Big Data: Discussion

Christian Hansen

August 18, 2015

What is “big data”?

Big data:

1. Large data sets: e.g. n observations, p variables, np too big to fit on a computer
2. High-dimensional models: n observations, p parameters, $n \approx p$ or $n \ll p$ [traditional nonparametrics; text data, big survey data sets; flexible “parametric models”, semiparametric models]
3. Statistical Learning/Data-mining: Exploit data to get good forecasting rules

Why the hubbub about “big data”?

Why might econometricians/economists care about “big data”?

- Data is getting bigger

Why the hubbub about “big data”?

Why might econometricians/economists care about “big data”?

- Data is getting bigger
- Want to learn “small” effects

Why the hubbub about “big data”?

Why might econometricians/economists care about “big data”?

- Data is getting bigger
- Want to learn “small” effects
  - e.g. Taddy et al. “A nonparametric Bayesian analysis of heterogeneous treatment effects in digital experimentation” - RCT, 21 million observations, treatment: 96 pixel picture in control, 140 pixel picture in treated

Why the hubbub about “big data”?

Why might econometricians/economists care about “big data”?

- Data is getting bigger
- Want to learn “small” effects
- Want to learn complex models with interesting data

Why the hubbub about “big data”?

Why might econometricians/economists care about “big data”?

- Data is getting bigger
- Want to learn “small” effects
- Want to learn complex models with interesting data
  - Curse of dimensionality

Why the hubbub about “big data”?

Why might econometricians/economists care about “big data”?

- Data is getting bigger
- Want to learn “small” effects
- Want to learn complex models with interesting data
- Want a framework and methods that face trade-offs made in real data analysis

Why the hubbub about “big data”?

Why might econometricians/economists care about “big data”?

- Data is getting bigger
- Want to learn “small” effects
- Want to learn complex models with interesting data
- Want a framework and methods that face trade-offs made in real data analysis
  - Traditional statistical approach that posits a low-dimensional fixed model (or set of models) seems unrealistic
    - Multiple testing, model choice/specification search, regularization ...

Why the hubbub about “big data”?

Why might econometricians/economists care about “big data”?

- Data is getting bigger
- Want to learn “small” effects
- Want to learn complex models with interesting data
- Want a framework and methods that face trade-offs made in real data analysis
  - Traditional statistical approach that posits a low-dimensional fixed model (or set of models) seems unrealistic
    - Multiple testing, model choice/specification search, regularization ...
  - History in econometrics: many instruments, incidental parameters, non- and semi-parametrics, partial identification ...

Why the hubbub about “big data”?

Why might econometricians/economists care about “big data”?

- Data is getting bigger
- Want to learn “small” effects
- Want to learn complex models with interesting data
- Want a framework and methods that face trade-offs made in real data analysis
  - Traditional statistical approach that posits a low-dimensional fixed model (or set of models) seems unrealistic
    - Multiple testing, model choice/specification search, regularization ...
  - History in econometrics: many instruments, incidental parameters, non- and semi-parametrics, partial identification ...
  - Statistical vs computational efficiency

Why the hubbub about “big data”?

Why might econometricians/economists care about “big data”?

- Data is getting bigger
- Want to learn “small” effects
- Want to learn complex models with interesting data
- Want a framework and methods that face trade-offs made in real data analysis
  - Traditional statistical approach that posits a low-dimensional fixed model (or set of models) seems unrealistic
    - Multiple testing, model choice/specification search, regularization ...
  - History in econometrics: many instruments, incidental parameters, non- and semi-parametrics, partial identification ...
  - Statistical vs computational efficiency

Fundamentally, theory and methods for “big data” are about efficient use of data facing real trade-offs while trying to be honest about the trade-offs - Your data are probably “big data”!

A Simple Example

$$y_i = d_i\alpha + x_i\beta + \epsilon_i, \quad d_i = x_i\gamma + v_i$$

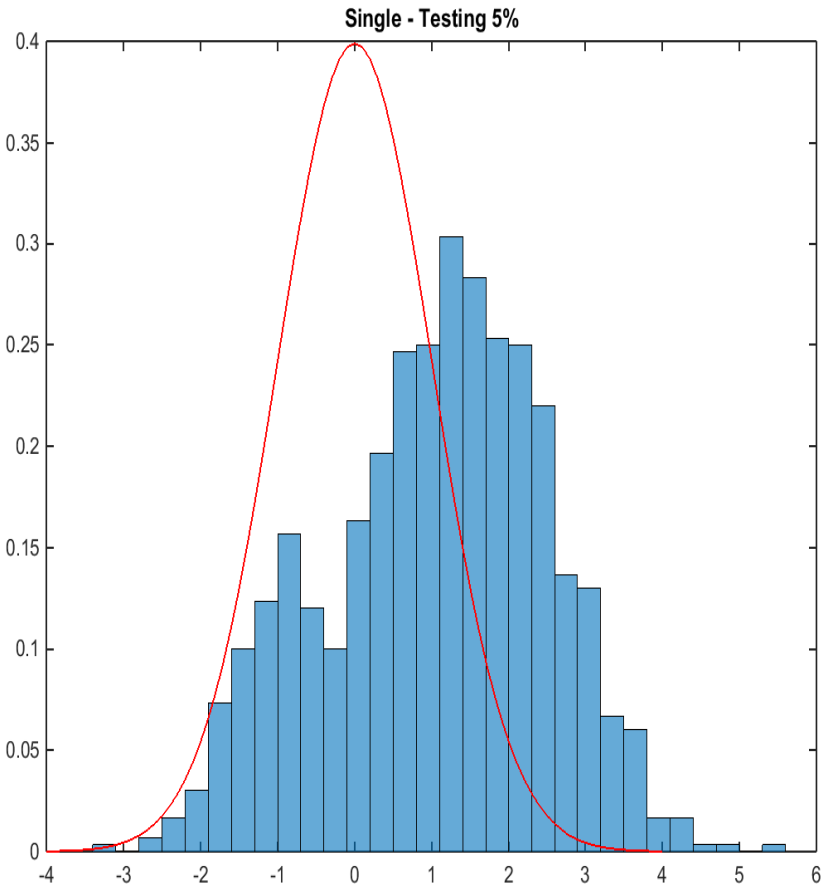
$$\alpha = \mathbf{0}, \quad \beta = .\mathbf{2}, \quad \gamma = .8,$$

$$n = 100$$

$$\epsilon_i \sim N(0,1)$$

$$(d_i, x_i) \sim N\left(0, \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix}\right)$$

- selection done by a **t-test**



Reject $H_0 : \alpha = 0$ (the truth) about 50% of the time for 5% level test

Challenges of “big data” - Inference

Inferential challenges:

- Inference following model selection/multiple testing/regularization is hard
  - (Valid) inference impossible in interesting setting without strong restrictions
  - What structures are we willing to buy? What learning is possible under these structures? What methods are appropriate for these structures?
    - Pretty well-understood under (i) (approximate) sparsity, (ii) smoothness and low-dimensional inputs. Is that enough?
  - Are there contexts where point forecasts are enough?
    - prediction (e.g. Kleinberg et al. “Prediction policy problems”)
    - inputs into other questions - semiparametric models, adaptive design of experiments

A Simple Example - Chernozhukov et al. approach

$$y_i = d_i\alpha + x_i\beta + \epsilon_i, \quad d_i = x_i\gamma + v_i$$

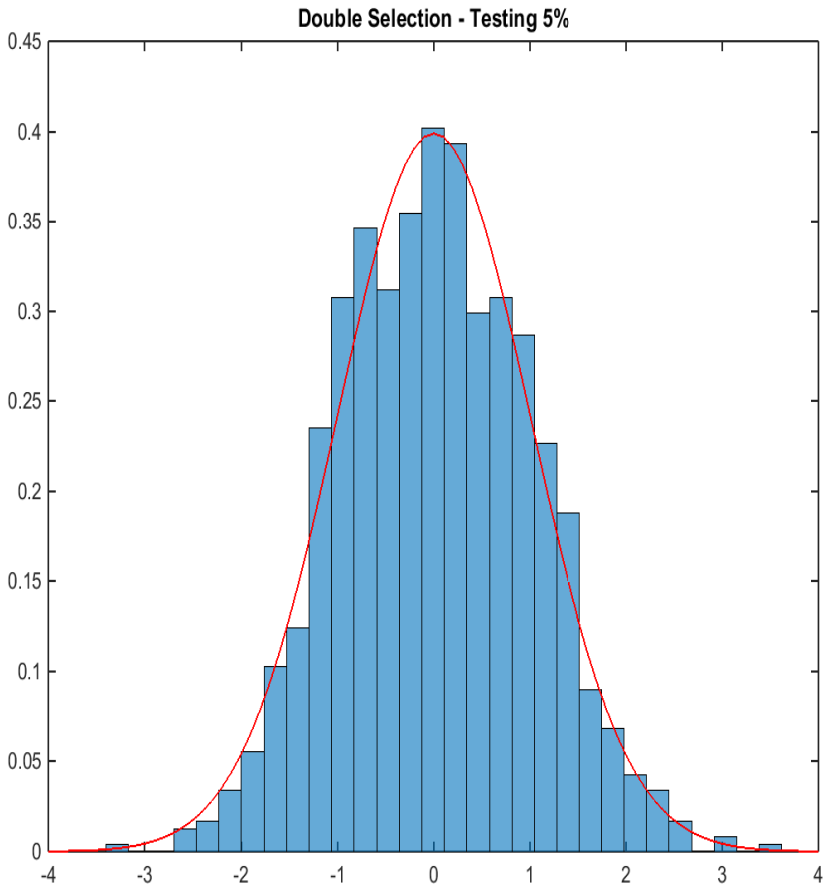
$$\alpha = \mathbf{0}, \quad \beta = .\mathbf{2}, \quad \gamma = .8,$$

$$n = 100$$

$$\epsilon_i \sim N(0,1)$$

$$(d_i, x_i) \sim N\left(0, \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix}\right)$$

- selection done by a **t-test**



Reject $H_0 : \alpha = 0$ (the truth) about 5% of the time for 5% level test

Challenges of “big data” - Computation

Computational challenges:

- Mechanical: Might need more computer science/programming understanding and more computing resources
- Theoretical: Statistical efficiency/computational efficiency tradeoff. Where do feasible algorithms fall?
  - e.g. Zhu and Lafferty “Quantized estimation of Gaussian sequence models in Euclidean balls;” Chandrasekarana and Jordan “Computation and statistical tradeoffs via convex relaxation;” Zhang et al. “Information-theoretic lower bounds for distributed statistical estimation with communication constraints”
- Theoretical: Sample splitting/subsampling
  - Managing information loss/optimal information discard/aggregation

Challenges of “big data” - Economic Data

Out-of-the-box methods from machine learning/stats are often for “simple” (e.g. iid Gaussian) data

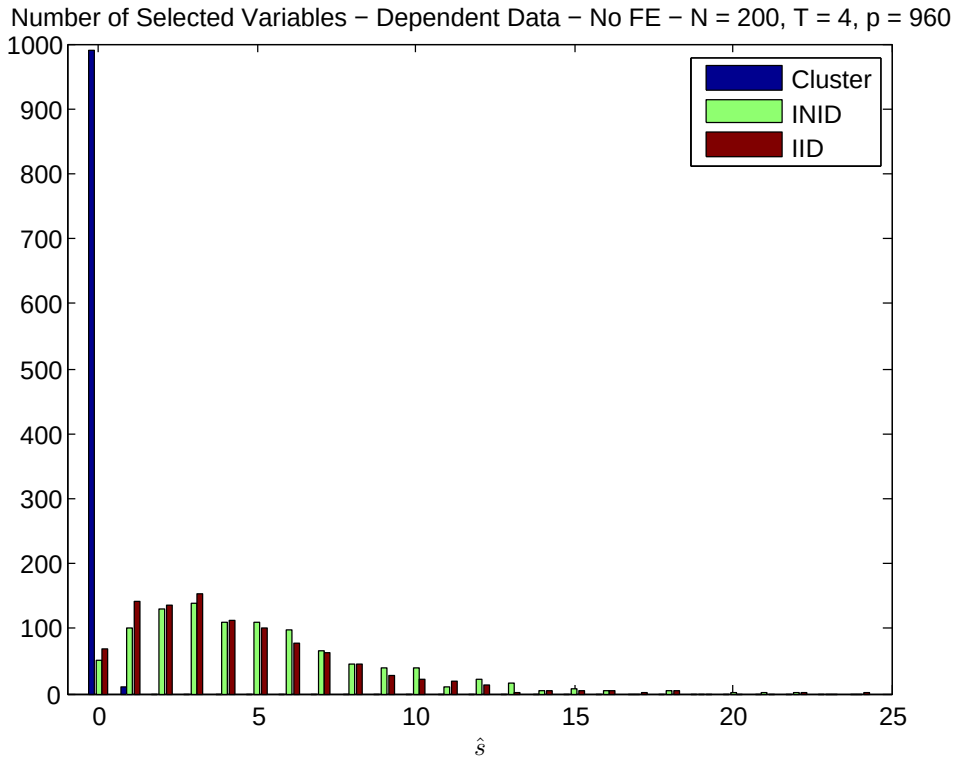
Economic data usually taken to be heterogeneous, possibly with complicated dependence, non-stationarity, etc.

- Seems likely that many methods will work with little or no modification but little theory and not deeply explored

Variable selection with Dependent Data

Number of selected variables in simulation where all coefficients are exactly zero.

- AR(1) Data ($\rho = .8$)
- $n = 200, T = 4, p = 960$



Challenges of “big data” - Economic Data

Out-of-the-box methods from machine learning/stats are often for “simple” (e.g. iid Gaussian) data

Economic data usually taken to be heterogeneous, possibly with complicated dependence, non-stationarity, etc.

- Seems likely that many methods will work with little or no modification but little theory and not deeply explored

Many economically interesting objects are not given by “reduced form objects” (conditional forecasts) and are not obtained as simple transformations of these objects (though many are)

- interplay between inferential issues, regularization, and computation seems delicate and interesting

Opinion: “Big data” is here to stay. Lots of opportunities for theoretical econometricians and empirical economists!