Centre de recherche en économie et statistique

Série des Documents de Travail

n° 2016-23

Distinguishing the Confounding Factors: Policy Evaluation, High-Dimension and Variable Selection

J.L'hour[1]

[1] ENSAE ParisTech – CREST. E-mail : Jeremy.L.HOUR@ensae-paristech.fr

Distinguishing the Confounding Factors: Policy Evaluation, High-Dimension and Variable Selection

Jérémy L'Hour*

ENSAE ParisTech - CREST

June 21, 2016

Abstract

Variable selection is an important question for policy evaluation when identification of the treatment effect relies on a conditional-on-observables strategy. Recent advances in variable selection methods, such as the Lasso, have been deemed useful for the econometrics of policy evaluation. The Lasso approach focuses on the computational feasibility of exhaustive model selection borrowing from procedures developed in a high-dimensional context. However, it has been seldom applied in policy evaluation works because it raises other difficulties such as the choice of a parameter that sets the trade-off between fit and sparsity. Two Lasso-based treatment effect estimators are reviewed and compared on an empirical application, on which they perform well. This paper also illustrates the pitfalls of variable selection in a policy evaluation context.†

*ENSAE ParisTech / CREST, Microeconometrics department, 15 boulevard Gabriel Péri, 92240 Malakoff, France. E-mail address: jeremy.l.hour@ensae-paristech.fr. I thank Xavier D'Haultfœuille, Alexandre Tsybakov, Max Farrell, Laurent Davezies and participants to CREST internal Microeconometrics seminar for advice and useful comments.

†JEL Classification: C01, C21, C52, C55, Keywords: treatment effect, variable selection, policy evaluation, semi-parametric estimation, high-dimension

I Introduction

Model selection and parsimony amongst explanatory variables are traditional scientific problems that have a particular echo in statistics and econometrics. They have received growing attention over the past decade, as high-dimensional datasets have become increasingly available to statisticians in various fields. But even with a small dataset, one can be quickly faced with a high-dimensional problem, for example when doing series estimation of a non-parametric model. In practice, applied econometricians often select features amongst a large design matrix by trial and error, guided by their intuition and report results based on the assumption that the selected model is the true. These results are often backed by further sensitivity analysis and robustness checks. However, the variable selection step of empirical work is rarely fully acknowledged and post-selection inference lacks uniformity. Leamer (1983) was one of the first econometric papers to address this problem. For a modern presentation, see Leeb and Pötscher (2005, 2006) and, in the context of policy evaluation, Belloni *et al.* (2014a).

This is particularly problematic in models that rely on a conditional-on-observables identification strategy for estimation, as it has been the case in a wide range of topics, including the labour market (LaLonde, 1986), development economics (Jalan and Ravallion, 2003), monetary policy (Angrist *et al.*, 2013), to name a few. In policy evaluation, estimation of the average causal treatment effect can rely on this type of assumption and may require a careful selection of the set of regressors to include in the model. For ease of exposition purpose, let me introduce the Rubin causal framework (Rubin, 1974). The quantity of interest is the effect of a treatment denoted D which is equal to 1 if the individual is treated and 0 otherwise. The potential outcomes with and without the treatment are denoted Y_1, Y_0. The predicament in estimating the treatment effect is that one never observe both outcomes, but only $Y = Y_0 + D(Y_1 - Y_0)$. The outcome which is not observed is called the counterfactual. Several estimands are of interest in empirical applications but the most common are the Average Treatment Effect (ATE) defined as $\Delta^{ATE} = \mathbb{E}(Y_1 - Y_0)$ and the Average Treatment Effect on the Treated (ATT) defined as $\Delta^{ATT} = \mathbb{E}(Y_1 - Y_0|D = 1)$. The behaviour of the agent makes estimation of such quantities difficult because it is reasonable to assume that agents self-select into the treatment, which is no longer exogenous to its outcomes and makes identification difficult.

The main challenge in this literature is to find a plausible estimate of the counterfactual. A central assumption to ensure identification of the quantities of interest in many empirical policy evaluation works is the Conditional Independence Assumption (CIA) or Selection on Observables:

$$Y_1, Y_0 \perp\!\!\!\perp D|X \tag{1}$$

This assumption states that conditionally on a set of well-identified observed covariates or confounding factors X, taking the treatment is independent of both potential outcomes. It implies that observations that are close in terms of X can be treated as having come from a randomized experiment. Hence, it is as if the treatment is exogenous and causal identification is possible since both potential outcomes are unaffected by being treated or not. Several estimators used in policy evaluation works rely on this assumption or on slightly weaker versions of it, such as OLS or Matching techniques. For a survey of policy evaluation methods see Imbens and Wooldridge (2009) and in French, Givord (2010), and see Imbens and Rubin (2015) for a textbook treatment.

However, this assumption appears to be fragile and results may be highly dependent on selecting the right set of covariates X to be included in the right-hand side of the regression equation. In practice, researchers are faced with datasets containing a large number of covariates, possibly even more than observations. They have two main problems: (1) the choice of the functional form, which is either dictated by the support of the outcome of interest or by simplicity (linear), (2) selection of the covariates, which is performed relying on economic theory or ad-hoc procedures. On the one hand, if variable selection is performed badly, the CIA may not hold and inference may not be valid. On the other hand, even if the correct set of covariates is used but the model is badly specified, inference will also suffer.

To deal with the model selection problem (in the widest sense) in a rigorous framework and overcome the shortcomings of usual methods like OLS, several methods have been developed. Traditional methods using information criteria (*e.g.* the BIC or the AIC) for exhaustive model selection are not computationally feasible as I will illustrate later. Fairly recently, a lot of attention has been devoted to estimators that penalize non-zero coefficients in a way that makes computation easier, namely with a convex function. Most of the time, the high-dimensional setting is used, even though this is not necessary. The most famous of these estimators is the Lasso (for Least Absolute Shrinkage and Selection Operator) of Tibshirani (1996) that uses the ℓ_1-norm. Proper-

ties and extensions of the Lasso have been developed and studies for example in Candes and Tao (2007); Van de Geer (2008); Bickel *et al.* (2009); Meinshausen and Yu (2009); Lounici (2008); Zhao and Yu (2006); Zhang and Huang (2008); Chatterjee (2013).

Automatic variable selection methods and Lasso-type methods have been deemed useful in the policy evaluation literature, even though they have seldom been applied due to several difficulties. Firstly, Belloni *et al.* (2014a,b) highlight the danger of selecting controls by only considering the outcome equation and propose a three-step procedure where Lasso selection is made firstly (1) on an equation where the treatment indicator is regressed on a set of controls, secondly (2) on an equation where the outcome variable is regressed on the same set of controls. Then, in a final step (3) the outcome is regressed on the treatment controlling for the union of the selected variables in both (1) and (2). This method helps selecting more controls and guards against omitted variable biases much more than a simple "post-single-selection" estimator, as it is usually done in the literature where only step (2) is used. This method is relatively simple. Farrell (2015) extends this approach by allowing for heterogeneous treatment effects, proposing an estimator that is robust to either model selection mistakes in propensity scores or in outcome regression, and dealing explicitly with a discrete treatment that is a more common setting in the policy evaluation literature.

Belloni *et al.* (2014b); Farrell (2015) and references cited earlier make a compelling case for dealing with model and covariate selection much more carefully in empirical work. The estimators they proposed have appealing theoretical properties. However, Lasso methods raise problems that are not yet fully understood, such as the trade-off between sparsity and fit. For example, the theoretically optimal penalty choice in Bickel *et al.* (2009) depends on unknown quantities such as the variance of the error terms that needs to be estimated. Other classical methods in Statistics can also be used but they do not necessarily belong to the applied economist's toolbox. The aim of this paper is to present these modern variable selection tools to the econometrician interested in evaluation of public policies in an intuitive way. It investigates the sensitivity to the penalty level using the dataset of LaLonde (1986) and shows that these tools perform well.

Section II recalls the main intuitions and ideas from recent advances in variable selection that come from the modern Statistics literature. I present a simple model

that nevertheless allows me to display the main intuitions, which remain valid in more complicated models. Section III presents two treatment effect estimators that rely on ℓ_1-penalized procedures. Section IV presents a short simulation study comparing several estimators. Section V presents an empirical application to compare the main estimators and illustrate their main advantages and difficulties.

II A Brief Introduction to Variable Selection Methods

To deal with the model selection problem in a rigorous framework and overcome the shortcomings of usual methods like OLS in terms of variable selection especially in a high-dimension context, several methods have been developed. This section is a brief introduction to the problem and offer intuitive examples before returning to the policy evaluation context.

II.1 Combinatorial Methods vs. ℓ_1-penalization

Consider the following linear regression model with the usual assumptions:

Model II.1 (Linear Model)

$$y_i = x_i^T \beta + \varepsilon_i, \ \forall i = 1, .., n \tag{2}$$

where y_i is the observed dependent variable for individual i, x_i is the (column) vector of p observed random regressors, β is the $p \times 1$ deterministic vector of coefficients and ε_i is the residual. For convenience, define the $n \times p$ matrix $X := \left(x_1^T, ..., x_n^T\right)$. Assume (y_i, x_i) to be iid and $\mathbb{E}(\varepsilon_i | x_i) = 0$.

The high-dimensional case arises when $p > n$, in which case the OLS estimator cannot be computed. A classical example in genomic is the detection of genes responsible for obesity, in which case n the number of patients can be a few hundreds while there may be $p = 3000$ genes to consider. For the applied econometrician, this problem may arise even if $p \leq n$ but p relatively large compared to n because $X^T X$ may be close to singular.

A convenient way to deal with high-dimensional models is to assume *sparsity* in β, *i.e.* only $s << n$ elements of β are non-zero. The set of non-zero components of β is called its *sparsity pattern* : $J(\beta) := \{j : \beta_j \neq 0\}$. In the gene case, it means that only two or three genes are to be responsible for obesity and all others to have an insignificant impact (*i.e.* to be outside the sparsity pattern). In a simple policy evaluation framework,

one would be typically interested in estimating an equation of the type:

$$y_i = d_i\alpha + x_i^T\beta + \varepsilon_i, \forall i = 1, .., n \tag{3}$$

Where α is the effect of the treatment and x_i a large vector of potential confounding factors or transformations of them. Here, the CIA translates into $\mathbb{E}(\varepsilon_i|x_i, d_i) = 0$. Acknowledging the variable selection step means that knowledge of $J(\beta)$ is not assumed but the uncertainty surrounding this set is also taken into account. In the not-so-large-dimensional case, x_i's could be included in the right-hand-side of the equation, but the variance of the estimator would blow up or the interpretation of some coefficients may drastically change (see Berk *et al.* (2013)).

The traditional, combinatorial, way of selecting a sparse model would be to estimate all the possible models that include only a given s number of variables by OLS and take the one that gives the best fit (denoted by $M_s^\star$). Do this for $s = 1, ..., \min(n, p)$. Then compare models $M_1^\star, ..., M_{\min(n,p)}^\star$ using a criterion that penalizes the number of variables included in the model such as the BIC of Schwarz (1978), *i.e.* solve the following program where $C(s)$ is a penalty function such as the AIC or BIC:

$$\min_s \min_{\beta:\|\beta\|_0=s} \left\{ \sum_{i=1}^n \left(y_i - x_i^T\beta\right)^2 \right\} + C(s) \tag{4}$$

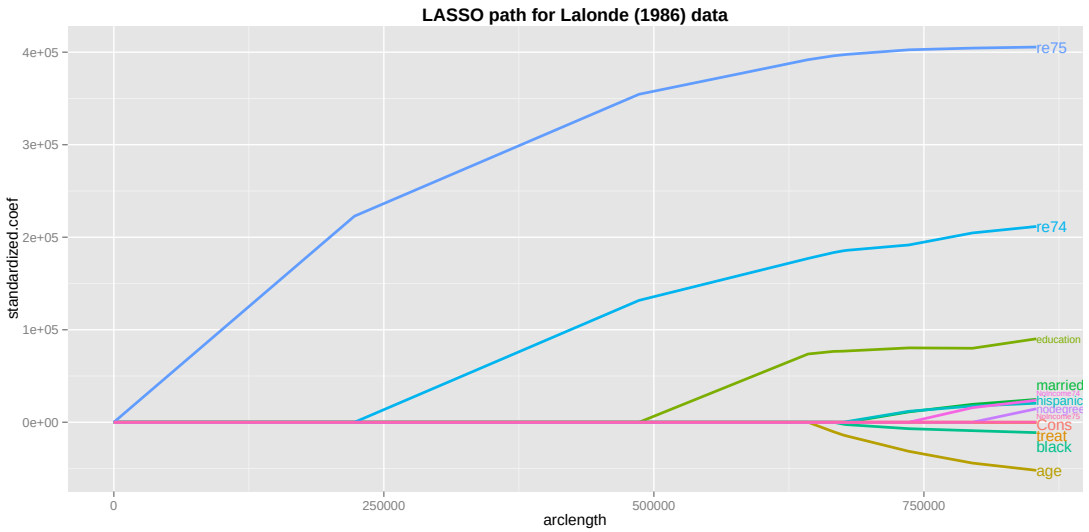
However, this would require to estimate $2^{\min(n,p)}$ models by least squares, which isn't technically feasible as soon as the number of possible regressors becomes moderately large. Indeed, this comes from the fact that non-zero elements are penalized by mean of the ℓ_0-norm which isn't convex. Fairly recently, a lot of attention has been devoted to estimators that penalize non-zero elements of β in a way that makes computation easier, namely with a convex function. The most famous of these estimators is the Lasso of Tibshirani (1996) that uses the closest convex function - the ℓ_1-norm. The Lasso estimator is defined as the result of the following minimization program ($\lambda > 0$):

$$\hat{\beta}^\lambda \in \arg\min_\beta \sum_{i=1}^n \left(y_i - x_i^T\beta\right)^2 + \lambda\|\beta\|_1 \tag{5}$$

λ sets the trade-off between a good fit and a sparse solution. When $\lambda = \infty$, the solution will be $\hat{\beta}^\infty = 0_p$. When $\lambda = 0$, the solution will be the OLS estimator if it can be computed $\hat{\beta}^0 = \hat{\beta}^{OLS}$. Between these two values, a typical Lasso estimator will give a solution which is sparse in the sense that several component of the estimated vector $\hat{\beta}^\lambda$ will be *exactly* zero. Computing the whole path $\left\{(\lambda, \hat{\beta}^\lambda), \lambda > 0\right\}$ can be

computationally greedy but efficient algorithms have been developed in Efron *et al.* (2004) and are available in usual statistical software such as R and Stata. Figure 1 plots an example of the Lasso regularization path as the penalty shrinks to zero (OLS solution, at the far right). This type of plot can provide an helpful guide when setting the penalty level, especially in a context where only raw variables (and not hard-to-interpret transformations of them) are included in the set of regressors.

Figure 1: Lasso Regularization Path: an Example



Note: LASSO path for the 1978 income regression using all the raw covariates. Data taken from the Lalonde (1986). The value of the coefficients are plotted against $|\hat{\beta}^{\lambda}|/|\hat{\beta}^{OLS}|$.

In a nutshell, the Lasso has gained success from three of its properties: it is well-defined in a high-dimensional setting $p > n$, it gives an exactly sparse solution and it is computationally efficient.

II.2 The Lasso: a Toy Example

This section presents the simple case of the Gaussian sequence model to illustrate how the Lasso works. It is taken from Tsybakov (2008, p. 164).

Model II.2 (Homoscedastic Gaussian Linear Model)
Consider the regular linear model in its vectorial form:

$$\tilde{Y} = X\beta + \varepsilon,\ \varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right),\ \varepsilon \perp\!\!\!\perp X \tag{6}$$

Assumption II.1 (ORT)

The design matrix has the following property:

$$\frac{1}{n}X^TX = I_p \tag{7}$$

This assumption means that the covariates are all uncorrelated to each other. This is of course a very strong assumption but in modern versions of the Lasso, it is relaxed and replaced by something less restrictive. Pre-multiplying the model by X^T/n transforms the regression problem into a signal-noise detection problem:

Model II.3 (Gaussian Sequence Model)

$$y_j = \beta_j + \xi_j,\ \xi_j \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right) \tag{8}$$

Under the (ORT) assumption, the Lasso program is equivalent (up to a constant) to the following program:

$$\min_{\beta} \sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{9}$$

And the following result gives the intuition of the Lasso, viewed as the soft-thresholding estimator:

Theorem II.1 (Soft-Thresholding Estimator)
The solution of the minimization program (9) *is given by the soft-thresholding estimator:*

$$\hat{\beta}_j = \left(1 - \frac{\lambda}{2|y_j|}\right)_+ y_j \tag{10}$$

Proof. A term-by-term minimization gives the result. From the square term: $\hat{\beta}_j$ and y_j will always have the same sign otherwise there is another β_j that minimizes the function. Hence, if $\beta_j \neq 0$ the following first-order condition arises:

$$\beta_j = y_j - \frac{\lambda}{2}\text{sign}(y_j)$$

At this point, the objective function takes the value $\left(\frac{\lambda}{2}\right)^2 + \lambda|y_j - \frac{\lambda}{2}\text{sign}(y_j)|$. At $\beta_j = 0$ the function take the value y_j^2 and consequently the solution is:

$$\hat{\beta}_j = 0,\ \text{if } |y_j| < \frac{\lambda}{2}$$
$$\hat{\beta}_j = y_j - \frac{\lambda}{2}\text{sign}(y_j),\ \text{if } |y_j| \geq \frac{\lambda}{2}$$

□

Two main observations must be made from this result. The first one is that the Lasso works as a thresholding estimator, *i.e.* relatively small values of the coefficients are set to zero. The larger the penalty level λ, the more coefficients will be set to zero. The second thing is that the Lasso necessarily induces a bias, contrary to the hard-thresholding estimator or the BIC estimator. This bias could be removed by running an OLS estimation on the set of variables detected as having a non-zero coefficient. Such an estimator is called Post-Lasso, or in general Post-selection estimator and has been studied in Belloni and Chernozhukov (2013). The Gaussian form of the error terms allows to use simple concentration inequalities, and easily yields results on the variable selection properties of the Lasso (*i.e.* the Lasso selects the true sparsity pattern with high probability). I am using the example in Tsybakov (2014) for this result. In many "true" applications, a sparsity assumption is first required. This is not strictly necessary to prove Theorem II.2 in this very simplistic case, but variable selection would not be useful if all variables had a non-zero coefficient!

Assumption II.2 (Sparsity)
The true sparsity pattern $J(\beta^*)$ *contains* $s < n$ *elements.*

Now the variable selection property of the Lasso follows:

Theorem II.2 (True Sparsity Selection Property)
Assume that $\forall j \in J(\beta^*)$, $|\beta_j| > \lambda$ *and set* $\lambda = 2\sigma\sqrt{2\log(p)/n}$. *With probability greater or equal to* $1 - 1/\sqrt{\pi \log(p)}$*:*

$$J(\hat{\beta}) = J(\beta^*) \tag{11}$$

Proof. Consider the following event:

$$\mathcal{A} = \left\{|y_j - \beta_j^*| < \lambda/2, j = 1, ..., p\right\}$$

$$\begin{aligned}
\mathbb{P}(\mathcal{A}) &= \mathbb{P}(|\xi_j| < \lambda/2, \forall j = 1, ..., p) \\
&= 1 - \mathbb{P}\left(\max |\eta_j| > \sqrt{2\log(p)}\right) \\
&\geq 1 - \frac{1}{\sqrt{\pi \log(p)}}.
\end{aligned}$$

with $\xi_j = \sigma\eta_j/\sqrt{n}$, $\eta_j \sim \mathcal{N}(0,1)$ and using lemma A.2. Now show by double inclusion that on this event, the two sparsity patterns coincide.

If $j \in J(\hat{\beta})$, $|y_j| > \lambda/2$ then on the event $\mathcal{A}$:

$$|\beta_j^*| \geq |y_j| - |y_j - \beta_j^*| > 0$$

Hence, $j \in J(\beta^*)$.

If $j \in J(\beta^*)$, $|\beta_j| > \lambda$ then on the event $\mathcal{A}$:

$$|y_j^*| \geq |\beta_j^*| - |y_j - \beta_j^*| > \lambda/2$$

Hence, $j \in J(\hat{\beta})$. Hence, on $\mathcal{A}$, $J(\hat{\beta}) = J(\beta^*)$. □

Two main observations must be made from this result. Firstly, the good news: the Lasso selects the true sparsity pattern with high probability. It is convenient because the Lasso is much easier to compute than the BIC and even if it is biased, at least it has good selection properties. Now the bad news: to ensure this property, λ the penalty term, is set to a value which is unknown is most empirical applications because the true variance of the error terms is unknown.

II.3 More General Cases and Extensions: a Literature Review

The previous section exposed the main advantages and problems regarding the Lasso estimator in a very simple case, made possible using the (ORT) assumption. These features however exist in more generic cases and several papers documented and extended them, using less restrictive assumptions.

A excellent textbook treatment of the application of Lasso and related methods can be found in Hastie *et al.* (2009). A more advanced and theoretical treatment can be found in Buhlmann and van de Geer (2011). For a presentation of the use of the Lasso in an economic context see Belloni and Chernozhukov (2009); Fan *et al.* (2011), and particularly in a policy evaluation context see Belloni *et al.* (2014a).

The selection and estimation properties of the Lasso and related ℓ_1-penalized regressions have been studied in numerous statistical papers, including Zhao and Yu (2006); Candes and Tao (2007); Van de Geer (2008); Meinshausen and Yu (2009); Bickel *et al.* (2009); Huang *et al.* (2010) and references therein. Some of these references have shown that the sparsity pattern of the Lasso estimator is only *asymptotically* identical to the true sparsity pattern under restrictive necessary conditions that are likely to be violated, jeopardizing the Lasso selection properties. Two of them are particularly fragile.

The first one is a so-called *irrepresentable condition* (Buhlmann and van de Geer, 2011, Section 7.5) that restricts the correlation between the covariates. An investigation of the topic and adequate relaxations have been proposed in Meinshausen and Yu (2009). Another one is the so-called *beta-min condition* (Buhlmann and van de Geer, 2011, Section 7.4) that requires the absolute value of all the non-zero coefficients to be above a certain threshold. The impossibility to detect small coefficient threatens the validity of inference particularly for post-selection estimators as described in Leeb and Pötscher (2005).

It is to be noted however that the econometric literature has switched its focus from good variable selection to adequate approximation of nuisance parameters such as control functions, propensity scores or outcome equations, see Belloni *et al.* (2014b,a, 2012). For example, in Farrell (2015), the Lasso is used because it can predict well the outcome equations and the propensity scores which is what matters for the chosen treatment effect estimator to work well. In the spirit of the econometric literature that highly values parsimony regarding assumptions on Data-Generating Processes (DGP), these results have also been extended in the non-parametric, non-Gaussian cases with mild assumptions regarding the DGP, for example in Belloni *et al.* (2012, 2014b).

Another departure from the traditional Lasso is the relaxation of the exact sparse structure of the model as embodied in Assumption II.2. Instead of assuming that the vector of coefficients is exactly sparse (*i.e.* has a small number of non-zero elements), a more general assumption introduced notably by Belloni *et al.* (2012) is that the vector of coefficients can be decomposed into two components: a sparse component and a small component. The second component is small in the sense that its ℓ_1-norms decays to zero as the sample size grows but is never exactly zero. The approximate sparse structure occurs in many possible cases and is relatively rich. The Lasso also performs well in this case.

The optimal penalty choice is another prominent issue that arises with the use of the Lasso. It will be illustrated latter in this paper, but I can nonetheless refer the reader to several key papers in Statistics. In theory, the optimal choice given for example in Bickel *et al.* (2009) depends on unknown quantities such as the variance of the error terms. Consequently, the optimal Lasso is infeasible. The dominant approach to make the Lasso feasible is given in Belloni *et al.* (2012); Belloni *et al.* (2014) and uses an iter-

ative procedure. The square-root Lasso is a nice alternative to circumvent the problem posed by the Lasso in the sense that the optimal penalty choice does not depend on the variance of the error term. It has been developed by Belloni *et al.* (2011). Along this line, Gautier and Tsybakov (2011); Lederer and Müller (2014) use more complex estimators that are more or less tuning-parameter free.

This section reviewed two methods for selecting variables in a regression model: the combinatorial approach or the ℓ_0-penalized approach, and the ℓ_1-penalized approach. They both perform well in terms of variable selection. They differ, however, in their feasibility: the combinatorial approach cannot be used as soon as the number of regressors is too large[1]. The Lasso, while being much more efficient, introduces a bias which can be removed. This is why it is the dominating approach. Next section reviews two contributions that use ℓ_1-penalized estimators to select the confounding factors in a policy evaluation context.

III Confounding Factors Selection for Treatment Effect Estimation

Sections III.1 and III.2 present the contributions of Belloni *et al.* (2014b) and Farrell (2015) respectively. Section III.3 deals with the delicate question of the penalty-level choice and review alternative selectors for which the optimal penalty level choice problem isn't as acute.

III.1 The Post-Double-Selection Estimator of Belloni *et al.* (2014b)

Belloni *et al.* (2014b) propose an estimator of the ATE that is robust to model selection mistakes in a semi-parametric framework. It is applicable when there are a very large number of covariates, possibly more than observations. The next section describes the model and the following one dives slightly deeper in the assumptions. The reader can skip this second section at first read.

III.1.1 Model and Estimation Strategy

The model is given by the outcome equation and the treatment equation :

[1]Some authors have tried, see for example Sala-i Martin (1997), but it is not a very convenient solution!

Model III.1 (BCH)

$$Y_i = D_i\alpha_0 + g(Z_i) + \xi_i \tag{12}$$

$$D_i = m(Z_i) + v_i \tag{13}$$

In this model, Z_i are raw covariates that enter on both right-hand sides of the equations through unknown functions $g(.)$ and $m(.)$. The implicit assumption is that when these covariates are included in the equations with the right functional form, there is no endogeneity concern, as the authors assume $\mathbb{E}(\xi_i|D_i, Z_i) = 0$ and $\mathbb{E}(v_i|Z_i) = 0$. It is an assumption that is similar albeit weaker than the CIA. This is their central identification assumption. The parameter of interest is α_0 the ATE of D_i, which is homogeneous. Linear combination of control terms $X_i = P(Z_i)$ are used to approximate $g(.)$ and $m(.)$:

Model III.2 (Linear BCH)

$$Y_i = D_i\alpha_0 + X_i^T\beta_{g0} + r_{gi} + \xi_i \tag{14}$$

$$D_i = X_i^T\beta_{m0} + r_{mi} + v_i \tag{15}$$

This model is assumed to be approximately sparse in the sense that only a small number of elements in β_{g0} and β_{m0} are needed to be different from zero in order to make both remainder terms r_{gi} and r_{mi} small.

Their method follows a three-step approach:

(1) Control variables that are useful for predicting the treatment are selected from a Lasso regression of D_i on the whole set of covariates. This step helps identifying confounding factors.

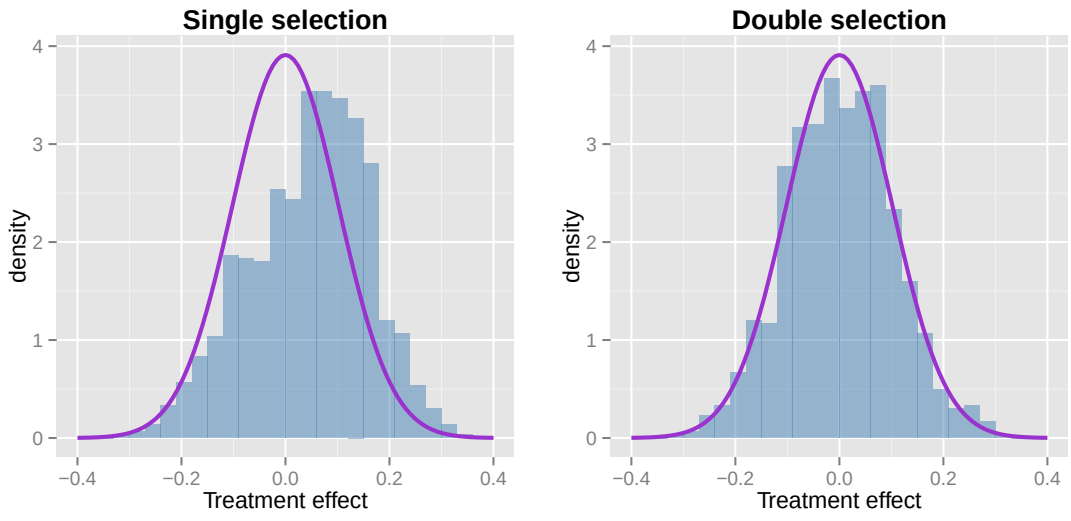
(2) Control variables that are useful for predicting the outcome are selected from a Lasso regression of Y_i on the whole set of covariates, excluding the treatment. This step is aimed at capturing the main determinants of the outcome.

(3) In the final step, the ATE α_0 is estimated by a linear regression of Y_i on D_i and the union of the selected controls from the first two steps.

At each selection steps (1) and (2), a regular linear Lasso is used. Their method has several strengths. The first one is that is it relatively simple in the case of a linear

model. Selection for step (1) and (2) is assumed to be performed using the Lasso and model selection properties of their method is proven using this selector. Moreover, using not only step (2) as it is often done in the literature, but also step (1) helps selecting more controls and guards against potential omitted variables biases much more than a simple "post-single-selection" estimator. The intuition is as follows: imagine performing variable selection over X_i by running a regression of Y_i on D_i and X_i and excluding the elements of X_i for which the t-stat is too low. Assume that some relevant elements of X_i in the equation of Y_i are also very correlated to D_i. Then you might miss them using the single-selection method because the information contained in those elements of X_i is also contained in D_i. And in the post-selection stage, inference regarding the treatment effect will be distorted by the missing variable bias. However, assume now that you run a regression of D_i on X_i: those elements of X_i that were previously missed because of their correlation to D_i will be now selected because of their correlation to D_i ! This is a Frisch-Waugh-Lovell partialling-out procedure for model selection. A graphical illustration from a Monte-Carlo experiment inspired by Belloni *et al.* (2014b) is displayed in Figure 2. Section IV explores these issues in more depth.

Figure 2: The Dangers of Post-Single Selection Inference



Note: Treatment effect estimates from two selection procedures. On the left panel single selection on equation of y_i. On the right panel, double selection on equation of y_i and d_i. Results from a Monte-Carlo experiment replicated 1,000 times. The simulated model is $y_i = 0d_i + .2x_i + \varepsilon_i$, $d_i = .8x_i + u_i$, with ε_i, u_i and x_i independent random variables of distribution $\mathcal{N}(0, 1)$. Sample size is $n = 100$.

III.1.2 Assumptions and Main Results

This section goes through the main assumptions so as to provide an intuition and open the black box of this model.

Assumption III.1 (Approximate Sparse Treatment Effects)
The authors consider the vector $X_i = P(Z_i)$ which is a p-dimensional transformation of the raw covariates. It is assumed that there exist a natural number $s \geq 1$ and β_{m0} and β_{g0} such that:

$$m(Z_i) = X_i^T \beta_{m0} + r_{mi}$$
$$g(Z_i) = X_i^T \beta_{g0} + r_{gi}$$

with $|\beta_{m0}|_0 \leq s$, $|\beta_{g0}|_0 \leq s$ and the empirical L2 norms of the remainder terms r_{mi} and r_{gi} are bounded from above by $C\sqrt{s/n}$. C is a real positive number such that $|\alpha_0| \leq C$.

This assumption means that the functions $g(.)$ and $m(.)$ can be approximated by only a few terms. It is equivalent to say that we can for example expand $g(.)$ in the spline basis and that only s terms will have a non-zero coefficients. A function that has a sparse representation in one basis may not have it in another, even if both of them span the same linear space. Of course, s should be smaller than n. This assumption allows to consider a transformation of model III.1 which becomes linear and allows to use the classical Lasso to deal with variable selection. The next condition is very technical and allows to improve the consistency rate of the Lasso.

Assumption III.2 (Sparse Eigenvalues on the Gram Sub-Matrices)
This assumption that deals with the behavior of the empirical Gram matrix $\mathbb{E}_n(X_i X_i^T)$ is classical in the high-dimension literature. Indeed, when $p > n$, $\mathbb{E}_n(X_i X_i^T)$ is not invertible and even if $p \leq n$ but p relatively large compared to n, $\mathbb{E}_n(X_i X_i^T)$ may be close to singular. Hence usual estimators such a Ordinary Least Squares cannot be used. Define the minimal and maximal m-spare eigenvalue of a semi-definite matrix M as:

$$\phi_{min}(m)[M] := \min_{1 \leq |\delta|_0 \leq m} \frac{\delta^T M \delta}{\|\delta\|^2}$$
$$\phi_{max}(m)[M] := \max_{1 \leq |\delta|_0 \leq m} \frac{\delta^T M \delta}{\|\delta\|^2}$$

A usual assumption in this literature is that $\phi_{min}(m)[\mathbb{E}_n(X_i X_i^T)] > Cst > 0$, i.e. any Gram matrix from a sub-component of dimension m of X_i should be bounded away

from zero. The Sparse Eigenvalues condition writes as follows:
There is an absolute sequence $\ell_n \to \infty$ *such that with a high-probability the minimal and the maximal* $\ell_n s$ *sparse eigenvalues are bounded from above and away from zero:*

$$\kappa' \leq \phi_{min}(\ell_n s)[\mathbb{E}_n(X_i X_i^T)] \leq \phi_{max}(\ell_n s)[\mathbb{E}_n(X_i X_i^T)] \leq \kappa''$$

with $0 < \kappa' \leq \kappa'' < \infty$.

This condition has been introduced and discussed by Bickel *et al.* (2009). It is a more advanced condition than condition (ORT) from Section II but achieves the same purpose: overcoming the singularity of the Gram matrix in high-dimensional models. Finally, the last condition is concerned with moments of the model and is also very technical. It allows notably to apply moderate deviation theorems from Jing *et al.* (2003).

Assumption III.3 (Structural Moments)
There are absolute constants $0 < c \leq C < \infty.$, *and* $0 < q < \infty$ *such that for* $(\tilde{Y}_i, \varepsilon_i) = (Y_i, \xi_i)$ *and* $(\tilde{Y}_i, \varepsilon_i) = (D_i, v_i)$ *the following conditions hold:*

(1) $\mathbb{E}(\mathbb{E}_n(|D_i|^q)) \leq C$, $c \leq \mathbb{E}(\xi_i^2|X_i, v_i) \leq C$ *and* $c \leq \mathbb{E}(v_i^2|X_i) \leq C$ *almost surely for all* i

(2) $\mathbb{E}(\mathbb{E}_n(|\varepsilon_i|^q)) + \mathbb{E}(\mathbb{E}_n(\tilde{Y}_i^2)) + \max_{1 \leq j \leq p} \{\mathbb{E}(\mathbb{E}_n(x_{ij}^2 \tilde{Y}_i^2)) + \mathbb{E}(\mathbb{E}_n(x_{ij}^3 \varepsilon^3))\} + 1/\mathbb{E}(\mathbb{E}_n(x_{ij}^2)) \leq C$

(3) $\log p = o(n^{1/3})$

(4) $\max_{1 \leq j \leq p} \{\mathbb{E}_n(x_{ij}^2 \varepsilon^2 - \mathbb{E}(\mathbb{E}_n(x_{ij}^2 \varepsilon^2)) + \mathbb{E}_n(x_{ij}^2 \tilde{Y}^2 - \mathbb{E}(\mathbb{E}_n(x_{ij}^2 \tilde{Y}^2))\} + \max_{1 \leq i \leq n} |X_i|_\infty^2 \frac{s \log(\max(n,p))}{n} \to 0$

These conditions are relatively mild and can be made in a large number of settings. Let us emphasize the condition on the growth of the number of regressors: $\log p = o(n^{1/3})$ which his very mild. Now we are ready to state the main result of Belloni *et al.* (2014b):

Theorem III.1 (Post-Selection Inference on Treatment Effect)
If assumptions III.1, III.2 and III.3 hold, then the post-double-estimator $\hat{\alpha}$ *satisfies:*

$$\hat{\sigma}_n^{-1} \sqrt{n} \, (\hat{\alpha} - \alpha_0) \to^d \mathcal{N}(0, 1) \tag{16}$$

where $\hat{\sigma}_n^2 = \mathbb{E}_n(v_i^2)^{-1} \mathbb{E}_n(\hat{v}_i^2 \hat{\xi}_i^2) \mathbb{E}_n(\hat{v}_i^2)^{-1}$, $\hat{\xi}_i = \sqrt{\frac{n}{n-\hat{s}-1}}(Y_i - D_i\hat{\alpha} - X_i^T \hat{\beta}_g)$, $\hat{v}_i = D_i - X_i^T \hat{\beta}_m$ *and* $\hat{\beta}_m \in \arg\min_\beta \{\mathbb{E}_n(D_i - X_i^T \beta)^2 : \beta_j = 0, \forall j \notin \hat{I}\}$ *where* $\hat{I}$ *is the set of indices of selected covariates from the first two steps of the procedure.*

Here $\hat{\alpha}$ is of course the Treatment effect estimator obtained by the three-step approach described above. Note that if we have iid data and conditional homoscedasticity on ξ_i, the estimator attains the semi-parametric efficiency bound of Robinson (1988).

The main take-away from the proposed estimator is that using double-selection by considering both the outcome equation and the treatment equation helps reducing the risk of making variable selection mistakes, compared to the usual approaches that only use variable selection for the controls in the outcome equation. Circumventing the results of Leeb and Pötscher (2005, 2006), this approach yields a uniformly valid inference procedure for the ATE.

III.2 The Doubly-Robust Estimator of Farrell (2015)

III.2.1 Treatment Effect Estimation Strategy and Double Robustness

The estimator proposed by Farrell (2015) is directly geared towards estimation of the treatment effect. It adopts the same point of view as Belloni *et al.* (2014b) in the sense that it focuses on a sparse approximation of the control functions and not on good selection of precise raw variables and uses the same variable selection techniques *i.e.* the Lasso. Nevertheless, it builds on the Doubly-Robust approach to allow for an heterogeneous treatment effect and uses a Logit propensity score to deal with an explicitly discrete treatment. These two features make it more appealing in the context of policy evaluation.

Although the paper considers a multi-valued discrete treatment and also deals with estimation of ATT, we consider the case of a binary treatment and focus on estimation of the ATE to make the presentation simpler. We call Z the original set of covariates and $X = P(Z)$ a p-dimensional transformation of these raw covariates. We are in a setting where we have an iid sample $\{(y_i, d_i, z_i^T)\}_{i=1}^n$ of (Y, D, Z^T), where as in the rest of this work, Y is the outcome of interest, D is the treatment and Z is a set of covariates. We denote $\mu_t = \mathbb{E}(Y_t)$, for $t = 0, 1$ describing either the control group or the treated. In this setting, $\mu_1 - \mu_0$ is the ATE. The outcome function and the propensity scores are denoted by $\mu_1(Z) = \mathbb{E}(Y|Z, D = 1)$, $\mu_0(Z) = \mathbb{E}(Y|Z, D = 0)$ and $p_t(Z) = \mathbb{P}(D = t|Z)$ for $t = 0, 1$. The estimator of the ATE considered here is the Doubly-Robust estimator:

$$\hat{\mu}_t = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}\{d_i = t\}(y_i - \hat{\mu}_t(z_i))}{\hat{p}_t(z_i)} + \hat{\mu}_t(z_i) \tag{17}$$

Following widespread empirical practice, the outcome functions are linear in the covariates and the propensity score is logistic.

The estimator proposed by Farrell (2015) also uses three steps that echo those of Belloni *et al.* (2014b):

(1) Estimate the propensity score using the Logit-Lasso to select the right confounding factors.

(2) Estimate both outcome equations using the Group-Lasso to select the main determinants of the outcome both for the treated and the control groups.

(3) Estimate the ATE using the Doubly Robust estimator with the propensity score obtained in (1) and outcome equations obtained in (2).

One of the nice feature of this estimator is its double robustness property. Indeed, it is still consistent even if either the outcome equations or the propensity score is ill-specified. The next section completes this more intuitive presentation and can be skipped at first.

III.2.2 Assumptions and Main Results

To ensure good estimation of the quantities of interest two assumptions are required:

Assumption III.4 (Identification)
For $t = 0, 1$ and almost surely, Z,P_n obeys:

(1) (Mean independence) $\mathbb{E}(Y_t|D, Z) = \mathbb{E}(Y_t|Z)$

(2) (Overlap) $\mathbb{P}(D = t|Z = z) \geq p_{\min} > 0$

This first part of the assumption is similar albeit weaker than the CIA and is very common in this literature as we have argued before. The second assumption is necessary to be able to use inverse weighing.

Assumption III.5 (Data-Generating Process)
For each n, P_n obeys:

(1) $\{(y_i, d_i, z_i^T)\}_{i=1}^n$ *is an iid sample of* (Y, D, Z^T).

(2) The transformed covariates X have bounded support, with $\max_{j=1,\dots,p} |X_j| \leq \mathbb{X} < \infty$ *uniformly in n. Transformations may depend on n but not the underlying data generating process.*

(3) $\mathbb{E}(|U|^4|Z) \leq \mathbb{U}^4$, *uniformly in* n.

(4) $\min_{j=1,...p;t=0,1} \mathbb{E}(X_j^2 U^2) \wedge \mathbb{E}(X_j^2(\mathbf{1}\{D=t\} - p_t(X))^2)$ *is bounded away from zero, uniformly in* n.

(5) *for some* $r > 0$: $\mathbb{E}(|\mu_1(z_i)\mu_0(z_i)|^{1+r})$ *and* $\mathbb{E}(|u_i|^{4+r})$ *are bounded, uniformly in* n.

Theorem III.2 (Uniformly Valid Treatment Effect Inference)
Consider a sequence $\{P_n\}$ *of data-generating processes that obeys, for each* n, *the two previous assumptions. Provided that the two following conditions hold:*

(1) $\sum_{i=1}^n (\hat{p}_t(z_i) - p_t(z_i))^2/n = o_{P_n}(1)$ *and* $\sum_{i=1}^n (\hat{\mu}_t(z_i) - \mu_t(z_i))^2/n = o_{P_n}(1)$

(2) $\left[\sum_{i=1}^n \mathbf{1}\{d_i = t\}(\hat{p}_t(z_i) - p_t(z_i))^2/n\right]^{1/2} \left[\sum_{i=1}^n \mathbf{1}\{d_i = t\}(\hat{\mu}_t(z_i) - \mu_t(z_i))^2/n\right]^{1/2} = o_{P_n}(n^{-1/2})$

Then:

$$\sup_{P \in P_n} \left| \mathbb{P}_P \left[\mu_t \in \left\{ \hat{\mu}_t \pm c_\alpha \sqrt{\hat{V}_t/n} \right\} \right] - (1-\alpha) \right| \to 0 \tag{18}$$

where c_α *is the quantile of level* $1 - \alpha/2$ *of a standard Normal variable.*

The first interest of this theorem is that this confidence interval is uniformly valid for all DGP, even though we have performed variable selection in previous steps. This is again circumvents the results of Leeb and Pötscher (2005, 2006). However, the first assumption of this theorem means that both the propensity scores and the outcome equations are consistently estimated: in order to have the asymptotic normality for the ATE estimator, *both* features of the model have to be well specified ! Hence, the double robustness property does not translate to the asymptotic normality. This was not the case to obtain consistency of the ATE, for which consistency of only one of the two features is essential. If either one outcome equation or the propensity score is ill-specified then there would still be asymptotic normality but there would remain a bias.

III.2.3 Propensity Score and Outcome Equation Estimators

In order to satisfy the two higher-level conditions of Theorem III.2, the model has to be specified more precisely. The propensity score is Logit:

$$\log\left(\frac{p_1(z)}{p_0(z)}\right) = x_D^T \gamma + B^D(x)$$

and both outcome equations are linear:

$$\mu_t(z) = x_Y^T \beta_t + B_t^Y(x),\ t = 0, 1$$

The terms B^D and B_t^Y are approximations errors due to the parametric specification. The approximate sparsity assumption in this context means that only a few number of variables X are needed to make the bias small. The remainder of this section exposes the main features of the estimators used for the outcome equations and the propensity score.

Propensity Score Estimation The propensity score is estimated through a Logit-Lasso which validity has been established in Van de Geer (2008). We denote γ the coefficients associated with the propensity score:

$$\hat{\gamma} \in \underset{\gamma \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} d_i \log(1 + e^{-x_i^T \gamma}) + (1 - d_i) \log(1 + e^{x_i^T \gamma}) + \lambda_D \sum_{j=1}^{p} |\gamma_j| \tag{19}$$

This program is similar to maximization of the log-likelihood except that non-zero entries in γ are penalized by the ℓ_1-norm to end up with a sparse solution.

Outcome Equations Estimation We denote by β_1, β_0 the coefficients associated with the outcome function in the treated and control cases. The outcome equations are estimated using a Group-Lasso:

$$(\hat{\beta}_1, \hat{\beta}_0) \in \underset{\beta_1 \in \mathbb{R}^p, \beta_0 \in \mathbb{R}^p}{\arg\min} \frac{1}{n_1} \sum_{i, D_i = 1} (y_i - x_i^T \beta_1)^2 + \frac{1}{n_0} \sum_{i, D_i = 0} (y_i - x_i^T \beta_0)^2 + \lambda_Y \sum_{j=1}^{p} \sqrt{\beta_{1,j}^2 + \beta_{0,j}^2} \tag{20}$$

This program is similar to the Lasso program we stated in Section II, but the penalty somewhat differs. The use of this mixed $\ell_{(1,2)}$ penalty reflect the belief that across the equations, the coefficients β_1 and β_0 are included in the same sparsity pattern. In other words, it means that the predictors that are relevant to predict the outcome are assumed to be the same across the treated and the control groups, which seems intuitively reasonable. We refer the reader to Lounici *et al.* (2011) for a more advanced discussion of this assumption.

Under several technical assumptions, both these estimators have good selection and estimation properties that we will not state to keep the discussion clear, and abide by all the conditions required to apply Theorem III.2.

III.3 Considerations on the Tuning Parameter Choice

III.3.1 Theoretical Results: Infeasible Choices

As we have seen with the toy example of Section II, the Lasso suffers from the fact that the penalty parameter λ has to be set in a very specific way to yield good properties. As both Farrell (2015) and Belloni *et al.* (2014b) use the Lasso as their selection device, this issue needs to be addressed.

In Belloni *et al.* (2014b) They consider a generic Lasso estimator with variable-specific penalty loadings:

$$\hat{\beta}^{\lambda} \in \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2 + \lambda \sum_{j=1}^{p} l_j |\beta_j|$$

For step (1) - the selection step - y_i would be replaced by d_i. The optimal choices for the overall penalty parameter λ and the penalty loadings l_j in each step that requires selection are:

$$\lambda = 2c\sqrt{n}\Phi^{-1}(1 - \gamma/2p) \tag{21}$$

$$l_j = \sqrt{\mathbb{E}_n(x_{ij}^2 \varepsilon_i^2)}$$

Unfortunately, these choices are not available since ε_i, the residuals of a given equation, is not observed. Moreover, c is user-defined constant for which the authors give the rule of thumb: $c \approx 1.1$. γ is a confidence level that should be small: the authors give the guideline $\gamma \approx 1/n \wedge .05$. These choices seem arbitrary, but work relatively well in practice.

In Farrell (2015) The paper is interesting in this regard since the optimal penalty choices are displayed and discussed. Intuitively, the penalty should be chosen so as to dominate the noise. We refer the reader to Section III.2 for the expression of the estimators. The optimal choices are:

$$\lambda_D = \frac{2\mathbb{X}}{\sqrt{n}} \left(1 + \log(p \vee n)^{3/2+\delta_D}\right)^{1/2}$$

$$\lambda_Y = \frac{4\sqrt{2}\mathbb{X}\mathbb{U}}{\sqrt{\min(n_1, n_0)}} \left(1 + \frac{\log(p \vee \min(n_1, n_0))^{3/2+\delta_Y}}{\sqrt{2}}\right)^{1/2}$$

for some strictly positive real number δ_D, δ_Y. The double robustness property of the treatment effect estimator must reduce the sensitivity to the penalty choice. However, it is still not free of arbitrary choices for the constants δ_D and δ_Y. Moreover, if estimating $\mathbb{X}$ is easy by taking $\max_{i=1,...,n;j=1,...,p} x_{ij}$, estimating $\mathbb{U}$ is not since the residuals are not observed.

III.3.2 The Usual Approach: Iterative Procedure

To overcome the infeasibility of these approaches, Belloni *et al.* (2012) (Lemma 11) have developed an iterative procedure which is used in Farrell (2015); Belloni *et al.* (2014b). We illustrate this algorithm in the case of Belloni *et al.* (2014b).

The penalty loadings are estimated by the following algorithm:
Set a small constant $v > 0$ and a maximal number of iterations K.

(1) Start by setting $l_j^{(0)} = \sqrt{\mathbb{E}_n(x_{ij}^2 y_i^2)}$, $j = 1, ..., p$. For step k, set $l_j^{(k)} = \sqrt{\mathbb{E}_n(x_{ij}^2 \hat{\varepsilon}_i^{(k)2})}$, $j = 1, ..., p$.

(2) Estimate the model by Lasso using the overall penalty level as in equation 21 and penalty loadings found previously, to obtain $\hat{\beta}^{(k)}$.

(3) Estimate the model by OLS only including variables contained in $\text{supp}(\hat{\beta}^{(k)})$, to obtain $\hat{\varepsilon}_i^{(k)}$.

(4) Stop if $\max_{j=1,...,p} |l_j^{(k)} - l_j^{(k-1)}| \leq v$ or $k > K$. Set k=k+1 and go to step 1 otherwise.

This procedure is asymptotically valid. However, applying it in practice requires to be careful when choosing the starting values as we have found empirically that the algorithm may get stuck in a sub-optimal interval. In practice, no more than five to ten iterations are necessary to obtain stable penalty loadings.

III.3.3 k-fold Cross-Validation

A popular method amongst practitioners to find the optimal penalty level in Lasso-type methods is to use k-fold cross-validation so as to estimate the Mean Squared Error (or the expected classification error for discrete outcome variables) and minimize it. We illustrate this procedure in the context of a generic linear Lasso estimator of the type:

$$\hat{\beta}^{\lambda} \in \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2 + \lambda \|\beta\|_1$$

(1) Randomly divide the dataset into k sub-samples and take a penalty level λ.

(2) For each of these sub-samples (denoted by $j = 1, ..., k$), do the following:

   (a) Take out sub-sample j

   (b) Estimate the model on the training sample constituted by the $k-1$ other sub-samples, for the penalty level λ, to get $\beta^{\lambda}_{-(j)}$

   (c) Compute the MSE or the classification error on sub-sample j using estimated parameter $\beta^{\lambda}_{-(j)}$

(3) Combine the error's estimates for each $j = 1, ..., k$ sub-sample to obtain a single number.

(4) Minimize in λ.

The procedure is acknowledged to work well in practice even though theoretical justification has yet to be proven outside very simple settings. In practice, many authors recommend using 5 or 10-fold cross-validation. See Hastie *et al.* (2009, p. 241) for a more detailed presentation of the subject.

III.3.4 Tuning Parameter-Free Estimators

Other approaches have tried to overcome the penalty choice problem by modifying the Lasso program in order to find an optimal solution which is free of any tuning parameter while keeping the appealing properties of the Lasso such as the kink that fixes some estimated coefficients at zero.

In an heteroskedastic case, the main problem is that the optimal penalty level is expressed as a function of the variance of the error term. Recall that in Theorem II.2, the penalty level is set at $\lambda = 2\sigma\sqrt{2\log(p)/n}$. The econometrician however does not know σ and could, for example, perform a two-step approach with a first estimate of σ and then use the Lasso. This is the basis of the iterative approach we have seen. In a slightly more general case, when the (ORT) assumption does not hold, the optimal penalty level is such that $\lambda \approx \sigma\|X^T\varepsilon\|_\infty/n$. It is obvious that since the error vector ε is not observed this optimal penalty choice is not feasible. The main approach to get rid of this unknown variance is the Square-Root Lasso of Belloni *et al.* (2011), which solves the following program:

$$\hat{\beta}^{\gamma} \in \arg\min_{\beta} \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^T\beta\right)^2}}{\sqrt{n}} + \gamma\|\beta\|_1 \tag{22}$$

The optimal γ choice in this set-up does not depends on σ, the standard deviation of the error terms, and should be set in the following way: $\gamma \approx \|X^T\varepsilon\|_\infty/n$. Intuitively, dividing by $\sqrt{\sum_{i=1}^n \left(y_i - x_i^T\beta\right)^2/n}$ the square-ℓ_2 loss in the Lasso program cancels the σ in the optimal penalty choice because $\sqrt{\sum_{i=1}^n \left(y_i - x_i^T\beta\right)^2/n}$ acts as an inherent estimator of σ. Gautier and Tsybakov (2011) also build along those lines.

In a promising new attempt to circumvent the tuning parameter choice, Lederer and Müller (2014) proposes a Lasso-type estimator which is completely free of any arbitrary calibration, called TREX. Building on the square-root Lasso by finding inherent estimators of the penalty choice, the idea is to use the following estimator:

$$\hat{\beta}^{TREX} \in \arg\min_{\beta} \frac{\sum_{i=1}^n \left(y_i - x_i^T\beta\right)^2}{\frac{1}{2}\max_{j=1,\dots,p}\sum_{i=1}^n x_{i,j}\left(y_i - x_i^T\beta\right)} + \|\beta\|_1 \tag{23}$$

The TREX is close to the Lasso with the optimal penalty choice and can outperform cross-validated Lasso in terms of variable selection and computational efficiency. We note however that this program isn't convex and may entail computational difficulties. Moreover, formal proofs of its efficiency have not been published yet.

These estimators could be used within the context of Farrell (2015) and Belloni *et al.* (2014b) provided that they verify the required higher-level assumptions.

IV Simulation Study

This section illustrates the benefits of the double selection and the Lasso in a classical policy evaluation setting, by a simple Monte Carlo experiment.

Model IV.1 (Monte Carlo Experiment: Data-Generating Process)

$$y_i = d_i\alpha + x_i^T\beta + \varepsilon_i$$
$$d_i = x_i^T\gamma + v_i$$

The sample size is $n = 200$. ε_i *and* v_i *are independent random variables of distribution* $\mathcal{N}(0,1)$. d_i *is of dimension 1.* x_i *is a vector of dimension* $p = 150$ *independent of* ε_i *and* v_i*, and distributed as* $\mathcal{N}(0,\Sigma)$ *with* $\Sigma_{jk} = .5^{|j-k|}$. *The treatment effect is set to be null* $\alpha = 0$.

Two scenarios are considered:

- An approximately sparse setting with coefficients: $\beta_j = (-1)^{j+1}B/j^b$ and $\gamma_j = (-1)^j G/j^g$, $j = 1, ..., p$.

- An exactly sparse setting were the previous coefficients are set to zero except for the first five coefficients that keep the same value as above.

The parameter of interest is the treatment effect α, while β and γ are nuisance parameters. Both scenarios are relatively high-dimension in the sense that the number of variables to be considered is proportional to the sample size. The first scenario is more complicated to deal with because even though coefficients decay to zero, they are not exactly zero, so selection is expected to be more uncertain. Several estimators of the treatment effects will be of interest:

(1) An OLS estimator from the full model, including all the possible covariates.

(2) A post-single selection estimator, where a Lasso regression of y_i on d_i and x_i is used to select relevant elements of x_i.

(3) A post-double selection estimator as in Belloni *et al.* (2014b).

An OLS estimator of α in the model including all the covariates is expected to be unbiased but with high-variance. A single post-selection estimator of α is likely to be very biased but with low variance. The post-double selection should offer a better trade-off. Results from the approximately sparse scenario are displayed in Figure 3 and results from the exactly sparse scenario are displayed in Figure 4.

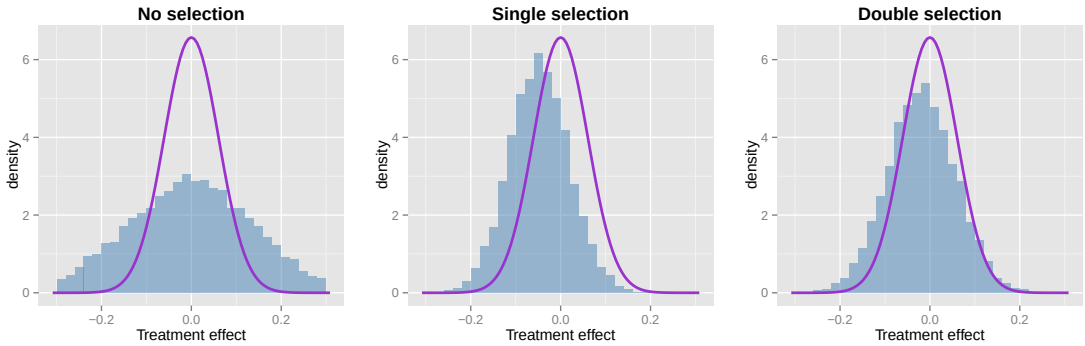
In both scenario, the post-double selection estimator offers a better bias-variance trade-off than the other two. The estimator that uses all the covariates is imprecise, while the single-selection estimator is very biased. The double-selection estimator performs well even in a case where all covariates have a non-zero coefficient.

V Empirical Application

V.1 Presentation of the Dataset

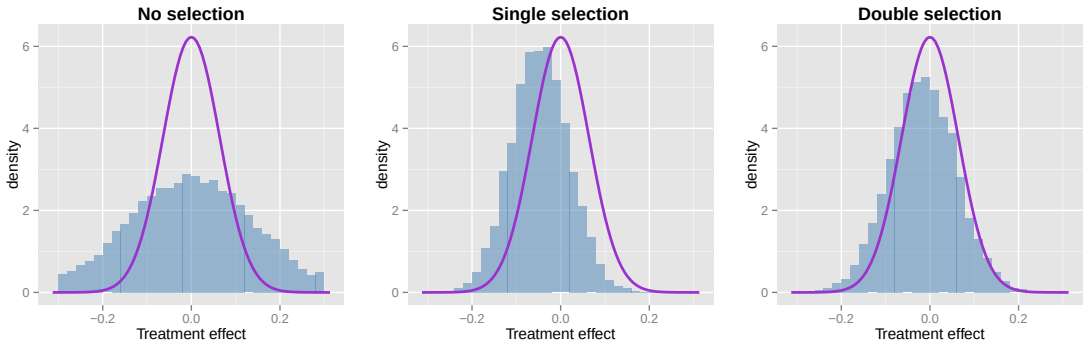
This last section is an empirical illustration of the theoretical propositions of Belloni *et al.* (2014b) and Farrell (2015) revisiting the famous LaLonde (1986) dataset. It is concerned with analysis of the National Supported Work (NSW) demonstration which is a transitional, subsidized work experience program targeted towards people with long-standing employment problems: ex-offenders, former drug addicts, women who were

Figure 3: Approximately Sparse Scenario



Note: Treatment effect estimates densities. On the left panel the full model with $p = 150$ covariates is considered. On the middle panel, single selection on equation of y_i. On the right panel, double selection on equation of y_i and d_i. The purple curve is the Normal density of mean 0 and standard deviation equals to the post-double selection asymptotic standard deviation. Results from a Monte-Carlo experiment replicated 10,000 times. Parameters are set as: $B = 1, G = 2, b = g = 2$. Sample size is $n = 200$.

Figure 4: Exactly Sparse Scenario



Note: See Figure 3.

long-term recipients of welfare benefits and school dropouts. The treated group gathers people who were randomly assigned to this program amongst the population at risk ($n_1 = 185$). The interesting feature of this dataset is that two control groups are available. The first one is experimental: it is directly comparable to the treated group as it has been generated by a Random Control Trial (RCT) ($n_0 = 260$). The second one comes from observationnal data: it is a sample from the Panel Study of Income Dynamics (PSID) ($n_0 = 2490$). Appendix B describes the data using charts and tables. The aim of the paper by LaLonde (1986) is to investigate whether results from carefully designed RCTs are reproducible using a selection-on-observables strategy on observational

data. In the original paper, the answer appears to be negative, potentially implying that selection into the treatment operates on unobservable characteristics. However, it might also be the case that a missing variable bias remains or that the control function is not well approximated. The interested readers are deferred to Dehejia and Wahba (1999, 2002); Smith and Todd (2005) for the controversy regarding econometric estimates of nonexperimental causal studies that used LaLonde (1986)'s data: I merely use the data as an illustration of the estimators we have presented in the paper.

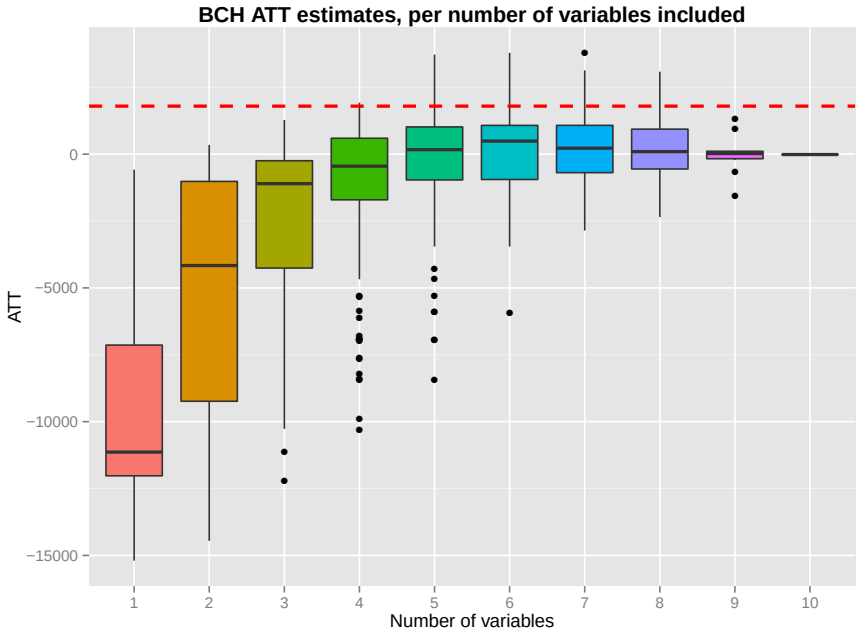
Here, the quantity of interest is the ATT defined as the impact of the participation into the program on 1978 yearly earnings in dollars: the specificity of the targeted population, people with longstanding employment problems who also suffer from intricate socio-economic problems, implies that this program was never designed to be applied to the whole U.S. labour force and that the ATE may not be a good indicator of its effectiveness. The distribution of the outcome across treated and non-treated is given in the left panel of Figure 9. It appears clearly that the control group earns overall more than the treated. Considering however the substantial differences between those two populations in terms of characteristics, it should not come as a surprise. If we take a look at the middle panel of Figure 9 which plots an estimate of the distribution of the difference between earnings in 1975 and 1978, we can see that the treated seem to have enjoyed a larger wage increase. These two graphs are of course no evidence in favour or against the NSW program since being part of the program is highly endogenous. Table II shows the tremendous socio-demographic differences between the two groups and the right panel of Figure 9 displays the schooling years distribution amongst the two groups.

V.2 Exhaustive Model Selection and Uncertainty Regarding ATT Estimates

The following exercise justifies the use of the selection devices advocated by Belloni *et al.* (2014b) and Farrell (2015) by showing how wide the range of possible ATT estimates can be depending on the model. For this, we estimate the ATT given by all the possible models with variables taken amongst the 10 variables of the original dataset (*age, education, black, hispanic, married, no degree, income in 1974, income in 1975, no earnings in 1974, no earnings in 1975*) in two cases. The first case, echoing Belloni *et al.* (2014b), is a linear regression model estimated by OLS. In this case, there are 2^{10} possible sets of regressors to be included in the controls, so 2^{10} possible values for the ATT. The result is displayed in Figure 5. The second case, echoing Farrell (2015), uses

a doubly-robust estimate of the ATT. In this case, there are $2^{10} \times 2^{10}$ possible models because two models have to be selected: the propensity score and the outcome functions. Results are displayed in Figures 6.

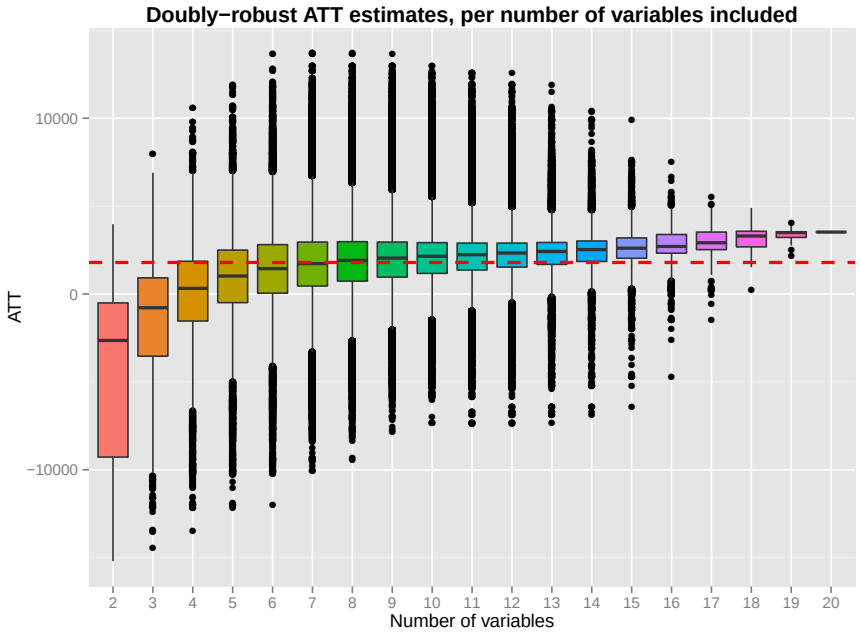
Figure 5: Distribution of ATT estimates, by number of variables, linear estimator



Note: Each boxplot shows the distribution of the ATT estimates for all models that include a given number of variables, using a linear regression. The red dotted horizontal line gives the experimental ATT benchmark. The x-axis unit is the number of variables included as controls in the right-hand side of the equation.

The value of $ -15,000, the difference between 1978 earnings for treated and control individuals is a lower bound for the ATT. Indeed, it is the naive estimator that suffers from a large downward bias because of the heterogeneity of the two groups. From Figure 5 it appears that there is a wide range of possible ATT estimates, depending on the number of variables to be included in the controls. Too few variables give an underestimated ATT. Although the ATT estimate increases with the number of variables, the experimental benchmark remains at the upper tail of the distribution. Moreover, in a usual empirical application where there is no experimental benchmark, the econometrician would not know how to select the ATT estimate just by looking at this chart. The same exercise conducted with a doubly robust estimators yields similar conclusions (see Figure 6). In this case, we have estimated more than one million models. The distribution of ATT estimated this way appear to be very roughly centered around the experimental benchmark. By breaking down these estimates by the number of variables included in the model, Figures 5 and 6 show how wide the range of possible ATT estimates can be

Figure 6: Distribution of ATT estimates, by number of variables, doubly-robust estimator



Note: Each boxplot shows the distribution of the ATT estimates for all models that include a given number of variables, using a doubly robust estimator. The red dotted horizontal line gives the experimental ATT benchmark. The x-axis unit is the sum of the number of variables included in the propensity score and the number of variables included in the outcome functions.

depending on the size of the model. A clear conclusion can be drawn from both charts: always including more variables in the model does not mean that the treatment effect estimate will be more plausible ! Here, the median ATT estimates for models with 7 to 10 variables fall close to the experimental benchmark. Both these charts also call for considering more flexible functional forms and potential transformations of raw covariates as a way to improve the ATT estimation. These exercises give us a good sense on how large the uncertainty regarding the ATT can be. On the other hand, the exercise is feasible because we only have ten variables. The Lasso-like selection devices provided by Belloni *et al.* (2014b) and Farrell (2015) must help in selecting a model that will give the correct ATT estimate amongst all those possible values.

V.3 Comparison of Several Estimators

This section compares the performance of the estimators reviewed in section III on LaLonde (1986) dataset. The `R` files for replications are available upon request. To allow for a flexible model, we take the raw covariates of the dataset (*age, education, black, hispanic, married, no degree, income in 1974, income in 1975, no earnings in 1974, no earnings in 1975*), two-by-two-interactions between the four continuous variables and

the dummies, two-by-two interactions between the dummies and up to a degree of order 5 polynomial transformations of continuous variables. All in all, we end up with 172 variables to select from.

We focus on each of the following estimators of the ATT that corresponds to each columns of Table I:

(1) The experimental estimator, obtained on experimental data from the RCT (benchmark).

(2) Farrell (2015) doubly-robust estimator from Section III.2, with penalty levels set as $\delta_Y = 2$ and $\delta_D = 3$ (author recommended values).

(3) Belloni *et al.* (2014b) post-double selection estimator relying on the Lasso as in Section III.1 with the theoretical penalty level estimated by the iterative procedure and tuning parameters set according to recommendations from the authors.

(4) An OLS estimation of the counterfactual with a Group-Lasso selection step for the outcome equation. The penalty level is set in the same way as in (2). Here, the outcome equation is assumed to be linear and the propensity score unknown and not estimated.

(5) An OLS estimation of the counterfactual with a Lasso selection step for the outcome equation on the control group. Then the counterfactual is an out-of-sample estimate of the baseline outcome for the treated, using the parameters estimated on the control group. The penalty level is set using 10-fold cross-validation. Both (4) and (5) are single-post-selection estimators as described (but criticized) by Belloni *et al.* (2014b).

(6) An Normalized Propensity Weighting (NPW) estimator (Imbens, 2004) where the propensity score is selected via a Logit-Lasso step with penalty level chosen as in (2) according to Farrell (2015)'s recommendation.

(7) An NPW estimator where the propensity score is selected via a Logit-Lasso step with penalty level chosen by 10-fold cross validation.

Table I displays the results.

Table I: Average Treatment Effect on the Treated (ATT) for several estimators

	Estimator:						
	Experimental benchmark	*Farrell (2013) Iterative*	*BCH (2012) Iterative*	*Linear Group Lasso*	*Linear Lasso CV*	*NPW Recomm.*	*NPW Lasso CV*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Estimate	1,794.34	1,737.07	947.27	2,073.86	2,418.64	2,989.31	2,798.02
Standard error (B)	(665.53)	(2,077.86)	(1,096.98)	(1,565.10)	(580.23)	(911.67)	(1,559.80)
Standard error (Asy.)	(671.00)	(736.66)	(816.74)				
Bias (B)	-6.12	-1023.99	110.69	1400.24	2031.19	29.89	-7.64
.95 confidence interval (B)	[519;3046]	[-97;8105]	[-1727;2828]	[-2336;3719]	[-795;1519]	[1084;4763]	[-84;5859]
# variables in Propensity Score	none	7	3	none	none	7	58
# variables in Outcome functions	none	21	18	21	38	none	none
Observations (n)	445	2,675	2,675	2,675	2,675	2,675	2,675
Variables (p)	none	172	172	172	172	172	172
# control retained	260	1,735	2,490	2,490	2,490	1,735	621
# treated retained	185	185	185	185	185	185	185

Note: The experimental estimate is computed on experimental data, column (1). (B) signals that the quantity is computed using a bootstrap method of 1,000 iterations. The bootstrap method used here encompasses the variable selection step but not the tuning parameter choice which is set before the loop starts. (Asy.) signals the asymptotic approximation estimator of the quantity is used. Columns (2) and (3) shows the estimators studied in the central part of the dissertation with penalty parameters set according to authors' recommendations. The size of the control group differs along columns (2)-(7) because when the propensity score is estimated we restrict the estimation on the control sample to individuals for which the common support assumption holds.

Several observations must be made. The first one is a word of caution: bootstrapped quantities are reported to provide a better sense of finite-sample behaviour of the estimators. We use a fully non-parametric bootstrap. The confidence interval bounds come from quantiles of the bootstrapped distribution. However, the validity of the usual non-parametric bootstrap for Lasso estimators has been put into questions (see for example Chatterjee and Lahiri (2011) and references therein). Nevertheless, the Lasso is only used as a selection device here. The quantities of interest, the ATT estimators, are post-selection estimators that are asymptotically normal, which increases confidence in the bootstrap method employed here.

When compared to the benchmark experimental estimator, two estimators give a good approximation of the ATT: Farrell (2015) and the Group Lasso for the outcome equation, in columns (2) and (4). Farrell (2015) in particular seems well-suited to reproduce these results. The good performance of columns (4)-(5) compared to columns (6)-(7) can be interpreted as evidence that the outcome equation is well approximated by a linear combination of transformations of the raw dataset, while a Logit modelling of the propensity score does not work well. As we have seen, the estimator from column (2) displays a double-robustness property which guards against one mistake in specification.

The somewhat disappointing performance of the post-double-selection estimator in column (3) in comparison to the experimental benchmark and to Farrell (2015) estimator could be due to two factors. The first one is the parametric assumption on the treatment effect which is given by a coefficient in a linear regression as we have already noticed. This is very restrictive. The second one is the linear selection step on the treatment dummy: we have computed the same estimator but this time replacing a linear Lasso selection step on the treatment dummy by a Logit-Lasso and found considerable improvement, the ATT being estimated at $ 1,578. Consequently, we conclude that the BCH procedure seems relatively adequate in the sense that the post-double selection seems to work to select a small number of covariates, but the form of the selectors chosen may be challenged. This is all the more acute as we had a hard time obtaining convergence of the iterative algorithm for the treatment dummy explanatory variable selection step.

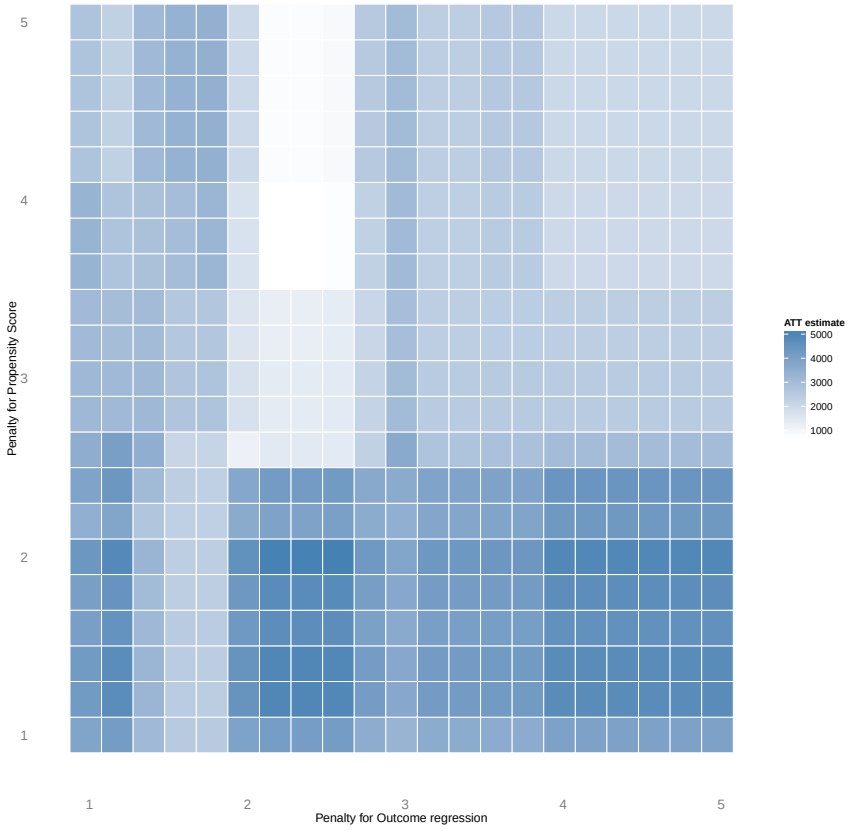
It is to be noted however that the relatively good performance of the estimators in columns (2), (4) and (5) must be questioned. Indeed, their bootstrapped variances

appear to be huge, especially compared to the experimental benchmark and we must not rule out that the fact that these estimators fall close to the experimental benchmark may be due to luck. In particular, column (5) offers poor results: the point estimate of the ATT is not even in the bootstrapped confidence interval ! This offers more evidence that single-selection estimators are indeed not very reliable as already noted in Belloni *et al.* (2014b,a). Table III gives further results, and notably offers a comparison with classical policy evaluation estimators using a model that includes only the original covariates.

V.4 Further Investigation of the Tuning Parameter Choice

In the previous section, for all estimators, the tuning parameter was left unchanged. Now we investigate the sensitivity of the ATT estimates to the tuning parameter.

Figure 7: ATT as a function of both penalty levels

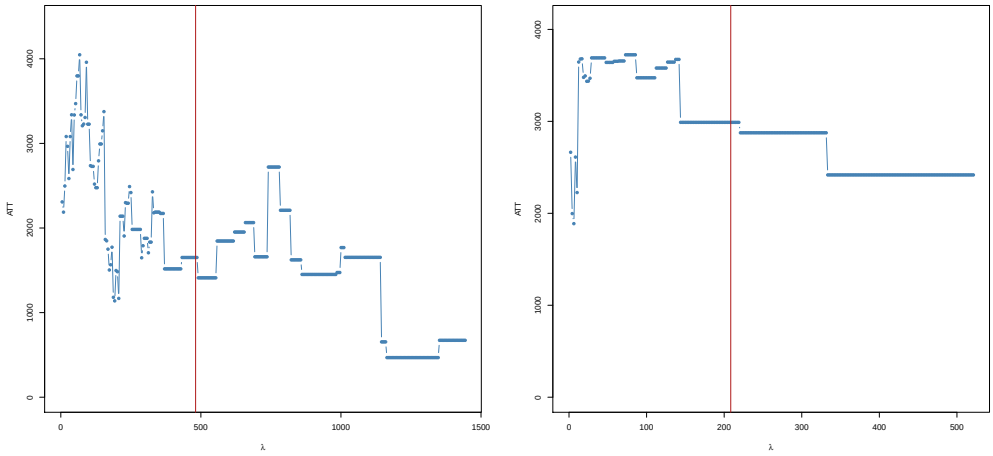


Note: Each square represents the ATT estimate corresponding the outcome functions selected with the x-coordinate penalty level and the propensity score selected with the y-coordinate penalty level.

Figure 7 plots the doubly-robust estimator (referred to as (2) in the previous subsection) of the ATT as we change δ_Y and δ_D, the penalty parameters levels for the outcome

regression and propensity score in Section III.3.1 respectively. Both penalties move on a grid that goes from 1 to 10 by .2 unit. Surprisingly enough, the results are found to be relatively sensitive to the penalty levels, despite the double-robustness property of the estimator. Indeed, it was expected that the double robustness would prevent sharp jumps in the treatment effect estimated as the penalty level moves for one Lasso estimator while it stays constant for the other. The maximum of the estimated ATT is $ 5,126, the minimum $ 749 and the median $ 2,755.

Figure 8: ATT as a function of penalty levels for the Lasso counterfactual (left) and the NPW with Logit propensity score (right)



Note: The red vertical lines represents the cross-validation selected penalty level (on the left) and the recommended value (on the right).

Figure 8 plots the ATT estimate with the linear estimator of counterfactual where the model is selected using a Lasso on the control group (column (5) in Table I, left panel in the Figure), and with the NPW estimator where the propensity score (column (6)-(7) in Table I), as a function of the penalty level for each. The linear Lasso estimator seems relatively sensitive to the chosen penalty level. At the left of the graph, the model selects too many controls and the ATT estimate appears very volatile. Towards the right, after the red line, the ATT estimate stabilises around the experimental benchmark but nevertheless displays substantial sensitivity. The NPW estimator appears more stable but very much biased.

VI Conclusion

Policy evaluation requires careful model and covariate selection that could gain transparency from more automatic procedures. Of course, these mathematical tools are not designed to avoid sound economic reasoning, but rather to complete it. The Lasso-based approach is emerging as a way to perform automatic model selection, approximate control functions when needed, and provide robust estimates of treatment effects. Borrowing from the high-dimension literature, it focuses on the practical implementation of variable selection and is computationally fast. In a policy evaluation context, the Lasso-based estimators proposed by Belloni *et al.* (2014b) and Farrell (2015) appear to be of great help when one needs to select amongst all possible models to come up with plausible treatment effect. These post-selection estimators also have a nice uniform post-selection inference property, which is a great solution to a problem that used to put econometricians in a deadlock for decades. In spite of their appealing theoretical properties, the difficulties posed by the choice of the tuning parameter cannot realistically be avoided. As the empirical part of the paper has shown, Estimation of the treatment effect where the outcome functions or propensity score estimation relies on ℓ_1-penalized selectors, even if conducted in a doubly-robust context are very sensitive to the tuning parameter choice for which no clearly-defined rule exists. Two alternative exist. On the one hand, the iterative procedure to estimate the optimal penalty level seems to work relatively well and in few iterations. The arbitrary parameters for which no rules exist but authors' recommendations provide a good start. On the other hand, cross-validation can provide an interesting tool to set these tuning parameters but relies on the randomness of the sample division.

A Mathematical Tools

This section gather mathematical results that bear no interest in themselves but are useful for several proofs.

Lemma A.1

If $\eta \sim \mathcal{N}(0,1)$*, then* $\forall x > 0$*:*

$$\mathbb{P}(|\eta| > x) \leq \sqrt{\frac{2}{\pi}} \frac{1}{x} e^{\frac{-x^2}{2}} \tag{24}$$

Proof.

$$\int_x^\infty e^{-\frac{u^2}{2}} du \leq \int_x^\infty \frac{u}{x} e^{-\frac{u^2}{2}} du = \frac{1}{x} e^{-\frac{x^2}{2}}$$

□

Lemma A.2 (Gaussian Concentration)

If $\eta_j \sim \mathcal{N}(0,1), \forall j$*, then* $\forall p \geq 2$*:*

$$\mathbb{P}(\max_{j=1,\dots,p} |\eta_j| > \sqrt{2\log(p)}) \leq \frac{1}{\sqrt{\pi \log(p)}} \tag{25}$$

Proof. Applying the union bound and the previous lemma gives the result. □

B Data Analysis of LaLonde (1986)

Figure 9: Income and Education Distributions for Treated and Non-Treated

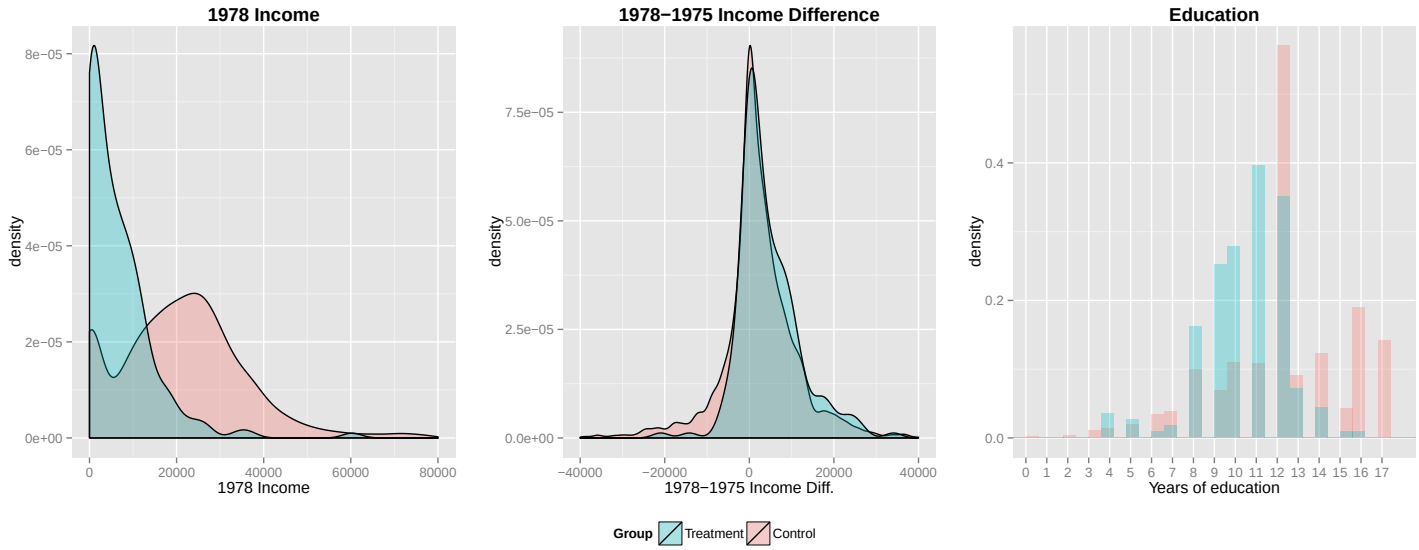


Table II: Descriptive statistics (mean per group)

	Group:	
	Treated	*Control*
	$(D = 1)$	$(D = 0)$
Age	25.82	34.85
Education	10.35	12.12
Black	0.84	0.25
Hispanic	0.06	0.03
Married	0.19	0.87
No degree	0.71	0.31
Income 1974	2,095.57	19,428.75
Income 1975	1,532.06	19,063.38
Income 1978	6,349.15	21,553.92
No Income in 1974	0.74	0.09
No Income in 1975	0.60	0.10
Observations	185	2490

C Empirical Application: Further Results

Table III: Average Treatment Effect on the Treated (ATT) for several estimators

	Estimator:				
	Experimental benchmark	*Linear benchmark*	*NPW benchmark*	*Farrell (2013) iterative*	*BCH (2012) iterative*
	(1)	(2)	(3)	(4)	(5)
Estimate	1,794.34	635.26	3,050.97	1,737.07	947.27
Standard error (B)	(665.53)	(1,155,89)	(888.46)	(2,077.86)	(1,096.98)
Standard error (Asy.)	(671.00)			(736.66)	(816.74)
Bias (B)	-6.12	577,35	-2,388.74	-1023.99	110.69
.95 confidence interval (B)	[519;3046]	[-1991;2255]	[922;4827]	[-97;8105]	[-1727;2828]
# variables in Propensity Score	none	none	10	7	3
# variables in Outcome functions	none	10	none	21	18
Observations (n)	445	2,675	2,675	2,675	2,675
Variables (p)	none	10	10	172	172
# control retained	260	2,490	1047	1,735	2,490
# treated retained	185	185	185	185	185

Note: The experimental estimate is computed on experimental data, column (1). (B) signals that the quantity is computed using a bootstrap method of 1,000 iterations. The bootstrap method used here encompasses the variable selection step but not the tuning parameter choice which is set before the loop starts. (Asy.) signals the asymptotic approximation estimator of the quantity is used. Column (2) displays the result for a conterfactual using parameters obtained by a linear regression of the outcome of the ten original covariates on the control group. Column (3) displays the result from a Normalized Propensity Weighting (NPW) where the propensity score includes the ten original covariates. Columns (4) and (5) shows the estimators studied in the central part of the dissertation with penalty parameters set according to authors' recommendations. The size of the control group differs along columns (2)-(5) because when the propensity score is estimated we restrict the estimation on the control sample to individuals for which the common support assumption holds.

D Bibliography

Angrist, J. D., Òscar Jordà, and Kuersteiner, G. (2013): "Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited". Working Paper 19355, National Bureau of Economic Research.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain". *Econometrica*, 80(6):2369–2429.

Belloni, A. and Chernozhukov, V. (2009): "High Dimensional Sparse Econometric Models: An Introduction". In P. Alquier, E. Gautier, and G. Stoltz, editors, *Inverse Problems and High-Dimensional Estimation*. Springer Publishing Company, Incorporated.

Belloni, A. and Chernozhukov, V. (2013): "Least squares after model selection in high-dimensional sparse models". *Bernoulli*, 19(2):521–547.

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

——— (2014b): "Inference on Treatment Effects after Selection among High-Dimensional Controls". *The Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2014): "Inference in High Dimensional Panel Models with an Application to Gun Control". *ArXiv e-prints*.

Belloni, A., Chernozhukov, V., and Wang, L. (2011): "Square-root lasso: pivotal recovery of sparse signals via conic programming". *Biometrika*, 98(4):791–806.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013): "Valid post-selection inference". *Ann. Statist.*, 41(2):802–837.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009): "Simultaneous analysis of Lasso and Dantzig selector". *The Annals of Statistics*, 37(4):1705–1732.

Buhlmann, P. and van de Geer, S. (2011): *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Publishing Company, Incorporated, 1st edition.

Candes, E. and Tao, T. (2007): "The Dantzig selector: Statistical estimation when p is much larger than n". *Ann. Statist.*, 35(6):2313–2351.

Chatterjee, A. and Lahiri, S. N. (2011): "Bootstrapping Lasso Estimators". *Journal of the American Statistical Association*, 106(494):608–625.

Chatterjee, S. (2013): "Assumptionless consistency of the Lasso". *ArXiv e-prints*.

Dehejia, R. H. and Wahba, S. (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs". *Journal of the American Statistical Association*, 94(448):pp. 1053–1062.

——— (2002): "Propensity Score-Matching Methods For Nonexperimental Causal Studies". *The Review of Economics and Statistics*, 84(1):151–161.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004): "Least angle regression". *Ann. Statist.*, 32(2):407–499.

Fan, J., Lv, J., and Qi, L. (2011): "Sparse High-Dimensional Models in Economics". *Annual Review of Economics*, 3(1):291–317.

Farrell, M. H. (2015): "Robust inference on average treatment effects with possibly more covariates than observations". *Journal of Econometrics*, 189(1):1 – 23.

Gautier, E. and Tsybakov, A. (2011): "High-dimensional instrumental variables regression and confidence sets". *ArXiv e-prints*.

Van de Geer, S. A. (2008): "High-dimensional generalized linear models and the lasso". *Ann. Statist.*, 36(2):614–645.

Givord, P. (2010): "Econometric Methods for Public Policies Evaluation". Technical report.

Hastie, T., Tibshirani, R., and Friedman, J. (2009): *The Elements of Statistical Learning: Data mining, Inference and Prediction*. Springer, 2nd edition.

Huang, J., Horowitz, J. L., and Wei, F. (2010): "Variable selection in nonparametric additive models". *Ann. Statist.*, 38(4):2282–2313.

Imbens, G. (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review". *The Review of Economics and Statistics*, 86(1):4–29.

Imbens, G. W. and Rubin, D. B. (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 1st edition.

Imbens, G. W. and Wooldridge, J. M. (2009): "Recent Developments in the Econometrics of Program Evaluation". *Journal of Economic Literature*, 47(1):5–86.

Jalan, J. and Ravallion, M. (2003): "Does piped water reduce diarrhea for children in rural India?" *Journal of Econometrics*, 112(1):153–173.

Jing, B.-Y., Shao, Q.-M., and Wang, Q. (2003): "Self-normalized Cramér-type large deviations for independent random variables". *Ann. Probab.*, 31(4):2167–2215.

LaLonde, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". *American Economic Review*, 76(4):604–20.

Leamer, E. E. (1983): "Let's Take the Con Out of Econometrics". *The American Economic Review*, 73(1):31–43.

Lederer, J. and Müller, C. (2014): "Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX". *ArXiv e-prints*.

Leeb, H. and Pötscher, B. M. (2005): "Model Selection and Inference: Facts and Fiction". *Econometric Theory*, null:21–59.

——— (2006): "Can one estimate the conditional distribution of post-model-selection estimators?" *Ann. Statist.*, 34(5):2554–2591.

Lounici, K. (2008): "Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators". *Electronic Journal of Statistics*, 2:90–102.

Lounici, K., Pontil, M., van de Geer, S., and Tsybakov, A. B. (2011): "Oracle inequalities and optimal inference under group sparsity". *The Annals of Statistics*, 39(4):2164–2204.

Sala-i Martin, X. (1997): "I Just Ran Two Million Regressions". *American Economic Review, American Economic Association*, 87(2):178–83.

Meinshausen, N. and Yu, B. (2009): "Lasso-type recovery of sparse representations for high-dimensional data". *The Annals of Statistics*, 37(1):246–270.

Robinson, P. (1988): "Root- N-Consistent Semiparametric Regression". *Econometrica*, 56(4):931–54.

Rubin, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies". *Journal of Educational Psychology*.

Schwarz, G. (1978): "Estimating the Dimension of a Model". *The Annals of Statistics*, 6(2):461–464.

Smith, J. and Todd, P. (2005): "Does matching overcome LaLonde's critique of nonexperimental estimators?" *Journal of Econometrics*, 125(1-2):305–353.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society (Series B)*, 58:267–288.

Tsybakov, A. B. (2008): *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition.

——— (2014): "Stat 681: Nonparametric Estimation and Statistical Learning". Yale University Department of Statistics.

Zhang, C.-H. and Huang, J. (2008): "The sparsity and bias of the Lasso selection in high-dimensional linear regression". *The Annals of Statistics*, 36(4):1567–1594.

Zhao, P. and Yu, B. (2006): "On Model Selection Consistency of Lasso". *J. Mach. Learn. Res.*, 7:2541–2563.