Review of High-Dimensional Methods and Inference on Structural and Treatment Effects

A Belloni, V Chernozhukov, and C Hansen - 2014

Ding Chih, Lin   Yi Ting, Lin

National Taipei University, Dept of Economics

June, 2020

Introduction

High-dimensional data:

- number of variables $p \gg$ sample size n
- potential variables formed by the set of small number of variables

Introduction (conti.)

Statistical methods for high-dimensional data:

- regularization methods, ...
    - LASSO
    - ...
- good at prediction
- incorrect conclusions when inference about parameter

Statisticians often make inference about model parameter

Introduction (conti.)

Main goal:

Provide an **overview** of modified methods to inference about model parameter with high-quality in approximately sparse linear model.

It's less math intensive compared the paper we read before.

國立臺北大學
National Taipei University

Approximately sparse regression models

$$y_i = g(w_i) + \varepsilon_i$$

- nonparametric $g(.)$ not discussed here
- treats $g(w_i)$ as a high-dimensional, approximately linear model

$$g(w_i) = \sum_{j=1}^{p} \beta_j x_{i,j} + \gamma_{p,i}$$

  - $p \gg n$
  - $\gamma_{p,i}$ is an approximation error
  - **approximate sparsity** of the high-dimensional linear model

Approximately sparse regression models (conti.)

Approximate sparsity:

- only s variables among all of x_j that have nonzero β_j
- $s \ll n$
- nonzero $\gamma_{p,i}$

Estimating the parameter of sparse linear model

Variant of the LASSO estimator, Belloni et al. (2012):

$$\hat{\beta} = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{i,j} b_j \right)^2 + \lambda \sum_{j=1}^{p} |b_j| \gamma_j$$

- λ controls the degree of penalization
- γ_j address heteroskedasticity, non-normality in model error, etc.
- $\hat{\beta}_j$ tend to be biased because of regularization

Casual inference not prediction (1)

Problem: LASSO are designed for forecasting, not for inference about parameter.

Solution: use Post-LASSO estimator or others

1. use LASSO to determine which variables can be dropped
2. use OLS to estimate the coefficients on the remaining variables

Causal inference not prediction (2)

Problem: model selection mistakes may occur

Solution: Develop inference procedures that are robust to such mistakes

1. doing model selection or regularization over **nuisance parts**
2. no model selection will be done with **parts of interest**
3. estimating equation for main parameters are *orthogonal* to changes in nuisance parts

Causal inference not prediction (2) conti.

- **Canonical instrumental variable model**:
  - Ng and Bas (2009)
  - Belloni, Chen, Chernozhukov and Hansen (2012)
  - Belloni, Chen, Chernozhukov and Hansen (2013)
- **Partially linear model**:
  - Belloni, Chen, Chernozhukov (2013)
  - Farrell (2013)
- **Inference about parameters after model selection is a topics of ongoing research.**

Casual inference not prediction (2) conti.

- One case in DML by Chernozhukov (2016)

  ... as the DML estimator can clearly be interpreted as a linear IV estimator, and to the more recent literature on debiased lasso in the context where g_0 is taken to be well approximated by as sparse linear combination of pre-specified function of X.

- Belloni et al. (2013, 2014a,b)
- Javanmard and Montanari (2014b)
- van De Geer et al. (2014)
- Zhang and Zhang (2014)

國立臺北大學 National Taipei University

Example 1: Selection among many IVs

Linear IV model:

$$y_i = \alpha d_i + \varepsilon_i$$

$$d_i = \boldsymbol{z}_i' \boldsymbol{\theta} + \gamma_i + v_i$$

- d_i: scalar endogenous variable of interest, $\mathbb{E}(\varepsilon_i v_i) \neq 0$
- $\boldsymbol{z}_i$: p-dimensional IVs, $p \gg n$
- γ_i: approximation error

Example 1: Estimating linear IV model

Belloni, Chen, Chernozhukov, and Hansen (2012)

1. select small number of instruments from z_i by LASSO
2. use conventional 2SLS to estimate α

Example 1: Estimating linear IV model (conti.)

Why the procedure proposed by Belloni et al. (2012) works?

- there's no selection over whether d_i will be included in model
- selection is limited in first stage $\Rightarrow$ predictive perspective
- the second-stage IV estimate is *orthogonal* to selection mistakes

Example 2: Selection Among Many Controls

Linear Model:

$$y_i = \alpha d_i + x_i'\theta_y + r_{yi} + \zeta_i$$

- α effect of trt on outcome, parameter of interest
- d_i treatment variable
- x_i p dimensional vector of controls, $p \gg n$
- r_{yi} approximation error

$$E[\zeta_i | d_i, x_i, r_{yi}] = 0$$

國立臺北大學 National Taipei University

Example 2: Naive Approach 1

Apply LASSO on the equation, excluding α from the penalty term

Problems

1. Delete high-correlated variables, if $\theta_y \neq 0$ then omitted-variables bias
2. Neglects the relationship between trt vars and control vars
3. Not a forecasting rule for y_i given d_i and x_i

Example 2: Naive Approach 1 (conti.)

Solution:

$$
\begin{aligned}
y_i &= x_i'(\alpha\theta_d + \theta_y) + (\alpha r_{di} + r_{yi}) + (\alpha\nu_i + \zeta_i) \\
&= x_i'\pi + r_{ci} + \epsilon_i
\end{aligned} \tag{1}
$$

$$
d_i = x_i'\theta_d + r_{di} + \nu_i \tag{2}
$$

where: $E[\nu_i | x_i, r_{di}] = 0$ and $E[\epsilon_i | x_i, r_{ci}] = 0$

Example 2: Naive Approach 2

Selecting vars with only one equation from solution of naive approach 1

Problems

1. Rely on no errors in the selection process
2. Select with Eq. (1), snatch out vars effective in predicting y_i, but not effective in predicting d_i
3. Select with Eq. (2), pick out variables effective in predicting d_i, but not effective in predicting y_i

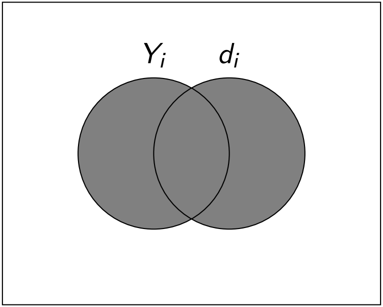
Omitted-variables Bias!

Example 2: Double Selection

Apply variable selection methods for **both** Eq.s

Estimate α by OLS y_i on d_i and **union** of both groups of selected variables.

Selected Variables



Example 2: Double Selection (conti.)

Virtues of the Double Selection Method

1. Discarded x_is are irrelevant of y_i and d_i
2. "Filter-out" concept of the DML article

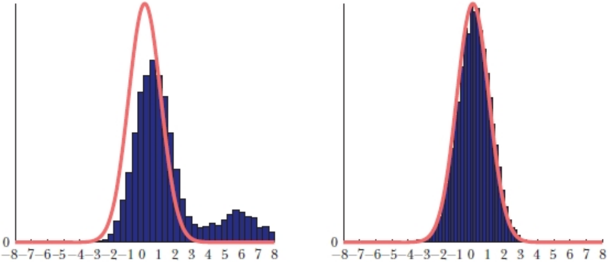
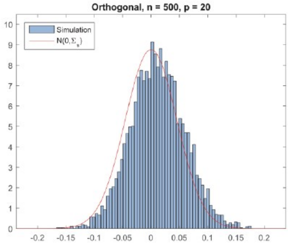


Figure: Naive 1 Selection (Left) V.S. Double Selection (Right)

Recall DML "Filter-out"

1. Predict Y and D using $\mathbf{X}$ obtained by ML method
2. Get residual $\widehat{U} = Y - \widehat{Y}$ and $\widehat{V} = D - \widehat{D}$
3. Build regression model of $\widehat{U}$ on $\widehat{V} \implies \breve{\theta}_0$

The "filter-out" concept is based on Frisch-Waugh-Lovell therom.



Conclusion

Double selection method in high-dimensional regression:

- Successfully perform dimension reduction
- Preclude omitted-variables bias

Follow ups:

Belloni, Chernozhukov, Fernandez-Val, and Hansen (2013):

valid inference when based on orthogonal estimating equations