y (quantitative vs. qualitative) has led to a naming convention for the prediction tasks.

- Regression : we predict quantitative output.
- Classification : we predict qualitative output (categorical / discrete).
- Coding of qualitative variables : $0/1$, $-1/+1$, or in general case via dummy variables (*i.e.* when there are more than two categories, several alternatives are available).
- Also input variables can be of different type : we can have some of each of qualitative and quantitative input variables.

Definitions : prediction and inference

- Prediction : Given inputs X, but not the output Y, we want to predict Y :

$$\hat{Y} = \hat{f}(X).$$

  We are interested in high quality predictions and not in the function f which is considered as given.

- Inference : the goal is to understand the relationship between Y and X and the form of f (testing and confidence estimation). Related questions are which predictors are associated with the response (model selection) and is the relationship linear or nonlinear.

Definitions : Prediction Accuracy vs. Model Interpretability

- Some methods are less flexible or more restrictive, *i.e.* the range of shapes of f they can estimate is restricted.
- Other methods are more flexible in the shape of f.
- Usually there is a trade-off between prediction accuracy and interpretability : flexible models often deliver good prediction accuracy but give models which are harder to interpret.

Definitions : the Bias-Variance trade-off. I

- The *training* mean squared error (MSE) is the most commonly used measure of quality of fit in regression analysis. It is defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2.$$

- The MSE is computed using the training data that was used to fit the model.
- But in general, we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data :
  - suppose that we fit our statistical method on our training observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ and we obtain the estimate $\hat{f}$;
  - then, we can compute $\hat{f}(x_0)$ and see whether $\hat{f}(x_0) \approx y_0$ where (x_0, y_0) is a previously unseen test observation not used to estimate the statistical method.

  Hence, we want to choose the method that gives the lowest *test* MSE (as opposed to the lowest *training* MSE) and we could compute the average squared prediction error for these test observations :

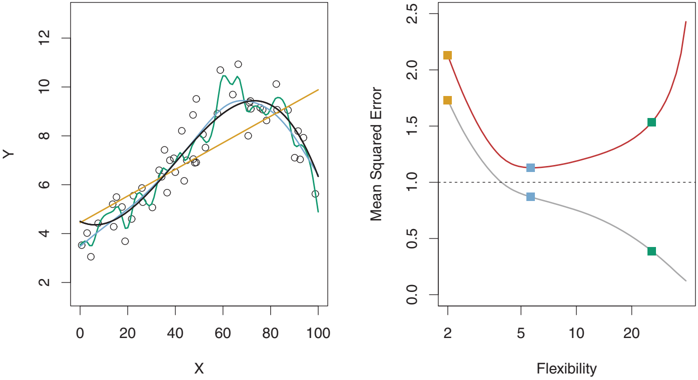
$$Ave(y_0 - \hat{f}(x_0))^2.$$

  We'd like to select the model for which the this quantity is as small as possible.

Definitions : the Bias-Variance trade-off. II

- Many statistical methods estimate coefficients so as to minimize the training MSE. For these methods, the training MSE can be quite small, but the test MSE is often much larger.
- When a given method yields a small *training* MSE but a large *test* MSE, we are said to be overfitting the data.
- Cross-validation is a method for estimating *test* MSE using the training data and can be used in practice to estimate the minimum of the flexibility level.

The Bias-Variance trade-off. III



Definitions : the Bias-Variance trade-off. IV

- The U-shape observed in the test MSE curve is the result of two competing properties of statistical learning methods : variance and bias.

- The expected test MSE, for a given value x_0, can always be decomposed into the sum of 3 quantities : the $Var(\hat{f}(x_0))$, the squared bias of $\hat{f}(x_0)$ and $Var(\epsilon)$:

$$\mathbf{E}(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon).$$

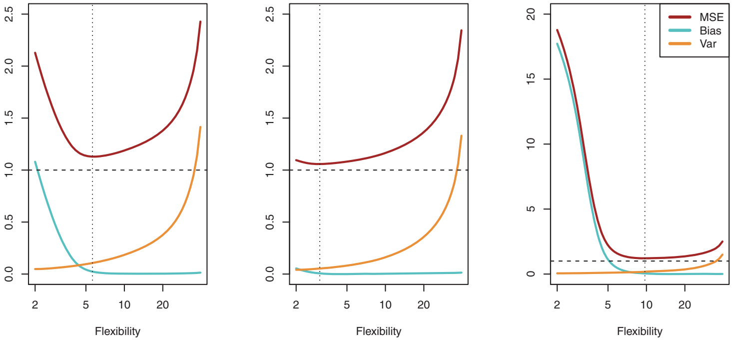
- Variance = amount by which $\hat{f}$ would change if we estimated it using a different training data set. If a method has high variance then small changes in the training data can result in large changes in $\hat{f}$.

- Bias = error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.

- In general, more flexible methods result in less bias and higher variance. The relative rate of change of these two quantities determines whether the test MSE increases or decreases.

- The challenge lies in finding a method for which both the variance and the squared bias are low.

The Bias-Variance trade-off. V



Definitions : causal inference

Prediction and causal inference are distinct (though closely related) problems. Causal questions :

- What is the causal relationship of interest ? (most interesting research in social science).
- What would happen if a policy-maker/firm . . . changes a policy ?
- Rubin causal model or potential outcome framework : one postulates the existence of two potential outcomes for each unit, with and without the treatment (Rubin 1974).
- "The critical step in any causal analysis is estimating the counterfactual - a prediction of what would have happened in the absence of the treatment. The powerful techniques used in machine learning may be useful for developing better estimates of the counterfactual, potentially improving causal inference." Varian 2016

Example in marketing (Varian 2015) I

- Data on ad spend and product sales in various cities.
- We want to predict how sales would respond to a contemplated change in ad spend.
- y_c= per capita sales in city c and x_c= per capita ad spend in city c. One can run the regression
$$y_c = bx_c + e_c.$$
- Such a regression is unlikely to provide a satisfactory estimate of the *causal effect of ad spend on sales*. Why ?
  - Suppose that y_c are per capita box office receipts for a movie about surfing and x_c are per capita television ads for that movie. There are only two cities in the dataset : Honolulu, Hawaii and Fargo, North Dakota.
  - Suppose that with these data we estimate the model : $y_c = 10x_c$.

Example in marketing (Varian 2015) II

- Problem : there is an omitted variable in our regression, which we may call “interest in surfing”. Interest in surfing is high in Honolulu and low in Fargo.
- Moreover,the marketing executives that determine ad spend presumably know this, and they choose to advertise more where interest is high and less where it is low.
- Therefore, this omitted variable - interest in surfing - affects both y_c and x_c. Such a variable is called a “confounding variable”.

- If we are primarily interested in predicting sales as a function of spend, and the advertiser’s behavior remains constant, the simple regression $y_c = bx_c + e_c$. is fine.

- However, usually a prediction of past behavior is not the goal : one wants to know how box office receipts would respond to a change in the advertiser’s behavior.

High-Dimensional models in Econometrics for causal inference. I

- High dimensional sparse (HDS) regression models in econometrics allows for a large number of regressors, $p \gg n \ldots$

- and imposes that the model is sparse : only $s \ll n$ of these regressors are important for capturing the main features of the regression function :

$$
\begin{aligned}
Y &= X'\beta_0 + \epsilon, \qquad \beta_0 \in \mathbb{R}^p \\
T &= \text{support}(\beta_0) \text{ has } s \text{ elements where } s < n
\end{aligned}
$$

  and $p > n$ is allowed. T is unknown.

- Sparsity can be motivated on economic grounds in situations where a researcher believes that the economic outcome could be well-predicted by a small (relative to the sample size) number of factors but is unsure about the identity of the relevant factors.

High-Dimensional models in Econometrics for causal inference. II

- The motivation for considering HDS models comes in part from the wide availability of data sets with many regressors :
  - the American Housing Survey records prices as well as a multitude of features of houses sold ;
  - scanner data-sets record prices and numerous characteristics of products sold at a store or on the internet.
- HDS models are also partly motivated by the use of series methods in econometrics : they use many constructed or series regressors - regressors formed as transformation of elementary regressors - to approximate regression functions. In these applications, it is important to have parsimonious yet accurate approximation of the regression function.

High-Dimensional models in Econometrics for causal inference. III

Example : returns to schooling (Angrist and Krueger (1991) Data)

$$
\begin{aligned}
Y_1 &= \theta_1 Y_2 + \gamma' W + U, \qquad \mathbf{E}[U|W,Z] = 0 && (1) \\
Y_2 &= \beta' Z + \delta' W + V, \qquad \mathbf{E}[V|W,Z] = 0 && (2)
\end{aligned}
$$

where

- Y_1 = log(wage) ;
- Y_2 = education ;
- W = a vector of control variables,
- Z = a vector of instrumental variables that affect education but do not directly affect the wage.

High-Dimensional models in Econometrics for causal inference. IV

Dataset :

- drawn from the 1980 U.S. Census and consist of $329,509$ men born between 1930 and 1939.
- W = set of 510 variables : a constant, 9 year-of-birth dummies, 50 state-of-birth dummies, and 450 state-of-birth $\times$ year-of-birth interactions.
- Z = three quarter-of-birth dummies and interactions of these quarter-of-birth dummies with the set of state-of-birth and year-of-birth controls in W giving a total of 1530 potential instruments.
- Angrist and Krueger (1991) discusses the endogeneity of schooling in the wage equation and provides an argument for the validity of Z as instruments based on compulsory schooling laws and the shape of the life-cycle earnings profile.
- The coefficient of interest is θ_1, which summarizes the causal impact of education on earnings.

Using sparse methods for the first-stage estimation offers an option for estimating θ_1 : only $s \ll n$ elements of β_1 are nonzero but the identities of these elements is unknown.

Outline

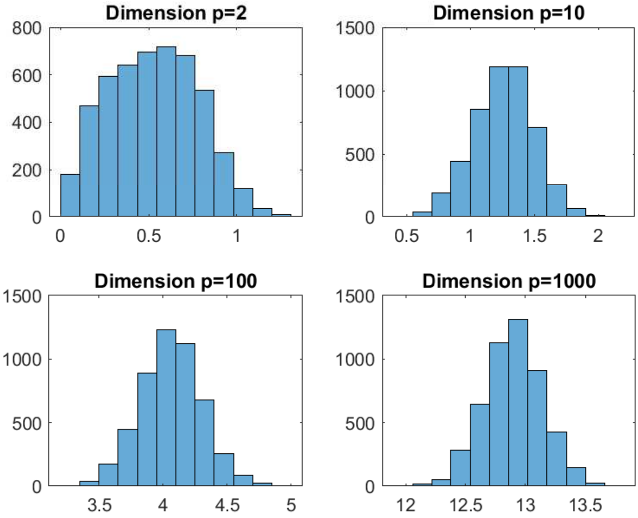
Problems / Challenges in High-Dimensions

- Lost in the immensity of high-dimensional spaces (curse of dimensionality).
- Fluctuations cumulate.
- A simple LS regression cannot be used : regardless of whether or not there truly is a relationship between X and Y, LS will yield a set of coefficient estimates that result in a perfect fit to the data (the residuals are zero).
- Computational complexity

Immensity of high-dimensional spaces

When the dimension p increases, the notion of “nearest points” vanishes. Below the histograms of the pairwise distances of n = 100 points randomly drawn (uniformly) from the unit cube are given.

Immensity of high-dimensional spaces



Immensity of high-dimensional spaces

```
close all
clear all

s=0;
for d = [2 10 100 1000]
X = rand(100,d);
dist_X = zeros(size(X,1)^2,1);
for i=1:size(X,1)
    for j=1:size(X,1)
        if i > j
            dist_X(size(X,1)*(i-1)+i+j-1,1) = sqrt((X(i,:)-X(j,:))*(X(i,:)'-X(j,:)'));
        end
    end
end
dist_X(dist_X==0)=[];
s=s+1;
%figure (1)
figure(1)
subplot(2,2,s);
histogram(dist_X,12);
title('Dimension p=')
end
```

Fluctuations accumulate

In the linear regression model $Y = X\beta + \varepsilon$ for the OLS estimate $\hat{\beta} = (X^T X)^{-1} X^T Y$ we have

$$\mathbf{E}\|\hat{\beta} - \beta\|^2 = \mathbf{E}\|(X^T X)^{-1} X^T \varepsilon\|^2 = tr((X^T X)^{-1})\sigma^2$$

where $Var(\varepsilon) = \sigma^2$.

In the case of orthogonal design :

$$\mathbf{E}\|\hat{\beta} - \beta\|^2 = p\sigma^2.$$

Hence the estimation error grows with the dimension p of the problem.

LS regression does not work

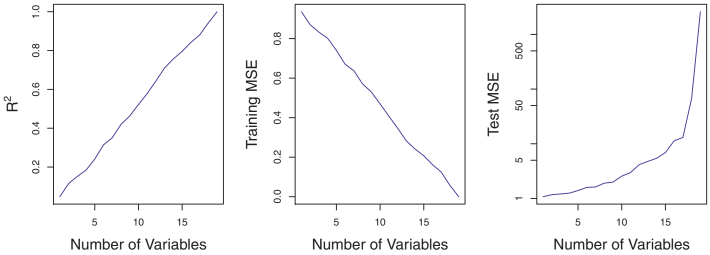


FIGURE 6.23. *On a simulated example with $n = 20$ training observations, features that are completely unrelated to the outcome are added to the model.* Left: *The R^2 increases to 1 as more features are included.* Center: *The training set MSE decreases to 0 as more features are included.* Right: *The test set MSE increases as more features are included.*

LS regression does not work I

- Someone who carelessly examines only the R_2 or the training set MSE might erroneously conclude that the model with the greatest number of variables is best.

- Methods as Ridge regression, the Lasso, and principal components regression, are particularly useful for performing regression in the high-dimensional setting : they avoid overfitting by using a less flexible fitting approach than LS.

- Three important points :
  1. regularization or shrinkage plays a key role in high-dimensional problems,
  2. appropriate tuning parameter selection is crucial for good predictive performance,
  3. the test MSE tends to $\uparrow$ as the dimensionality of the problem (*i.e.* p) $\uparrow$, unless the additional features are truly associated with the response (curse of dimensionality).

LS regression does not work II

- In the high-dimensional setting, the multicollinearity problem is extreme.
- One should never use *sum of squared errors*, *p-values*, R_2 on the training data as a measure of goodness of fit $\Rightarrow$ Instead, report results on an *independent test set*, or *cross-validation errors*.

Computational Complexity

With increasing dimension, numerical computations can become very demanding and exceed the available computing resources.

Example : When we have p potential regressors, than the number of submodels is $2p$ which grows exponentially with the number of regressors.