Metrics and High-Dimensional Model Selection

Mirko Draca

Warwick University

21st October 2020

Metrics Meets Regularization

- Tools such as Lasso and Ridge Regression look useful, but how do we integrate them with what we know about model building in applied econometrics?
- What we'll look at here is the 'Double Machine Learning' approach outlined by Belloni, Chernozhukov, Hansen (hereafter BCH) and others in a broad programme of research.
- Then we'll move onto examples, both 'demonstrative' pieces from BCH that dissect previous work and new, original work using the Double ML approach.

Key Takeaways

- Human model selection and framing still matters. In particular, we still focus on key 'target' variables (usually denoted as d) that we do not subject to model selection.
- Naive approach of doing OLS 'post-Lasso' after dropping variables that were not selected in some initial Lasso exercise doesn't work in terms of good inference.
- This is because target variable d is exposed to omitted variable bias (OVB) problems via the correlated high dimensional $X's$. We therefore need to 'purge' d of this OVB using a partialling out procedure.

Key References.

There are two bits of material that you need to get a working knowledge of the Double ML approach.

- Belloni, A; Chernozhukov, V and Hansen, C (2014) 'High-Dimensional Methods and Inference on Structural and Treatment Effects'. *Journal of Economic Perspectives.* 28(2): 29-50.

- *NBER Summer Institute 2013 Econometrics Lectures.* 'Econometric Methods for High-Dimensional Data'.

Amongst the lectures, focus on those by Chernozhukov and Hansen. They give a deeper perspective than the JEP paper which is a broad overview. The lectures are a good mid-point between the JEP piece and the various full papers where they developed the double ML approach.

Plan of Operation

- 'Approximate Sparsity' set-up.
- The Double ML Approach.
- Examples.

Approximate Sparsity - Set-Up.

Consider the general model for outcome y_i, controls w_i and error ζ:

$$y_i = g(\mathbf{w}) + \zeta_i \tag{1}$$

with $n = 1, ..., n$ independent observations and the aim of $E(\zeta_i | w_i) = 0$.

As part of regularization we posit $g(.)$ as a high-dimensional, approximately linear model:

$$g(\mathbf{w_i}) = \sum_{j=1}^{p} \beta_j x_{i,j} + r_{p,i} \tag{2}$$

where $x_i = (x_{i,j}, ..., x_{i,p})'$ can include the basic regressors or transformations (eg: polynomials, interactions). The $r_{p,i}$ term is 'approximation error'. We allow $p > n$ in this framework.

Approximate Sparsity - Implications.

- Think of approximation error $r_{p,i}$ as a a source of noise that is both (i) distinct from and (ii) small relative to the irreducible or 'true' error ζ_i.

- Practically, sparsity means that we think that only a subset s of all variables x_{ij} have $\beta_j \neq 0$. Hence, a low-dimensional sub-model of size s approximates the full p-dimensional model.

- **Example:** In empirical network interaction models such as Manresa(2016) we plausible think that only a subset of firm-to-firm realistically matter. This is due to well-established stylised facts such as 'small worlds' networks.

Lasso Re-statement

Recall the basic Lasso optimization problem:

$$argmin_b \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{i,j} b_j)^2 + \lambda \sum_{j=1}^{p} \mid b_j \mid \gamma_j$$

...the new element here is the γ_j term for 'penalty loadings'.

This is where econometrics has made a contribution in terms of addressing issues of heteroscedasticity, clustering, and non-normality in model errors. See the Chernozhukov NBER lecture for an introduction.

Inference with Many Controls - Set-up.

Consider a linear model with treatment d_i as follows:

$$y_i = \alpha d_i + x_i'\theta_y + r_{yi} + \zeta_i \tag{3}$$

where we take d_i as exogenous after conditioning $E(\zeta_i | d_i, x_i, r_{yi}) = 0$ and allow $p > n$. Note that d_i is a 'target' variable that we include no matter what (ie: it is not subject to variable selection) since we're interested in it economically.

We can also think of a 'reduced form' for d_i whereby it is itself potentially influenced by many controls:

$$d_i = x_i'\theta_d + r_{di} + v_i \tag{4}$$

where v_i is irreducible error and r_{di} is approximation error specifically for the d_i equation.

Inference with Many Controls - Further Set-up.

Now think about combining these equations by plugging the d_i expression into the y_i 'structural' equation to get a single reduced form equation:

$$y_i = x_i'(\alpha\theta_d + \theta_y) + (\alpha r_{di} + r_{yi}) + (\alpha v_i + \zeta_i) \tag{5}$$

Where we can simplify this with notation to reflect the composite nature of the parameters:

$$y_i = x_i'\Pi + r_{ci} + \epsilon_i \tag{6}$$

..where we've just re-written the expression immediately above. Armed with this set-up we can think about estimation...

Empirical Modelling with Many Controls - Naive Approach.

- The naive approach **first** involves running Lasso on the model above in (6) to study which x_j are good predictors with potential non-zero β_j coefficients.
- The **second** 'post-Lasso' step is to just keep the x_j that were selected by Lasso in the first step and run OLS just with those variables and target variable d_i.
- But this has proven to be problematic in terms of the inference on the final estimated β_j's. The core problem relates to the fact that we are estimating model model (6) when implementing the first Lasso selection step and bump into composite parameters as a result.

Key Problem with Naive Approach.

Think about what is selected when we estimate model (6):

- Within the Π-vector we are selecting those x_j that have larger effects on y_i but we will miss out the x_j with more moderate effects because they get hit harder by the penalization.
- These 'moderate effect x_j' are then omitted from the next post-Lasso step. The problem is that these omitted x_j might be correlated with the d_i variable.
- What will then happen is that our estimated α will pick up the effects of these missing x_j's following the logic of omitted variable bias (OVB). If there are many p variables floating around in our data then this could create a big channel for OVB to have an effect.

Alternative Double ML Approach.

The Double ML approach is designed to get rid of this OVB effect and involves the following steps:

- **Step 1:** Run a Lasso model of outcome variable y_i on the x_{ij}'s in order to select the x_{ij}'s that best predict y_i.
- **Step 2:** Run a Lasso model of target variable d_i on the x_{ij}'s in order to select the x_{ij}'s that best predict d_i.
- **Step 3:** Run OLS of y_i on d_i and the **union** of the controls found in the first two steps. That is, include every x_j that is a good predictor of either y_i or d_i.

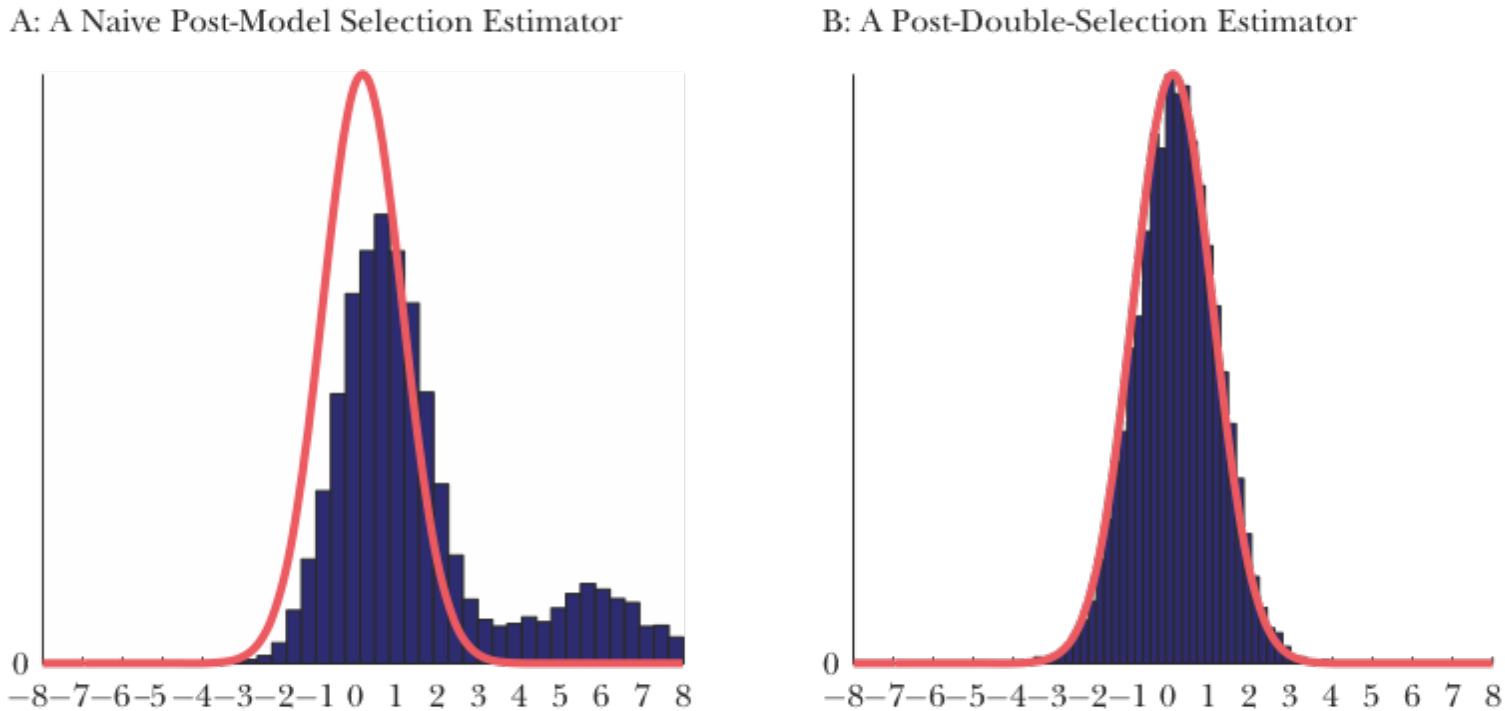
Double ML Approach.

- The key to the Double ML approach is that it effectively manages to partial out the OVB by making sure that any important x_j predictor of d_i is included in our final y_i outcome model.
- This prevents any 'missing in action' x_j from exerting an indirect influence on the important α coefficient via the OVB channel.
- Simulations show that the naive approach is badly biased relative to the Double ML approach...

Figure 1

The "Double Selection" Approach to Estimation and Inference versus a Naive Approach: A Simulation from Belloni, Chernozhukov, and Hansen (forthcoming)

(distributions of estimators from each approach)

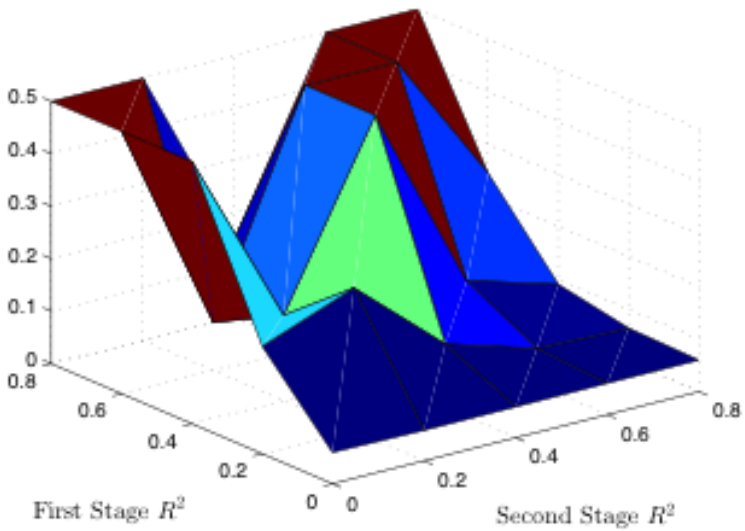


Source: Belloni, Chernozhukov, and Hansen (forthcoming).

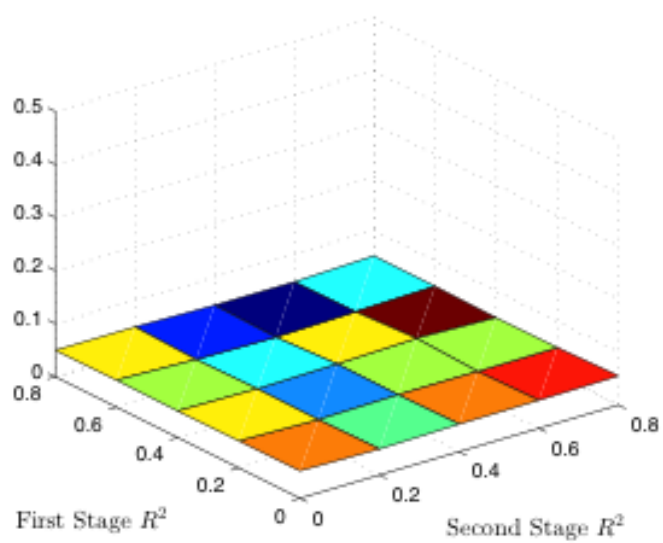
Notes: The left panel shows the sampling distribution of the estimator of α based on the first naive procedure described in this section: applying LASSO to the equation $y_i = d_i + x_i' \theta_y + r_{yi} + \zeta_i$ while forcing the treatment variable to remain in the model by excluding α from the LASSO penalty. The right panel shows the sampling distribution of the "double selection" estimator (see text for details) as in Belloni, Chernozhukov, and Hansen (forthcoming). The distributions are given for centered and studentized quantities.

Look at the rejection probabilities of a true hypothesis.

Naive/Textbook Selection

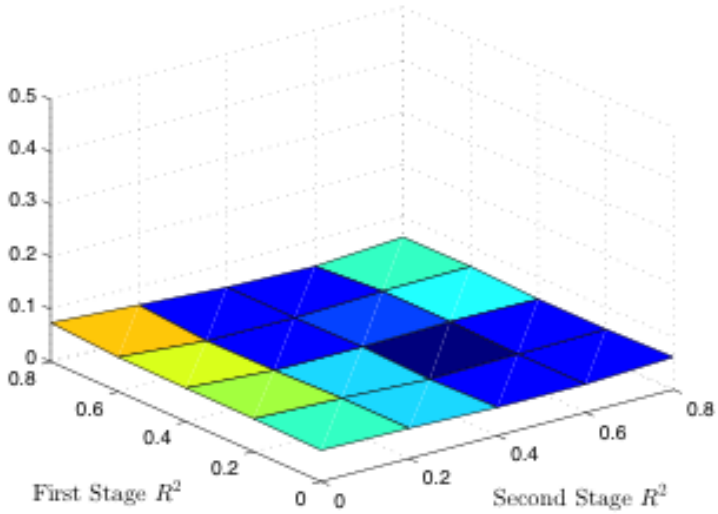


Ideal

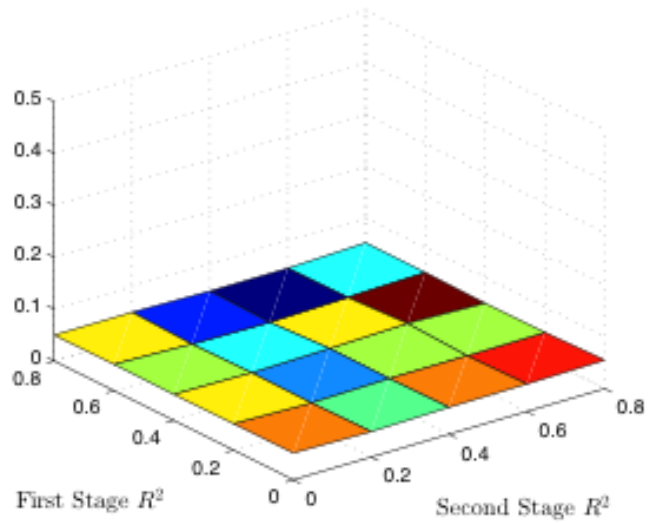


actual rejection probability (LEFT) is far off the nominal rejection probability (RIGHT)
consistent with results of Leeb and Pötscher (2009)

Double Selection



Ideal



the left plot is rejection frequency of the t-test based on the post-double-selection estimator

Belloni, Chernozhukov, Hansen (2011)

Example: Abortion and Crime

Based on famous Donohue and Levitt (2001) paper.

- **Goal:** Understand the causal effect of abortion (d_{it}) on crime (y_{it})
- **Problem:** Abortion rates are not randomly assigned. States are different for many reasons and crime rates might evolve on the basis of various initial conditions (eg: demographics, baseline crime levels).
- A common modelling response to this concern is to include unit-specific trends, either plain linear or some other functional form. I call researcher-driven specification like this the 'heuristic approach'.

Baseline Abortion-Crime Model.

Original Donohue-Levitt model is:

$$y_{it} = \alpha_0 d_{it} + x'_{it}\beta_g + \gamma_t + \delta_i + \epsilon_{it} \tag{7}$$

where y_{it} is the crime rate, d_{it} is the abortion rate, γ_t is a common time period effect across all states and ϵ_{it} is an error term.

The x_{it} controls are: log of lagged prisoners per capita, log of lagged police per capita, the unemployment rate, per capita income, poverty rate and a $(t-15)$ measure of welfare payments ('AFDC generosity').

The exercise by Belloni et al (2014) JEP modifies this baseline by formulating first differences to take account of the state fixed effects instead of running levels with state dummies.

Applying a Variable Selection Approach.

- Linear state-specific trends have the issue of (i) being unrealistic and (ii) adding many potentially irrelevant variables, in the process pushing up estimated error variance $(\frac{1}{N-k}\hat{\sigma}^2)$.

- We therefore want to model an appropriate flexible, nonlinear function to reflect trend movements in the data. In practice , this will be an interaction of initial conditions and state averages with t and t^2 terms.

- The common time effects γ_t are included by default and not selected over. The state time effects are dealt with by first differencing and are therefore also included by 'default'.

Formal Dimension Reduction.

Researcher input still comes in deciding on a key set of relevant x_j. The following sets up a cubic trend in various baseline or average characteristics:

1. Differences in the 8 original controls.
2. Initial conditions of controls, abortion rate, crime rate.
3. Within state averages of controls, abortion rate.
4. t, t^2
5. Interactions of 1-3 with time terms in 4.

This gives a general model with $p = 284$ and $n = 576$. The Double ML approach is then applied to equations for Δy_{it} and Δd_{it}.

So What Does Double ML Select?

This example is for the property crime equation and using a 'plug-in' method for deciding on the value of the λ penalization parameter.

- **Abortion Rate equation:** lagged prisoners, lagged police, lagged unemployment, initial income, initial income difference $\times$ t, initial beer consumption difference $\times$ t, initial income $\times$ t, initial prisoners squared $\times$ t^2, average income, average income $\times$ t, initial abortion rate (12 variables).

- **Crime equation:** Initial income squared $\times$ t, Initial income squared $\times$ t^2, average AFDC squared.

- **Does it make sense?** Kind of. One problem is the selection of higher-order interactions but dropping of the lower-order terms. You would want to do a manual or 'eyeball' sense check and benchmark against similar heuristic models.

The Moment of Truth.

- The 'all controls' approach with the full $p = 284$ completely shreds the data (point estimate falls, std errors blow up) as expected. Note that this specification is only feasible because $p < n$ in this application.

- Double selection gives us a similar point estimate to the heuristic model but overall it is imprecise. Abortion is not significantly associated with crime!

- The result is qualitatively similar to a critique by Foote and Goetz (2008). But note that Donohue and Levitt (2008) responded with new estimates based on a longer t sample.

Table 1

Effect of Abortion on Crime

	Type of crime					
	Violent		*Property*		*Murder*	
Estimator	*Effect*	*Std. error*	*Effect*	*Std. error*	*Effect*	*Std. error*
First-difference	−.157	.034	−.106	.021	−.218	.068
All controls	.071	.284	−.161	.106	−1.327	.932
Double selection	−.171	.117	−.061	.057	−.189	.177

Notes: This table reports results from estimating the effect of abortion on violent crime, property crime, and murder. The row labeled "First-difference" gives baseline first-difference estimates using the controls from Donohue and Levitt (2001). The row labeled "All controls" includes a broad set of controls meant to allow flexible trends that vary with state-level characteristics. The row labeled "Double selection" reports results based on the double selection method outlined in this paper and selecting among the variables used in the "All controls" results.

Other Examples from Belloni et al (2014) JEP.

- **Acemoglu and Robinson (2001):** Famous settler mortality paper. Double ML ultimately gives a weaker first stage and attenuated second stage. But not disastrous. Note that since this is an IV setting we have ML selection over 3 equations ('triple ML'?).

- **Chen and Yeh (2012):** Effect of eminent domain (govt seizure of property) on regional ('Case-Shiller') house price index. Judges determining the policy are randomly assigned so we can use judge $X's$ as instruments. get stronger first stage with extra Z's selected but unreliable second stage.