# A Parametric Generalization of the Synthetic Control Method, with High Dimension*

Marianne Bléhaut† Xavier D'Haultfœuille‡ Jérémy L'Hour§
Alexandre Tsybakov¶

February 17, 2017
*Preliminary and incomplete*

**Abstract**

The synthetic control method developed by Abadie *et al.* (2010) is a an econometric tool to evaluate causal effects when only one unit is treated. While initially aimed at evaluating the effect of large-scale macroeconomic changes with very few available control units, it has increasingly been used in place of more well-known microeconometric tools in a broad range of applications, but lacks statistical foundations. This paper proposes a parametric generalization of the synthetic control, which is developed both in the usual asymptotic framework and in the high-dimensional one. The proposed estimator is doubly robust, consistent and asymptotically normal uniformly over a large class of data-generating processes. It is also immunized against first-step selection mistakes. We illustrate these properties using Monte Carlo simulations and applications to both standard and potentially high-dimensional settings, and offer a comparison with the synthetic control method.

**JEL Classification**: C01, C21, C52, C55.

**Keywords**: treatment effect, synthetic control, covariate balancing, high-dimension.

---

*We thank Alberto Abadie, Jann Spiess, Joshua Angrist, Jamie Robins, Arthur Lewbel, Stefan Hoderlein, Richard Blundell, David Margolis, Hyunseung Kang, Laurent Gobillon and participants at the 2016 North American and European Meetings of the Econometric Society, and CREST internal seminars for their useful comments and discussions. We acknowledge funding from Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

†CREST and Université Paris-Sud, Email address: marianne.blehaut@ensae-paristech.fr.

‡CREST, ENSAE, Université Paris-Saclay, Email address: xavier.dhaultfoeuille@ensae-paristech.fr.

§CREST, ENSAE, Université Paris-Saclay, Email address: jeremy.l.hour@ensae-paristech.fr.

¶CREST, ENSAE, Université Paris-Saclay, Email address: alexandre.tsybakov@ensae-paristech.fr.

# 1 Introduction

The synthetic control method developed by Abadie and Gardeazabal (2003); Abadie *et al.* (2010, 2015) is an econometric tool to quantify the effects of a policy change that affects only one or very few aggregate units, using aggregate-level data. The idea is to construct a counterfactual treated unit by taking a convex combination of non-treated units, labeled "synthetic control unit", that closely recreates the characteristics of the treated. The weights given to each control unit are computed by matching the mean of ancillary variables that are relevant to predict the outcome of interest between the treated and the synthetic control unit. The synthetic control method has been used to evaluate causal impacts in a wide range of applications such as terrorism, civil wars and social unrest (Acemoglu *et al.*, 2014), political and monetary unions (Abadie *et al.*, 2015; Wassmann, 2015), minimum wage (Allegretto *et al.*, 2013; Addison *et al.*, 2014), health (Bilgel and Galle, 2015), fiscal policies (Dietrichson and Ellegård, 2015), geographical and regional policies (Gobillon and Magnac, 2014), immigration policy (Bohn *et al.*, 2011), international trade (Nannicini and Billmeier, 2011) and many more. Contrasting with usual microeconometric approaches, the synthetic control method lacks theoretical foundations at the present time and its properties when the number of control units tends to infinity is unknown. While initially aimed at evaluating the effect of large-scale macroeconomic changes with very few available units of comparison (most of the time these units being states or regions), the synthetic control method has increasingly been used in place of more well-known microeconometric tools without clear justification.

This paper proposes a generalization to the synthetic method by using a parametric form for the weight given to each control unit, in hopes to better ground that approach both numerically and theoretically as well as to improve the interpretability of these weights. In the small-dimensional case where the number of observations is much larger than the number of ancillary variables, our approach amounts to a well-known two-step GMM estimator, where the parameters governing the synthetic control weights are computed in a first step so as to match some features of the data between the treated and the synthetic control unit. The proposed estimator is doubly robust in the sense that misspecifications in the synthetic control weights do not prevent valid inference if the outcome equation is linear for the control group. This approach is also extended to the high-dimensional case where the number of observations is smaller or proportional to the number of ancillary variables, and to cases where variable selection is performed. This extension makes the proposed estimator suitable for comparative case studies and macroeconomic applications. Here, the double robustness property helps constructing an estimator which is "immunized" against first-step selection mistakes. In both cases, it is consistent and asymptotically normal uniformly over a large class data-generating processes. Consequently, we develop inference based on asymptotic approximation, linking the synthetic control method with more classical microeconometric tools.

The present paper builds mainly along two lines of the treatment effect literature. The first one is the literature related to propensity score weighting and covariate balancing propensity scores. Several recent efforts have been made to target balance between covariates as an explicit objective with or without relation to the propensity score (*e.g.* Hainmueller (2012); Graham *et al.* (2012)). Very recently, Imai and Ratkovic (2014) convincingly integrated propensity score estimation and covariate balancing in the same framework. Their covariate balancing propensity score method is estimated with GMM and yields much better estimates than traditional propensity score related methods. Indeed, they show that this method is less impacted by potential misspecifications and retains the theoretical properties of GMM estimators. It is to be noted that this idea is related to the *calibration on margins* method used in survey sampling, see for example Deville *et al.* (1993).

It also partakes in the econometric literature that addresses variable selection when estimating a treatment effect, especially but not exclusively in a high-dimensional framework. The lack of uniformity for inference after a selection step has been raised in a series of papers by Leeb and Pötscher (2005, 2008a,b), echoing earlier papers by Leamer (1983) who put into question the credibility of many empirical policy evaluation results. One recent path-breaking solution proposed to circumvent this post-selection conundrum is the use of a double-selection procedure (Belloni and Chernozhukov, 2013; Farrell, 2015; Chernozhukov *et al.*, 2015a). For example, Belloni *et al.* (2014a,b) highlight the danger of selecting controls by only considering the outcome equation and propose a three-step procedure that helps selecting more controls and guards against omitted variable biases much more than a simple "post-single-selection" estimator, as it is usually done by selecting covariates based on either their relation with the outcome or with the treatment variable, but rarely both. 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. In addition, he deals explicitly with a discrete treatment that is a more common setting in the policy evaluation literature. Very recently, Chernozhukov *et al.* (2015b) have theorized this approach by showing how using moments that are first-order-insensitive to the selection step help "immunizing" the inference against selection mistakes. A different path to deal with the problem of propensity score specification has been followed by Kitagawa and Muris (2015) using the Focused Information Criterion (FIC) of Claeskens and Hjort (2003), but it does not explicitly accommodate for a high-dimensional nuisance parameter and assumes that the researcher knows the true model.

The paper is organized as follows. Section 2 reviews the synthetic control method developed by Abadie *et al.* (2010) and underlines the main problems related to that approach. Section 3 introduces a parametric generalization of the Synthetic Control method and states the property of that estimator is cases where the number of variables is lower than the number of observations. Section 4 extends the previous section to the high-dimensional case and

Section 4.2 studies the asymptotic properties of the estimator is that particular case. Section 5 illustrates the good inference properties of the estimator in a Monte Carlo experiment. In Section 6, we use our estimator and compare it with existing ones to Lalonde's dataset and to the evaluation of a large-scale tobacco control program on California tobacco consumption. The appendix gathers the proofs.

# 2 The Synthetic Control method as a tool to perform inference when data is scant

## 2.1 Synthetic Control Estimator

The synthetic control method Abadie and Gardeazabal (2003); Abadie *et al.* (2010, 2015) has initially been developed to answer a research question in a very specific but common setting. Firstly, it aims at estimating the impact of a macroeconomic change that affects every individual of a given group (*e.g.* state, county, city) without leaving anyone untreated within this group. Examples of such shifts are the Mariel Boatlift during which the Miami labor force suddenly grew by 7% (Card, 1990), natural disasters such as hurricanes (Belasen and Polachek, 2008), any state-level policy such as the setting of a minimum wage (Card and Krueger, 1994; Allegretto *et al.*, 2013) or a tobacco control program (Abadie *et al.*, 2010). In many of the previous settings, comparison of individuals that are located across each side of the border should help identifying the effect of the policy. However, the SC starts from the initial constraint that only aggregate data are available to the researcher, meaning that the outcome and/or the covariates will only be observed at the group-level. The scarcity of data at the individual level is a second feature of the synthetic control method. Indeed, in many of the above examples, the classical asymptotic framework (large and growing number of observations) that justifies microeconometric methods is ill-suited or at least insufficient. For example, displaying confidence intervals based on asymptotic approximations when the level of observation is a U.S. State or a European Union member lacks credibility. The synthetic control method is more akin to time series methods in the sense that some theoretical results displayed in Abadie *et al.* (2010) have been obtained by assuming that the number of pre-treatment periods grows. Lastly, most of the time the policy of interest only affects one or a small number of groups. Because there are very few treated units, estimating a propensity score using regular methods such as a logit regression can be difficult and will likely be very imprecise.

To overcome these difficulties, the synthetic control method builds on the intuition that a counterfactual group can be created by taking a convex combination of non-treated groups such that it matches the pre-treatment characteristics of the treated groups. This method

appears to work well in practice and is less demanding in term of data quality than classical treatment effect estimators, which explains its success especially for comparative case studies. It has also enjoyed a recent popularity among econometricians using microdata [CITE REFERENCES] but the benefits of using that specific method in these settings are unclear.

This paragraph presents the method in details, abstracting from the temporal dimension of the original method Abadie *et al.* (2010). The setting is an iid sample of $n$ units (states, countries, groups) indexed by $i$, $(Y_i, D_i, X_i)_{i=1,...,n}$. The sample is divided between treated units for which $D_i = 1$ and control units for which $D_i = 0$. In a typical synthetic control setting, treated units are very few. For simplicity, assume that only unit one is treated. In a case where several units are treated, one can take a an average of the treated unit outcomes. $Y_{0i}$ and $Y_{1i}$ are the corresponding potential outcomes. $Y_i$ is the outcome measurement which is such that: $Y_i = D_i Y_{1i} + (1 - D_i)Y_{0i}$. The observable characteristics or covariates are denoted by $X_i$. The index $i$ is dropped when unnecessary. The quantity of interest is the Average Treatment Effect on the Treated (ATET) defined as:

$$\theta_0 = \mathbb{E}[Y_1 - Y_0 | D = 1],$$

where $Y_1$ is the potential outcome of a unit when it is treated and $Y_0$ is the potential outcome when it is not treated. $Y_1$ is observed but $Y_0$ is missing. This quantity is the value of the potential outcome, had the policy intervention not been undertaken. To recover a plausible counterfactual, the synthetic control uses a weighted mean of non-treated units outcome, and measure the policy effect using:

$$\hat{\theta} = Y_1 - \sum_{i=2}^{n} w_i Y_i.$$

The problem is to choose weights $w_2, ..., w_n$ that take values between zero and one, and sum to one. These restrictions on the weights prevent data extrapolation, which ADH define as giving the counterfactual values that are outside the convex hull defined by the control units. With $p$ observed pre-intervention covariates denoted by $X$, set the weights so as to match the empirical average of the covariates between the treated and the control groups:

$$\begin{aligned} &\min_w (X_1 - X_0^T w)^T V (X_1 - X_0^T w) \\ &\text{s.t.} \sum_{i=2}^{n} w_i = 1, w_i \geq 0, i = 2, ..., n. \end{aligned}$$

where $X_1$ is a vector of dimension $p$ collecting the value of the covariates for the treated unit, $X_0$ is a matrix of dimension $(n-1) \times p$ given by $X_0 = \left(X_2^T, ..., X_{n+1}^T\right)$, $V$ is a $p \times p$ diagonal matrix that weights each balancing equation according to its importance. ADH recommend including the pre-intervention outcomes in the set of covariates to be matched

andadvocates the use of a data-driven choice for $V$ so as to minimize the mean squared prediction error over the pre-treatment period.

## 2.2 Limitations of the Synthetic Control Method

Despite its popularity, the synthetic control method is not well understood and has several limitations. Firstly, the uniqueness of the solution is not always assured when the number of control units is larger than the number of covariates ($n > p$). Indeed, particularly in cases where the treated unit lies within the convex hull of the control units, there exist an infinite number of convex combinations of the $n$ control units that allow to reproduce the $p$ features of the treated unit. Secondly, because it operates in a framework that refrains from considering a growing number of control units, the classical asymptotic properties of the synthetic control method are unknown. In particular, adding more control units to the dataset does not necessarily helps the estimation in the sense that more parameters needs to be estimated. In a time series setting, ADH show that the bias of the synthetic control estimator decreases as the number of pre-treatment periods increases, provided that the DGP for the outcome is a linear factor model. As a direct consequence of the lack of asymptotically properties, inference with synthetic control estimators has been constructed by inverting Fisher tests which means testing a much stronger null hypothesis and is computationally demanding (Imbens and Rubin, 2015). Furthermore, specifying the sharp null hypothesis can be a daunting task, especially when the treatment effect over several periods of time is considered. Lastly, the procedure advocated by ADH to find the weights is not standard and potentially not convex. This procedure can be viewed as a two convoluted optimization procedures akin to cross-validation. The high-level optimization consists in finding a matrix $\hat{V}$ such that the synthetic control group defined by weights $\hat{w}(\hat{V})$ closely reproduce the pre-treatment evolution of the treated unit outcome. So $\hat{V}$ is chosen to minimize the mean squared prediction error of the outcome variable for the pre-treatment period, i.e. it aims at minimizing $\sum_{t=1}^{T_0} \hat{\theta}_t^2$. Within this high-level optimization lies a lower-level one which: for a given level of $V$ find the weights $\hat{w}(\hat{V})$ to match the treated unit covariates as defined by the above minimization program. Besides potential numerical complications, a side effect of this procedure is that the link between the individual weights and the value of an individual covariates is unclear and could be of potential interest.

To overcome these limitations and offer a more flexible tool, this paper introduces an estimator that builds on the idea of the Synthetic Control while giving a parametric form to the individual weights.

# 3 A Parametric Generalization of the Synthetic Control Estimator

## 3.1 Covariate Balancing Weights and Double Robustness

The setting is an iid sample of $n$ individuals indexed by $i$, $(Y_i, D_i, X_i)_{i=1,...,n}$. The sample is divided between treated individuals for which $D_i = 1$ and the others for which $D_i = 0$ (control individuals). $Y_{0i}$ and $Y_{1i}$ are the corresponding potential outcomes. $Y_i$ is the outcome measurement which is such that: $Y_i = D_i Y_{1i} + (1 - D_i)Y_{0i}$. The observable characteristics or covariates are denoted by $X_i$. The index $i$ is dropped when unnecessary.

The quantity of interest is the Average Treatment Effect on the Treated (ATET) defined as:

$$\theta_0 = \mathbb{E}[Y_1 - Y_0|D = 1].$$

$\mathbb{E}(Y_1|D = 1)$ is straightforward to recover from the data. However, since no individual is observed in both states (treated and not treated), identification of the counterfactual $\mathbb{E}(Y_0|D = 1)$ is more complicated. This is achieved here through the following two classical conditions.

**Assumption 3.1** Nested Treated Support
$\mathbb{P}(D = 1|X) < 1$ *almost surely and* $\pi := \mathbb{P}(D = 1) \in (0, 1)$.

**Assumption 3.2** Mean Independence
$\mathbb{E}(Y_0|X, D = 1) = \mathbb{E}(Y_0|X, D = 0)$.

Assumption 3.1, which is a weak version of the usual common support condition, requires that there are control units for any possible value of the covariates in the population. Because the ATET is the parameter of interest, we are never reconstructing a counterfactual for control units so $\mathbb{P}(D = 1|X) > 0$ is not required. Assumption 3.2 states that conditional on a set of well-identified observed covariates or confounding factors $X$, the expected potential outcome without the treatment is the same for treated and control individuals. This assumption is a weaker form of the classical Conditional Independence Assumption (CIA) : $(Y_0, Y_1) \perp\!\!\!\perp D|X$.

As often in policy evaluation settings (Smith and Todd, 2005), we propose to identify and estimate the counterfactual as a weighted average of non-treated unit outcomes:

$$\theta_0 = \mathbb{E}[Y_1|D = 1] - \mathbb{E}[WY_0|D = 0], \tag{1}$$

where $W$ is a random variable. Popular choices for the weights are the following:

(1) Matching: see Smith and Todd (2005) for more details,

(2) Linear regression: $W = \mathbb{E}[DX^T]\mathbb{E}[(1-D)XX^T]^{-1}X$, which can also be referred to as the Oaxaca-Blinder counterfactual (Kline, 2011),

(3) Propensity score: $W = P(X|D=1)/[1-P(X|D=1)]$,

(4) Synthetic Control: see above.

This paper proposes a $W$ which is a particular solution of the synthetic control. Formally, we look for weights $W$ that (i) satisfy a balancing condition as in the synthetic control method, are (ii) positive and (iii) function of the covariates. The first condition writes:

$$\mathbb{E}[DX] = \mathbb{E}[W(1-D)X]. \tag{2}$$

This condition means that $W$ balances the first moment of the observed covariates between the treated and the control group. Here the definition of the observable covariates $X$ is left to the econometrician and can include transformation of the original covariates so as to match more features of their distribution. The idea behind such weights relies on the idea of "covariate balancing" as in Imai and Ratkovic (2014); Fong *et al.* (2015) and references therein. The following lemma shows that under Assumption 3.1, weights satisfying the balancing condition always exist.

**Lemma 3.1**
*If Assumption 3.1 holds, the propensity score weight* $W_0 := \mathbb{P}(D=1|X)/[1-\mathbb{P}(D=1|X)]$ *satisfies the balancing condition* (2).

It is straightforward to verify by plugging this expression in equation (2) and using the law of iterated expectations. Note that the linear regression weight $W = \mathbb{E}[DX^T]\mathbb{E}[(1-D)XX^T]^{-1}X$ also verifies the balancing condition but can be negative. The lemma suggests running a regression to obtain $\mathbb{P}(D=1|X)$ and estimate weights $W_0$ as a first step, and plugging them to estimate $\theta_0$ in a second step. The usual solution used in the literature on propensity weighting is to estimate a binary choice model of $D$ on $X$. However, an inconsistent estimate of the propensity score leads to an inconsistent estimator of $\theta_0$. Moreover, estimation of a propensity score can be problematic when there is only one or very few treated units as in the synthetic control setting. Finally, running a regression of $D$ on $X$ does not guarantee that the implied weights will achieve covariate balancing. For these reasons, we consider instead an estimation directly based on balancing equations:

$$\mathbb{E}\left[(D-(1-D)W_0)\,X\right] = 0. \tag{3}$$

An important advantage of this approach over the usual one based on the propensity score estimation through maximum likelihood is its double robustness (for a definition, see, e.g., Bang and Robins, 2005). Intuitively, let $W_1$ denote the weights identified by (3) and a

misspecified model on the propensity score. Because the balancing equations (3) still hold for $W_1$, the estimated treatment effect will still be consistent provided that $\mathbb{E}[Y_0|X]$ is linear in $X$. The formal result is provided in Theorem 3.1 below.

We consider a parametric estimator of $W_0$. Suppose that $P(D = 1|X) = G(X^T\beta_0)$ for some unknown $\beta_0 \in \mathbb{R}^p$ and some known, strictly increasing cumulative distribution function $G$. Then $W_0 = h(X^T\beta_0)$ with $h = G/(1 - G)$ and $\beta_0$ is identified by (3). $h$ is a positive increasing function, meaning that its primitive $H$ is convex and its derivative (if it exists) is positive. A classical example of $h$ would be $h = \exp$, corresponding to a logistic distribution for $G$. In such an example, $h = h' = H$. In any case, the convexity of $H$ implies that $\beta_0$ solves the following strictly convex program:

$$\beta_0 = \arg\min_{\beta} \mathbb{E}\left[(1 - D)H(X^T\beta) - DX^T\beta\right]. \tag{4}$$

Note that this program is well-defined, whether or not $P(D = 1|X) = G(X^T\beta_0)$.

We are now ready to state the main identification theorem that justifies the use of the ATET estimand of equation (1):

**Theorem 3.1** Double Robustness
*Suppose that Assumptions 3.1-3.2 hold and let $\beta_0$ be defined by:*

$$\beta_0 = \arg\min_{\beta} \mathbb{E}\left[(1 - D)H(X^T\beta) - DX^T\beta\right],$$

*for some positive, strictly increasing convex function $H$. Then, for any $\mu \in \mathbb{R}^p$, $\theta_0$ solves the following moment condition:*

$$\mathbb{E}\left[\left(D - (1 - D)h(X^T\beta_0)\right)(Y - X^T\mu) - D\theta_0\right] = 0 \tag{5}$$

*in two cases:*

*(1) the outcome equation for the control is linear, i.e. there exist $\mu_0 \in \mathbb{R}^p$ such that $\mathbb{E}(Y_0|X) = X^T\mu_0$, or*

*(2) the propensity score is given by $P(D = 1|X) = G(X^T\beta_0)$, with $G = h/(1 + h)$.*

Theorem 3.1 highlights the doubly robustness property of using an estimate of the propensity score based on the balancing approach. This result is very similar to the one obtained by Kline (2011) for the Oaxaca-Blinder estimator, but his requires the propensity score to follow specifically log-logistic model in the propensity-score-well-specified case. In consequence, Theorem 3.1 is more general.

At this stage, $\mu$ does not play any role and could be set to 0. However, we will see below that choosing carefully $\mu$ is important when variable selection is performed in a first step to obtain an "immunized" estimator of $\theta_0$. This is particularly relevant in high-dimensional settings.

## 3.2 Properties in a Low-Dimensional Setting

This section considers the usual low-dimensional case where the dimension $p$ of the parameter $\beta_0$ is fixed, while the sample size tends to infinity. $\beta_0$ verifies the first-order condition $\mathbb{E}\left[((1-D)h(X^T\beta_0) - D)X\right] = 0$, which is a system of $p$ equations with $p$ unknowns. In a low-dimensional setting, an estimator of $\beta_0$ is obtained by taking the empirical counterpart of (4):

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}(1-D_i)H(X_i^T\beta) - D_i X_i^T\beta \tag{6}$$

Including an intercept among the $X$ is strongly advised as it ensures that estimated weights sum to one. This estimator is plugged in the empirical counterpart of (5) to estimate $\theta_0$:

$$\widetilde{\theta} = \frac{1}{\sum_{i=1}^{n} D_i}\sum_{i=1}^{n}[D_i - (1-D_i)h(X_i^T\widehat{\beta})]Y_i.$$

For the sake of brevity, let us denote the observed data by $Z_i := (Y_i, D_i, X_i)$. Also denote the estimating moment for $\theta_0$ by $g(Z,\theta,\beta,\mu) := [D-(1-D)h(X^T\beta)][Y - X^T\mu] - D\theta$. In the low-dimensional case, $\hat{\theta}$ is such that $\mathbb{E}_n g(Z_i, \widetilde{\theta}, \hat{\beta}, 0) = 0$. This estimator is nothing more than a two-step GMM and the next theorem gives its asymptotic distribution.

**Theorem 3.2** Asymptotic Normality of $\widetilde{\theta}$ in low-dimension
*Assume $(Z_i)_{i=1,\dots,n}$ are iid copies of $Z = (Y, D, X)$. Suppose that Assumptions 3.1-3.2 hold. Let $\theta_0$ and $\beta_0$ be defined as in Theorem 3.1 and further assume that they are interior points of $\mathbb{R}$ and $\mathbb{R}^p$ respectively. $g$ is constinuoulsy differentiable in $\theta$ and suppose $\mathbb{E}[g^2(Z,\theta,\beta,0)] < \infty$. Finally, assume that $\mathbb{E}[(1-D)h'(X^T\beta_0)XX^T]$ is non-singular. Then $\widetilde{\theta}$ verifies:*

$$\hat{\sigma}^{-2}\sqrt{n}(\widetilde{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

*where $\hat{\sigma}^2 := \mathbb{E}_n\left[g(Z_i, \widetilde{\theta}, \hat{\beta}, \hat{\mu})^2\right] / \mathbb{E}_n(D_i)^2$ is a consistent estimator of the asymptotic variance, with $\widetilde{\theta}$ and $\hat{\beta}$ defined as above and $\hat{\mu} := \mathbb{E}_n[(1-D_i)h'(X_i^T\hat{\beta})X_iX_i^T]^T\mathbb{E}_n[(1-D_i)h'(X_i^T\hat{\beta})X_iY_i]$.*

The proof can be found in Newey and McFadden (1994, Section 6). The optimal variance uses the quantity $\hat{\mu}$ which is the coefficient of the weighted regression of $Y$ on $X$ for the control group. The next section will use this observation to adapt the estimation in the high-dimensional case.

# 4 Extension to the High-Dimensional Case and Variable Selection

## 4.1 $\ell_1$-Regularized Estimation

This section considers a high-dimensional setting where the number of covariates increases with and could even be larger than the sample size. It emcompasses several situations:

(1) There are some applications where the researcher is faced with a large dataset in the sense that many covariates are to be considered with respect to the relatively small sample size. It is a natural setting that often occurs in macroeconomic problems. For example, in the Tobacco control program application by Abadie *et al.* (2010) the control group size is limited due to the fact that the observational unit is the state but many pre-treatment outcomes are included among the covariates. Section 2 revisits this example (*Natural high-dimension setting*).

(2) Sometimes the researcher also wants to consider a flexible form for the weights and instead of only using the raw covariates, she also wants to include transformations of them. This arises for example when categorical variables are interacted with other categorical variables or with continuous variables, or when a discrete variable such as the number of schooling years is broken down into binary variables to have a very flexible non-linear effect. This case can be labeled as "non-parametric" (*Artificial high-dimension setting*).

(3) More specifically in our estimation strategy, one may want not only to balance the first moments of the distribution of the covariates but also the second moments, the covariances, the third moments and so on to make the distribution more similar between the treated and the control group. In this case, a high-dimension setting appears to be desirable (*Many moments balancing setting*).

An inherent element of the high-dimensional literature is the notion of *sparsity*, *i.e.* the assumption that although we consider many variables, only a small number of elements in the vector of parameter is different from zero. This assumption amounts to recasting the problem in a variable selection framework where a good estimator should be able to correctly select the relevant variables or approximate the quantities of interest and be consistent at a rate close to $\sqrt{n}$, only paying a price proportional to the number of non-zero elements. A less restrictive concept has been introduced by Belloni *et al.* (2012). Called *approximate sparsity*, it assumes that the high-dimensional parameter can be decomposed into a sparse component, which has a lot of zero elements and some large elements, and a small component for which all

elements are small and decaying towards zero without never exactly being zero. It has been shown that in both contexts, Lasso-type estimators can provide a good approximation of the relevant quantities that are subject to a sparse structure, be it finite or infinite dimensional parameters. Consequently, consider the program 6, regularized by penalizing the $\ell_1$-norm of $\beta$:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (1 - D_i) H(X_i^T \beta) - D_i X_i^T \beta + \lambda_d \sum_{j=1}^{p} \psi_j^d |\beta_j|, \tag{7}$$

where $\lambda_d > 0$ is an overall penalty parameter set to dominate the noise that stems from the gradient of the function, and $\{\psi_j^d\}_{j=1}^d$ are covariate specific penalty loadings set as to insure good asymptotic properties. The penalty loadings are estimated using the algorithm presented in Appendix A. For empirical applications, we advise not to penalize the intercept in order to obtain final weights that sum to one by construction.

The form of this minimization program is one of the main contributions of this paper to the existing literature on high-dimensional models. On the one hand, this program targets covariate balancing as the main objective because equating the derivative of the loss function to zero yields a balancing condition as in equation (3). On the other hand, this objective function includes a term that penalizes the complexity of the model. Such an objective function borrows from the Lasso estimator of Tibshirani (1994), further studied and generalized most notably in Candes and Tao (2007); van de Geer (2008); Bickel *et al.* (2009). It has been specifically studied in the econometric literature by Belloni and Chernozhukov (2011); Belloni *et al.* (2012); Belloni and Chernozhukov (2013). This type of penalization offers multiple advantages: it regularizes the program so as to make it solvable contrary to a non-penalized GMM estimator, it yields strict sparsity in the sense that some elements of the estimated coefficients will be set exactly to zero if the penalty is large enough contrary to an $\ell_2$-penalization, it is computationally feasible because it gives rise to a convex program contrary to an $\ell_0$-penalization.[1] The use of covariate-specific penalty loadings borrows from the approach of Belloni *et al.* (2012) that adapts the Lasso to the non-Gaussian, non-homoscedastic case.

The drawback of penalizing by the $\ell_1$-norm is the bias that it induces in the estimation of the coefficients. To remove it, a popular solution given in the Lasso-related econometric literature is the use of a Post-Lasso estimator. Such an estimator would use a second step where variables corresponding to non-zero elements of $\hat{\beta}$ are kept in the model and the others are discarded. Then estimation is done a second time using only these variables and no penalization to compute the Post-Lasso solution, see Belloni and Chernozhukov (2011). Our strategy allows this estimator to be used, although we do not pursue this avenue here.

[1] Examples of such penalization are the BIC and AIC.

The estimator of $\beta_0$ above will be consistent as $n$ tends to infinity under classical assumptions used for the Lasso with quadratic loss. As suggested above, we could then consider the plug-in estimator for the ATT, based on Equation (5) with $\mu = 0$:

$$\widetilde{\theta} = \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} [D_i - (1 - D_i)h(X_i^T \widehat{\beta})]Y_i.$$

We refer to this estimator as the *naive plug-in estimator*. Because $p$ also grows with $n$, the Lasso estimator of the nuisance parameter $\beta_0$ will not be asymptotically normal. This translates into the fact that the naive plug-in estimator will be badly biased and may not be asymptotically normal, as illustrated for example in Belloni *et al.* (2014b); Chernozhukov *et al.* (2015b). The following section considers another estimator based on another choice of $\mu$, which is not affected by this problem.

Following Chernozhukov *et al.* (2015b), consider an immunized estimator that is first-order insensitive to $\widehat{\beta}$. This estimator will be asymptotically normal with a very simple asymptotic variance that does not depend on the properties of the first-step estimator. The idea is to choose a $\mu$ in (5) so that the derivative of this moment with respect to $\beta$ is zero when taken at $(\theta_0, \beta_0)$. This holds for $\mu = \mu_0$, where $\mu_0$ satisfies

$$\mathbb{E}\left[(1 - D)h'(X_i^T \beta_0)(Y - X^T \mu_0)X\right] = 0.$$

Notice that since $h$ is a strictly increasing function of its argument, $h'$ will be positive. Recognizing the first order condition of a least-squares program, $\mu_0$ can be obtained as the coefficient of a weighted regression of $Y$ onto $X$:

$$\mu_0 = \arg\min_{\mu} \mathbb{E}\left[(1 - D)h'(X^T \beta_0)(Y - X^T \mu)^2\right]. \tag{8}$$

Note that we have to estimate $\mu_0$. Then one may be worried that because $\mu_0$ is high-dimensional, it will affect the estimation of $\theta_0$ just as the high-dimensional estimator of $\beta_0$ affects the naive plug-in estimator of $\theta_0$. However, by construction the derivative of the moment condition (5) with respect to $(\beta, \mu)$ is equal to zero at the true values $(\beta_0, \mu_0)$.

Because of the large dimension of $X$, we consider once again a Lasso-type estimator for $\mu_0$:

$$\hat{\mu} = \arg\min_{\mu} \frac{1}{n} \sum_{i=1}^{n} (1 - D_i)h'\left(X_i^T \hat{\beta}\right)\left(Y_i - X_i^T \mu\right)^2 + \lambda_y \sum_{j=1}^{p} \psi_j^y |\mu_j|. \tag{9}$$

As previously, $\lambda_y > 0$ is an overall penalty parameter set to dominate the noise that stems from the gradient of the function, and $\left\{\psi_j^y\right\}_{j=1}^{p}$ are covariate-specific penalty loadings. Those tuning parameters will not be set in the same way as for the estimation of $\beta_0$. Finally, the

immunized ATT estimator is:

$$\widehat{\theta} = \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} \left[ D_i - (1 - D_i) h(X_i^T \hat{\beta}) \right] (Y_i - X_i^T \hat{\mu})$$

$$= \underbrace{\tilde{\theta}}_{\text{Naive Plug-In}} - \underbrace{\hat{\mu}_0^T \frac{1}{n_1} \sum_{i=1}^{n} \left[ D_i - (1 - D_i) h(X_i^T \hat{\beta}) \right] X_i}_{\text{Correction = Outcome-related } X \times \text{ Imbalance in } X} .$$

Intuitively, the immunized moment brings a correction to the naive plug-in estimate in the case where the balancing program has "missed" a covariate which appears to be very important to predict the outcome. This result has a flavor of Frish-Waugh-Lowell partialling-out procedure for model selection as put under the spotlights most notably by Belloni *et al.* (2014a) and further theorized in Chernozhukov *et al.* (2015b). Indeed, the estimating moment 5 for $\theta_0$ can be re-written so as to highlight the partialling out of $X$ from both $Y$ and $D$:

$$\mathbb{E}\left( \underbrace{[D - (1 - D) h(X^T \beta_0)]}_{\text{Residual Imbalance}} \underbrace{[Y - X^T \mu_0]}_{\text{Regression Residual}} \right) = \mathbb{E}\left( D \theta_0 \right)$$

Here, the effect of $X$ is taken out from $Y$ in a linear fashion, while the effect of $X$ on $D$ is taken out by re-weighting the control group so as to yield the same mean for $X$.

To summarize, the estimator in the high-dimensional case comprises the three following steps. Each step is simple to obtain as it involves at most to minimize a convex (and in general strictly convex) function:

(1) *(Calibration step)* For a given level of penalty $\lambda_d$ and positive covariate-specific penalty loadings $\{\psi_j^d\}_{j=1}^p$ solve the following:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (1 - D_i) H(X_i^T \beta) - D_i X_i^T \beta + \lambda_d \sum_{j=1}^{p} \psi_j^d |\beta_j|, \tag{10}$$

(2) *(Immunization step)* For a given level of penalty $\lambda_y$ and covariate-specific penalty loadings $\{\psi_j^y\}_{j=1}^p$ solve the following, using $\hat{\beta}$ estimated in the previous step:

$$\hat{\mu} = \arg\min_{\mu} \frac{1}{n} \sum_{i=1}^{n} (1 - D_i) h'(X_i^T \hat{\beta}) \left( Y_i - X_i^T \mu \right)^2 + \lambda_y \sum_{j=1}^{p} \psi_j^y |\mu_j|, \tag{11}$$

(3) *(ATT estimation)* Estimate the ATT using the immunized moment estimator:

$$\widehat{\theta} = \frac{1}{\sum_{i=1}^{n} D_i} \sum_{i=1}^{n} \left[ D_i - (1 - D_i) h(X_i^T \hat{\beta}) \right] (Y_i - X_i^T \hat{\mu}). \tag{12}$$

We will refer to this estimator as the *immunized* estimator.

## 4.2 Asymptotic Properties

This section deals with the asymptotic properties of the immunized ATT estimator. Our current framework poses several challenges to achieving inference that would be uniform on a large class of DGP. Firstly, $X$ is of high-dimension, since we allows $p > n$ and $p$ can grow with $n$ as long as $\log p = o(n^{1/3})$. Secondly, the ATT estimation is polluted by estimation of nuisance parameters $\beta_0$ and $\mu_0$ and we wish to neutralize the influence of this first step. Finally and closely related to high-dimensional statistical problems, the $\ell_1$-penalized estimators we use for $\beta_0$ and $\mu_0$ are not conventional. The estimator of $\beta_0$ relies on a convex but potentially non-Lipschitz loss function, contrary to the cases considered by van de Geer (2008). The estimation of $\mu_0$ is close to a standard Lasso except that it relies on weights depending on $\hat{\beta}$.

For the sake of brevity, we still denote the observed data by $Z_i := (Y_i, D_i, X_i)$ and let $\eta_0 := (\beta_0^T, \mu_0^T)^T$ denote the vector gathering the two nuisance parameters. Also denote the estimating moment for $\theta_0$ by $g(Z, \theta, \eta) := [D - (1-D)h(X^T\beta)](Y - X^T\mu) - D\theta$. We recall that the true values satisfy

$$\mathbb{E}g(Z, \theta_0, \eta_0) = 0. \tag{13}$$

The main theorem of the paper is inspired by Proposition 2 in Chernozhukov *et al.* (2015b).

**Assumption 4.1** High-level Conditions

*(1)* *Adaptivity:* $\sqrt{n}[\mathbb{E}_n g(Z_i, \theta_0, \hat{\eta}) - \mathbb{E}_n g(Z_i, \theta_0, \eta_0)] \xrightarrow{\mathbb{P}_n} 0.$

*(2)* *Normality:* $Var\left(g(Z, \theta_0, \eta_0)\right)^{-1/2} \sqrt{n}\mathbb{E}_n g(Z_i, \theta_0, \eta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$

*(3)* *Variance consistency:* $\mathbb{E}_n\left[g(Z_i, \hat{\theta}, \hat{\eta})^2\right]^{-1} \mathbb{E}\left[g(Z, \theta_0, \eta_0)^2\right] \xrightarrow{\mathbb{P}_n} 1.$

**Theorem 4.1** Asymptotic Normality of the Immunized Estimator
*Consider a sequence* $\{\boldsymbol{P_n}\}$ *of sets of probability measures such that for each sequence* $\{\mathbb{P}_n\} \in \{\boldsymbol{P_n}\}$ *the high-level conditions above hold. The immunized estimator* $\hat{\theta}$ *verifies:*

$$\hat{\sigma}^{-2}\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1),$$

*where* $\hat{\sigma}^2 := \mathbb{E}_n\left[g(Z_i, \hat{\theta}, \hat{\eta})^2\right] / \mathbb{E}_n(D_i)^2$ *is a consistent estimator of the asymptotic variance.*

Among the three high-level conditions above, the adaptivity is the less obvious one. Achieving adaptivity requires putting some structure on the problem and characterizing the class of DGPs we are dealing with. We consider for that purpose the following conditions.

**Assumption 4.2** Approximate Sparsity and Dimension Restrictions
*(i)* $\log(p) = o(n^{1/3})$

*(ii) The nuisance parameter $\eta_0$ can be decomposed between two components: a sparse component and a small component in the following sense:*

$$\eta_0 = \eta_0^a + \eta_0^b, \text{ with } support(\eta_0^a) \cap support(\eta_0^b) = \varnothing.$$
$$\|\eta_0^a\|_0 \leq s, \ \|\eta_0^a\|_1 := C_s < \infty, \|\eta_0^b\|_1 \leq c_1\sqrt{s^2/n}, \ \|\eta_0^b\|_2 \leq c_1\sqrt{s/n}.$$

*At most $s$ elements of the sparse components are non-zero and $\ell_1$ and $\ell_2$ norms are bounded and decreasing to zero for the small component.*

**Assumption 4.3** Penalty Loadings

$$\lambda_d := c\Phi^{-1}(1-\gamma/2p)/\sqrt{n}$$
$$\psi_j^d := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(1-D_i)h(X_i^T\beta_0) - D_i\right]^2 X_{i,j}^2}, \text{ for } j = 1, ..., p,$$

*where $c > 1$ is an absolute constant, $\gamma \lesssim \log(p \vee n)$ and $\beta_0$ is the true coefficient. Moreover,*

$$\bar{\psi} := \max_{1\leq j\leq p}\psi_j < \infty, \ \underline{\psi} := \min_{1\leq j\leq p}\psi_j < \infty. \tag{14}$$

$c$ and $\gamma$ are arbitrary choices and Belloni *et al.* (2012) set $\gamma := .1/\log(p \vee n)$ and $c := 1.1$. We let $c_\psi := \frac{\bar{\psi}}{\underline{\psi}}$.

**Assumption 4.4** Bounded Covariates

*For every $i$:*

$$\max_{1\leq j\leq p}|X_{ij}| := K_n < \infty \tag{15}$$

Define $\Sigma := \mathbb{E}((1-D_1)X_1X_1^T)$, the theoretical Gram matrix on the control group. For a non-empty subset $S \subset \{1, ..., p\}$ and $\alpha > 0$, we also define the set:

$$\mathcal{C}[S, \alpha] := \{v \in \mathbb{R}^p : \|v_{S^C}\|_1 \leq \alpha\|v_S\|_1, v \neq 0\} \tag{16}$$

Finally, let $c_0 = (c+1)/(c-1)$.

**Assumption 4.5** Restricted Eigenvalue on Gram matrix

*For $\alpha \in \{c_0c_\psi, 2c_0c_\psi\}$,*

$$\kappa_\alpha^2(\Sigma) := \min_{J\subset\{1,...,p\}, |J|_0\leq s}\min_{\delta\in\mathcal{C}[J,\alpha]}\frac{\delta^T\Sigma\delta}{\|\delta_J\|_2^2} > 0.$$

*Moreover, there exists $h \in ]1, +\infty[$ such that for all $v \in \mathbb{R}^p$,*

$$\sqrt{\mathbb{E}[(v^T(1-D_1)X_1X_1^Tv)^2]} \leq hv^T\Sigma v.$$

The following theorem shows that the adaptivity condition is satisfied in this context.

**Theorem 4.2**

*Suppose that Assumptions 3.1, B.1-B.4 hold. Then the adaptivity condition (1) in Assumption 4.1 holds.*

# 5 Monte Carlo Simulations

This section reports results from a Monte Carlo experiment. The aim is two-fold: illustrate the better properties of the BEAST (immunized) estimator over the naive plug-in, and compare it with other competitors. In particular, we compare it with an inverse propensity score weighting estimator where the propensity score is estimated using a Logit-Lasso (van de Geer, 2008).

## 5.1 Data-Generating Process

This subsection describes the data-generating processes (DGP) used for the simulation exercises. In our main specification (DGP1), the outcome equation is linear and given by: $y_i = d_i\alpha + x_i^T\mu + \varepsilon_i$, where $\alpha = 0$, $\varepsilon_i \perp\!\!\!\perp x_i$, and $\varepsilon_i \sim \mathcal{N}(0,1)$. The treatment equation follows a Probit model, $d_i \sim \text{Probit}\left(x_i^T\gamma\right)$. The covariates are simulated as $x_i \sim \mathcal{N}(0,\Sigma)$, where each entry of the variance-covariance matrix is set as follows: $\Sigma_{j,k} = .5^{|j-k|}$. The most interesting part of the DGP is the form of the coefficients $\gamma$ and $\mu$:

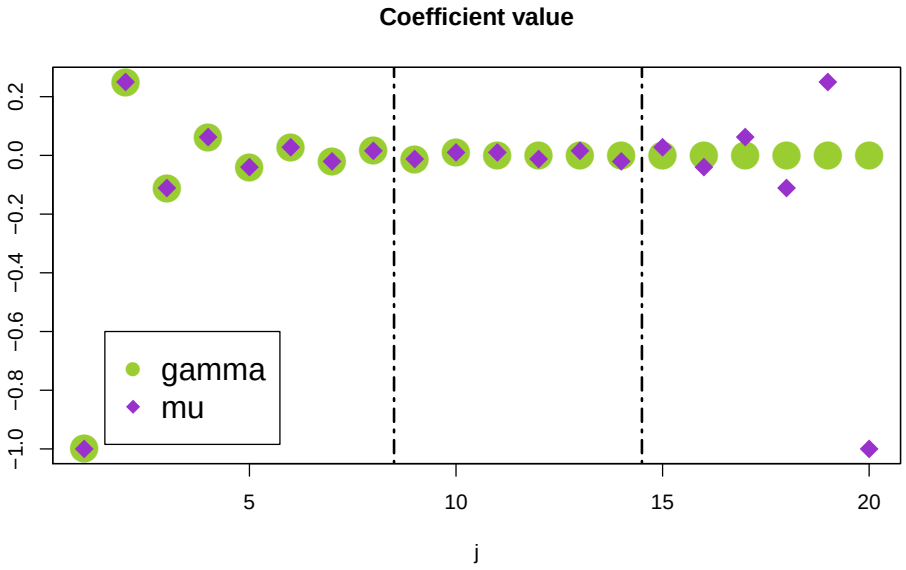
$$\gamma_j = \begin{cases} \rho_d(-1)^j/j^2,\ j < p/2 \\ 0,\ \text{elsewhere} \end{cases},\ \mu_j = \begin{cases} \rho_y(-1)^j/j^2,\ j < p/2 \\ \rho_y(-1)^{j+1}/(p-j+1)^2,\ \text{elsewhere} \end{cases}$$

We are in an approximately sparse setting for both equations. $\rho_y$ and $\rho_d$ are constants that are set to yield a certain value for the $R^2$ in both equations. It is a way to fix the signal-to-noise ratio, in the sense that a larger constant $\rho_y$ will mean that the covariates play a larger role. The trick here, is that some variables that matter a lot in the outcome equation are irrelevant in the treatment assignment rule. The fact that the balancing program will miss some relevant variables for the outcome should create a bias, a non-Normal behavior or at least a wider variance. Figure 1 depicts the sparsity pattern of both coefficients for $p = 20$.

In other words, we expect a naive plug-in estimator to miss the variables located at the far-right of the plot, thereby creating a bias in the treatment effect estimate. The immunize procedure is expected to correct this bias.

We explore three alternatives to DGP1. DGP2 is similar in all respects, except that $\mu_j$ is equal to 0 for all $j$. In other words, the outcome equation does not depend on the observed characteristics $X$. In this specification, the naive plug-in estimator is theoretically correct, and we can check whether the BEAST performs as well as the naive version. DGP3 explores a situation with a heterogeneous treatment effect: the outcome equation is specified as $y_i = d_i\alpha + x_i^T\mu + d_i x_i^T\gamma + \varepsilon_i$, with $\gamma_j = 10$ for all $j$. Since $x_i \sim \mathcal{N}(0,\Sigma)$, this setting still yields an ATT equal to zero. DGP4 relaxes the linearity of the outcome equation and allows to check for double robustness. In this DGP, we specify the outcome equation as follows: $y_i = d_i\alpha + \left(x_i^T\mu\right)^2 + \varepsilon_i$, but only the covariates $x_i$ are used in the estimation procedure.

**Figure 1:** Sparsity patterns of both coefficients




Note: In this example, $\rho_d = \rho_y = 1$. The central region of the graph represents the coefficients $\gamma$ and $\mu$ associated with variables that do not play an important role in either the equation equation or the outcome equation. The left region shows the coefficients associated with variables that are important for both equations. In the left region, only $\mu$ is different from 0, meaning that the variables determine the outcome equation but not the selection equation.

## 5.2 Results

Tables 1 to 4 show the results of Monte-Carlo simulations for both the naive plug-in estimator and the immunized estimator, compared with 3 potential alternatives: inverse propensity score weighting and Farrell's estimate (with either Lasso or post-lasso procedures). We present results of 1,000 replications with each DGP introduced in the previous paragraph, for values of $n$ and $p$ varying between 50 and 500. All other parameters are held fixed.

The first striking characteristic of these results is that in our baseline DGP (table 1), the bias of the plug-in estimator is almost always larger in absolute value than the bias of the immunized estimator, as predicted. In addition, the difference in bias grows with the sample size. For example, for $n = 500$, the bias of the plug-in estimator is about twice as big as the bias of the immunized BEAST estimator. Similarly, the root mean squared error (RMSE) is always higher for the naive estimator than for the immunized one. The difference again increases with sample size. The p-value of the Shapiro test is also usually quite high, showing that the null hypothesis of normality cannot be rejected. These results illustrate the theoretical asymptotic properties of our estimator. Moreover, table 2 also shows that in a setting in which the naive plug-in would be appropriate, the BEAST performs exactly as well as the naive estimator. Table 3 shows that these findings are robust to

heterogeneous treatment effects, as both estimators perform exactly the same way with or without heterogeneity.

Figure 2 further illustrates these properties by displaying the empirical distribution of both the naive plug-in estimator and the immunized estimator for $n = 500$ and $p = 500$. For this example, one can clearly see that the empirical variance of the estimate is much higher in the first case. Moreover, the immunized estimator much better fits the normal density curve, illustrating the asymptotic normality of the immunized estimator.

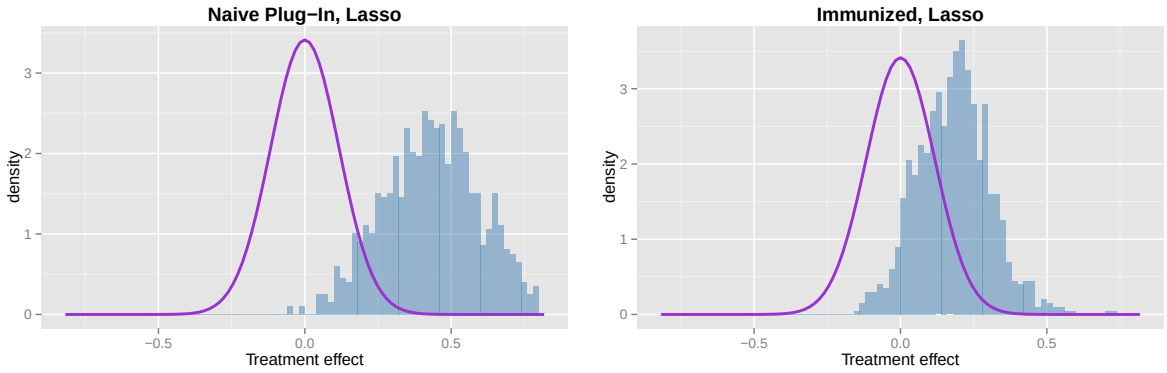



**Figure 2:** Empirical distribution

Note: The solid purple curve is a Normal distribution with mean $\theta_0 = 0$ and standard deviation the asymptotic standard error of the immunized estimator. The data generating process is such that the $R^2$ of the outcome equation is .8 and the $R^2$ of the treatment equation is .2. The estimates are computed with parameters $c_y = .2$ and $c_d = .7$ on 1000 replications, for $n = 100$ and $p = 200$.

**Table 1:** Monte-Carlo simulations (DGP1)

| | $p = 50$ | | | $p = 100$ | | | $p = 200$ | | | $p = 500$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RMSE | Bias | Shapiro | RMSE | Bias | Shapiro | RMSE | Bias | Shapiro | RMSE | Bias | Shapiro |
| $n = 50$ | | | | | | | | | | | | |
| Naive Plug-in | 0.993 | 0.829 | 0.326 | 1.060 | 0.891 | 0.831 | – | – | – | – | – | – |
| BEAST | 0.689 | 0.474 | 0.010 | 0.793 | 0.589 | 0.016 | – | – | – | – | – | – |
| Inv. prop. weighting | 0.925 | 0.771 | 0.491 | 0.99 | 0.828 | 0.28 | – | – | – | – | – | – |
| BCH | 0.873 | 0.567 | 0.074 | 1.048 | 0.811 | 0.715 | – | – | – | – | – | – |
| Farrell | 0.710 | 0.509 | 0.006 | 0.816 | 0.616 | 0.008 | – | – | – | – | – | – |
| Farrell PL | 0.657 | 0.210 | 0.000 | 0.794 | 0.301 | 0.000 | – | – | – | – | – | – |
| $n = 100$ | | | | | | | | | | | | |
| Naive Plug-in | 0.816 | 0.719 | 0.649 | 0.824 | 0.739 | 0.462 | 0.843 | 0.758 | 0.710 | – | – | – |
| BEAST | 0.396 | 0.264 | 0.002 | 0.395 | 0.270 | 0.030 | 0.424 | 0.300 | 0.002 | – | – | – |
| Inv. prop. weighting | 0.749 | 0.653 | 0.900 | 0.761 | 0.674 | 0.488 | 0.780 | 0.697 | 0.591 | – | – | – |
| BCH | 0.248 | 0.057 | 0.318 | 0.253 | 0.042 | 0.087 | 0.299 | 0.076 | 0.000 | – | – | – |
| Farrell | 0.396 | 0.266 | 0.008 | 0.406 | 0.282 | 0.011 | 0.440 | 0.318 | 0.000 | – | – | – |
| Farrell PL | 0.323 | 0.102 | 0.000 | 0.339 | 0.104 | 0.000 | 0.414 | 0.144 | 0.000 | – | – | – |
| $n = 200$ | | | | | | | | | | | | |
| Naive Plug-in | 0.616 | 0.556 | 0.659 | 0.637 | 0.581 | 0.844 | 0.645 | 0.591 | 0.853 | 0.678 | 0.628 | 0.383 |
| BEAST | 0.249 | 0.160 | 0.013 | 0.261 | 0.174 | 0.360 | 0.264 | 0.179 | 0.005 | 0.277 | 0.204 | 0.958 |
| Inv. prop. weighting | 0.571 | 0.507 | 0.984 | 0.587 | 0.529 | 0.98 | 0.596 | 0.539 | 0.849 | 0.627 | 0.576 | 0.575 |
| BCH | 0.176 | 0.055 | 0.010 | 0.177 | 0.056 | 0.725 | 0.176 | 0.056 | 0.536 | 0.178 | 0.061 | 0.311 |
| Farrell | 0.241 | 0.154 | 0.161 | 0.253 | 0.168 | 0.743 | 0.262 | 0.179 | 0.022 | 0.277 | 0.206 | 0.798 |
| Farrell PL | 0.206 | 0.062 | 0.000 | 0.233 | 0.076 | 0.198 | 0.251 | 0.103 | 0.019 | 0.297 | 0.140 | 0.000 |
| $n = 500$ | | | | | | | | | | | | |
| Naive Plug-in | 0.434 | 0.399 | 0.712 | 0.431 | 0.398 | 0.560 | 0.452 | 0.421 | 0.430 | 0.469 | 0.437 | 0.576 |
| BEAST | 0.154 | 0.099 | 0.825 | 0.151 | 0.102 | 0.975 | 0.159 | 0.110 | 0.330 | 0.162 | 0.116 | 0.819 |
| Inv. prop. weighting | 0.397 | 0.359 | 0.803 | 0.397 | 0.360 | 0.765 | 0.418 | 0.383 | 0.802 | 0.436 | 0.402 | 0.275 |
| BCH | 0.120 | 0.051 | 0.514 | 0.115 | 0.056 | 0.752 | 0.118 | 0.054 | 0.617 | 0.115 | 0.055 | 0.135 |
| Farrell | 0.144 | 0.088 | 0.791 | 0.143 | 0.092 | 0.980 | 0.150 | 0.100 | 0.372 | 0.155 | 0.108 | 0.910 |
| Farrell PL | 0.121 | 0.026 | 0.803 | 0.121 | 0.034 | 0.403 | 0.130 | 0.042 | 0.137 | 0.142 | 0.059 | 0.478 |

Note: Results based on 1,000 replications. The data generating process is such that the $R^2$ of the outcome equation is .8 and the $R^2$ of the treatment equation is .2. The naive plug-in and BEAST estimates are computed with parameters $c_y = .2$ and $c_d = .7$. The inverse propensity score weighting estimates are computed by estimating the propensity score using a Logit-Lasso algorithm. The Shapiro columns display the p-value of the Shapiro test (the null hypothesis is normality of the distribution).

**Table 2:** Monte-Carlo simulations (DGP2: outcome independent from $X$)

| | $p = 50$ | | | $p = 100$ | | | $p = 200$ | | | $p = 500$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RMSE | Bias | Shapiro | RMSE | Bias | Shapiro | RMSE | Bias | Shapiro | RMSE | Bias | Shapiro |
| $n = 50$ | | | | | | | | | | | | |
| Naive Plug-in | 0.287 | 0.004 | 0.484 | 0.293 | 0.001 | 0.102 | – | – | – | – | – | – |
| BEAST | 0.295 | 0.002 | 0.572 | 0.300 | 0.002 | 0.187 | – | – | – | – | – | – |
| Inv. prop. weighting | 0.273 | 0.005 | 0.424 | 0.279 | 0.001 | 0.037 | – | – | – | – | – | – |
| BCH | 0.293 | 0.000 | 0.347 | 0.293 | 0.002 | 0.115 | – | – | – | – | – | – |
| Farrell | 0.293 | 0.001 | 0.562 | 0.298 | 0.002 | 0.191 | – | – | – | – | – | – |
| Farrell PL | 0.446 | -0.012 | 0.000 | 0.475 | -0.009 | 0.000 | – | – | – | – | – | – |
| $n = 100$ | | | | | | | | | | | | |
| Naive Plug-in | 0.204 | -0.001 | 0.364 | 0.208 | -0.009 | 0.427 | 0.206 | -0.002 | 0.476 | – | – | – |
| BEAST | 0.206 | 0.000 | 0.339 | 0.211 | -0.009 | 0.208 | 0.209 | -0.001 | 0.306 | – | – | – |
| Inv. prop. weighting | 0.195 | -0.001 | 0.208 | 0.199 | -0.009 | 0.481 | 0.197 | -0.003 | 0.531 | – | – | – |
| BCH | 0.215 | 0.000 | 0.189 | 0.215 | -0.010 | 0.586 | 0.211 | -0.002 | 0.826 | – | – | – |
| Farrell | 0.204 | -0.001 | 0.436 | 0.210 | -0.010 | 0.270 | 0.209 | -0.002 | 0.383 | – | – | – |
| Farrell PL | 0.266 | -0.003 | 0.000 | 0.290 | -0.011 | 0.000 | 0.386 | -0.022 | 0.000 | – | – | – |
| $n = 200$ | | | | | | | | | | | | |
| Naive Plug-in | 0.146 | 0.000 | 0.071 | 0.150 | 0.000 | 0.552 | 0.148 | 0.005 | 0.712 | 0.147 | 0.007 | 0.945 |
| BEAST | 0.147 | -0.001 | 0.129 | 0.152 | 0.000 | 0.442 | 0.149 | 0.005 | 0.679 | 0.148 | 0.008 | 0.955 |
| Inv. prop. weighting | 0.141 | 0.000 | 0.217 | 0.146 | 0.000 | 0.57 | 0.141 | 0.005 | 0.684 | 0.140 | 0.007 | 0.726 |
| BCH | 0.152 | 0.000 | 0.077 | 0.151 | 0.000 | 0.901 | 0.151 | 0.004 | 0.346 | 0.153 | 0.006 | 0.345 |
| Farrell | 0.147 | 0.000 | 0.211 | 0.152 | 0.000 | 0.444 | 0.148 | 0.005 | 0.673 | 0.148 | 0.008 | 0.794 |
| Farrell PL | 0.180 | 0.002 | 0.000 | 0.196 | 0.001 | 0.172 | 0.203 | 0.005 | 0.320 | 0.224 | 0.002 | 0.000 |
| $n = 500$ | | | | | | | | | | | | |
| Naive Plug-in | 0.098 | 0.001 | 0.598 | 0.092 | -0.002 | 0.468 | 0.095 | -0.002 | 0.624 | 0.093 | -0.001 | 0.976 |
| BEAST | 0.098 | 0.001 | 0.572 | 0.092 | -0.001 | 0.319 | 0.095 | -0.001 | 0.618 | 0.093 | -0.001 | 0.979 |
| Inv. prop. weighting | 0.095 | 0.001 | 0.552 | 0.089 | -0.002 | 0.499 | 0.092 | -0.001 | 0.784 | 0.090 | -0.001 | 0.960 |
| BCH | 0.099 | 0.001 | 0.732 | 0.092 | 0.001 | 0.543 | 0.095 | -0.001 | 0.157 | 0.094 | -0.001 | 0.280 |
| Farrell | 0.098 | 0.001 | 0.552 | 0.092 | -0.002 | 0.380 | 0.096 | -0.001 | 0.630 | 0.093 | -0.001 | 0.981 |
| Farrell PL | 0.110 | 0.000 | 0.822 | 0.106 | -0.002 | 0.519 | 0.116 | -0.001 | 0.555 | 0.119 | -0.001 | 0.546 |

Note: Results based on 1,000 replications. The data generating process is such that the $R^2$ of the outcome equation is .8 and the $R^2$ of the treatment equation is .2. The naive plug-in and BEAST estimates are computed with parameters $c_y = .2$ and $c_d = .7$. The inverse propensity score weighting estimates are computed by estimating the propensity score using a Logit-Lasso algorithm. The Shapiro columns display the p-value of the Shapiro test (the null hypothesis is normality of the distribution).

**Table 3:** Monte-Carlo simulations (DGP3: heterogeneous treatment effect)

| | $p = 50$ | | | $p = 100$ | | | $p = 200$ | | | $p = 500$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RMSE | Bias | Shapiro | RMSE | Bias | Shapiro | RMSE | Bias | Shapiro | RMSE | Bias | Shapiro |
| $n = 50$ | | | | | | | | | | | | |
| Naive Plug-in | 0.993 | 0.829 | 0.326 | 1.060 | 0.891 | 0.831 | – | – | – | – | – | – |
| BEAST | 0.689 | 0.474 | 0.010 | 0.793 | 0.589 | 0.016 | – | – | – | – | – | – |
| Inv. prop. weighting | 0.925 | 0.771 | 0.491 | 0.990 | 0.828 | 0.280 | – | – | – | – | – | – |
| BCH | 7.772 | 2.303 | 0.000 | 7.709 | 1.449 | 0.000 | – | – | – | – | – | – |
| Farrell | 0.710 | 0.509 | 0.006 | 0.816 | 0.616 | 0.008 | – | – | – | – | – | – |
| Farrell PL | 0.657 | 0.210 | 0.000 | 0.794 | 0.301 | 0.000 | – | – | – | – | – | – |
| $n = 100$ | | | | | | | | | | | | |
| Naive Plug-in | 0.816 | 0.719 | 0.649 | 0.824 | 0.739 | 0.462 | 0.843 | 0.758 | 0.710 | – | – | – |
| BEAST | 0.396 | 0.264 | 0.002 | 0.395 | 0.270 | 0.030 | 0.424 | 0.300 | 0.002 | – | – | – |
| Inv. prop. weighting | 0.749 | 0.653 | 0.900 | 0.761 | 0.674 | 0.488 | 0.780 | 0.697 | 0.591 | – | – | – |
| BCH | 9.512 | 5.468 | 0.000 | 10.32 | 4.331 | 0.000 | 11.979 | 3.66 | 0.000 | – | – | – |
| Farrell | 0.396 | 0.266 | 0.008 | 0.406 | 0.282 | 0.011 | 0.440 | 0.318 | 0.000 | – | – | – |
| Farrell PL | 0.323 | 0.102 | 0.000 | 0.339 | 0.104 | 0.000 | 0.414 | 0.144 | 0.000 | – | – | – |
| $n = 200$ | | | | | | | | | | | | |
| Naive Plug-in | 0.616 | 0.556 | 0.659 | 0.637 | 0.581 | 0.844 | 0.645 | 0.591 | 0.853 | 0.678 | 0.628 | 0.383 |
| BEAST | 0.249 | 0.160 | 0.013 | 0.261 | 0.174 | 0.360 | 0.264 | 0.179 | 0.005 | 0.277 | 0.204 | 0.958 |
| Inv. prop. weighting | 0.571 | 0.507 | 0.984 | 0.587 | 0.529 | 0.98 | 0.596 | 0.539 | 0.849 | 0.627 | 0.576 | 0.575 |
| BCH | 9.275 | 6.688 | 0.000 | 10.049 | 6.676 | 0.000 | 12.283 | 7.141 | 0.000 | 16.616 | 7.341 | 0.000 |
| Farrell | 0.241 | 0.154 | 0.161 | 0.253 | 0.168 | 0.743 | 0.262 | 0.179 | 0.022 | 0.277 | 0.206 | 0.798 |
| Farrell PL | 0.206 | 0.062 | 0.000 | 0.233 | 0.076 | 0.198 | 0.251 | 0.103 | 0.019 | 0.297 | 0.140 | 0.000 |
| $n = 500$ | | | | | | | | | | | | |
| Naive Plug-in | 0.434 | 0.399 | 0.712 | 0.431 | 0.398 | 0.560 | 0.452 | 0.421 | 0.430 | 0.469 | 0.437 | 0.576 |
| BEAST | 0.154 | 0.099 | 0.825 | 0.151 | 0.102 | 0.975 | 0.159 | 0.110 | 0.330 | 0.162 | 0.116 | 0.819 |
| Inv. prop. weighting | 0.397 | 0.359 | 0.803 | 0.397 | 0.360 | 0.765 | 0.418 | 0.383 | 0.802 | 0.436 | 0.402 | 0.275 |
| BCH | 8.163 | 5.970 | 0.171 | 8.729 | 6.148 | 0.647 | 9.664 | 7.291 | 0.379 | 11.467 | 7.182 | 0.045 |
| Farrell | 0.144 | 0.088 | 0.791 | 0.143 | 0.092 | 0.980 | 0.150 | 0.100 | 0.372 | 0.155 | 0.108 | 0.910 |
| Farrell PL | 0.121 | 0.026 | 0.803 | 0.121 | 0.034 | 0.403 | 0.13 | 0.042 | 0.137 | 0.142 | 0.059 | 0.478 |

Note: Results based on 1,000 replications. The data generating process is such that the $R^2$ of the outcome equation is .8 and the $R^2$ of the treatment equation is .2. The naive plug-in and BEAST estimates are computed with parameters $c_y = .2$ and $c_d = .7$. The inverse propensity score weighting estimates are computed by estimating the propensity score using a Logit-Lasso algorithm. The Shapiro columns display the p-value of the Shapiro test (the null hypothesis is normality of the distribution).

**Table 4:** Monte-Carlo simulations (DGP4: non-linear outcome equation)

| | $p = 50$ | | | $p = 100$ | | | $p = 200$ | | | $p = 500$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RMSE | Bias | Shapiro | RMSE | Bias | Shapiro | RMSE | Bias | Shapiro | RMSE | Bias | Shapiro |
| $n = 50$ | | | | | | | | | | | | |
| Naive Plug-in | 1.692 | 0.119 | 0.749 | 1.676 | 0.200 | 0.088 | – | – | – | – | – | – |
| BEAST | 1.720 | 0.052 | 0.375 | 1.710 | 0.094 | 0.026 | – | – | – | – | – | – |
| Inv. prop. weighting | 1.734 | 0.487 | 0.864 | 1.739 | 0.590 | 0.057 | – | – | – | – | – | – |
| BCH | 1.669 | -0.052 | 0.853 | 1.657 | 0.073 | 0.030 | – | – | – | – | – | – |
| Farrell | 1.703 | 0.021 | 0.345 | 1.689 | 0.084 | 0.073 | – | – | – | – | – | – |
| Farrell PL | 2.812 | -0.164 | 0.000 | 3.346 | -0.112 | 0.000 | – | – | – | – | – | – |
| $n = 100$ | | | | | | | | | | | | |
| Naive Plug-in | 1.148 | 0.218 | 0.947 | 1.152 | 0.263 | 0.004 | 1.195 | 0.224 | 0.420 | – | – | – |
| BEAST | 1.158 | 0.209 | 0.843 | 1.168 | 0.245 | 0.011 | 1.207 | 0.172 | 0.200 | – | – | – |
| Inv. prop. weighting | 1.256 | 0.548 | 0.499 | 1.271 | 0.607 | 0.012 | 1.315 | 0.602 | 0.179 | – | – | – |
| BCH | 1.089 | -0.037 | 0.338 | 1.089 | 0.032 | 0.009 | 1.130 | 0.032 | 0.052 | – | – | – |
| Farrell | 1.163 | 0.176 | 0.946 | 1.142 | 0.205 | 0.027 | 1.199 | 0.130 | 0.084 | – | – | – |
| Farrell PL | 1.911 | 0.045 | 0.000 | 1.852 | 0.037 | 0.000 | 2.482 | -0.165 | 0.000 | – | – | – |
| $n = 200$ | | | | | | | | | | | | |
| Naive Plug-in | 0.935 | 0.328 | 0.694 | 0.847 | 0.280 | 0.091 | 0.863 | 0.289 | 0.399 | 0.889 | 0.249 | 0.008 |
| BEAST | 0.952 | 0.332 | 0.318 | 0.862 | 0.280 | 0.376 | 0.869 | 0.279 | 0.380 | 0.895 | 0.210 | 0.025 |
| Inv. prop. weighting | 1.075 | 0.597 | 0.652 | 0.988 | 0.576 | 0.047 | 1.025 | 0.614 | 0.279 | 1.058 | 0.629 | 0.009 |
| BCH | 0.799 | 0.008 | 0.723 | 0.772 | -0.005 | 0.162 | 0.760 | -0.014 | 0.959 | 0.810 | -0.020 | 0.224 |
| Farrell | 0.971 | 0.322 | 0.919 | 0.876 | 0.265 | 0.188 | 0.870 | 0.252 | 0.132 | 0.888 | 0.180 | 0.045 |
| Farrell PL | 1.438 | 0.171 | 0.000 | 1.403 | 0.097 | 0.000 | 1.485 | 0.06 | 0.000 | 1.36 | 0.031 | 0.000 |
| $n = 500$ | | | | | | | | | | | | |
| Naive Plug-in | 0.634 | 0.286 | 0.191 | 0.625 | 0.282 | 0.007 | 0.618 | 0.276 | 0.204 | 0.616 | 0.296 | 0.299 |
| BEAST | 0.637 | 0.288 | 0.188 | 0.626 | 0.280 | 0.007 | 0.621 | 0.274 | 0.211 | 0.613 | 0.288 | 0.338 |
| Inv. prop. weighting | 0.768 | 0.487 | 0.001 | 0.764 | 0.502 | 0.000 | 0.770 | 0.514 | 0.067 | 0.791 | 0.561 | 0.070 |
| BCH | 0.504 | -0.002 | 0.108 | 0.498 | 0.013 | 0.801 | 0.491 | -0.019 | 0.129 | 0.479 | 0.006 | 0.324 |
| Farrell | 0.722 | 0.337 | 0.043 | 0.689 | 0.315 | 0.005 | 0.682 | 0.300 | 0.092 | 0.658 | 0.302 | 0.145 |
| Farrell PL | 0.882 | 0.165 | 0.000 | 0.829 | 0.136 | 0.000 | 0.890 | 0.133 | 0.000 | 0.867 | 0.119 | 0.000 |

Note: Results based on 1,000 replications. The data generating process is such that the $R^2$ of the outcome equation is .8 and the $R^2$ of the treatment equation is .2. The naive plug-in and BEAST estimates are computed with parameters $c_y = .2$ and $c_d = .7$. The inverse propensity score weighting estimates are computed by estimating the propensity score using a Logit-Lasso algorithm. The Shapiro columns display the p-value of the Shapiro test (the null hypothesis is normality of the distribution).

How does the BEAST estimator fare compared with alternative methods? Tables 1 to 4 also show the performance of four alternative methods. The first one is inverse propensity-score weighting, estimating the propensity score with a Logit-Lasso (van de Geer, 2008). Overall, this method performs similarly to the naive plug-in estimator. This is not surprising, as it is likely to suffer from the same bias as the latter. The second alternative is the method introduced by Belloni *et al.* (2014a) (denoted as "BCH" in our tables). The main difference between this estimator and the BEAST is that the former relies on the assumption that the treatment effect is homogeneous. We thus expect the third DGP to yield better results for the BEAST than for BCH. Indeed, table 3 shows that BCH has both a very high bias and RMSE in this case. As an order of magnitude, they tend to be 10 times as high as for the BEAST estimator, whatever the size of the sample or the covariate set. In addition, the p-value of the Shapiro test tends to be smaller than for the BEAST, indicating that it is often less likely to be asymptotically normal.

The third and fourth alternatives are the Farrell estimator in its two versions: one based on a Lasso procedure (denoted as "Farrell" in our tables), and one based on a post-Lasso procedure ("Farrell PL"). The theory predicts that the latter has a smaller bias, because it directly tackles the bias introduced by variable selection. With DGP1 to DGP3, we find that the BEAST performs similarly to Farrell's method, and that the post-lasso procedure indeed reduces the bias compared to our estimator. However, it has two drawbacks. First, the Shapiro test tends to be rejected more often for the post-Lasso procedure than for the BEAST estimator. Second, table 4 shows that Farrell's methods are not as robust to non-linear outcome equations as the BEAST. Indeed, with DGP4 we find that the RMSE is systematically smaller for the BEAST than for Farrell's estimators. In particular, the post-Lasso procedure yields much higher RMSE for small samples (twice as high for $n = 50$). This illustrates the advantages of the BEAST's double robustness.

# 6 Applications

## 6.1 Job Training Program

Our first application revisits the famous application by LaLonde (1986). This dataset was first built to assess the impact of the National Supported Work (NSW) program. The NSW is a transitional, subsidized work experience program targeted towards people with longstanding employment problems: ex-offenders, former drug addicts, women who were long-term recipients of welfare benefits and school dropouts. Here, the quantity of interest is the ATET, defined as the impact of the participation into the program on 1978 yearly earnings in dollars. The treated group gathers people who were randomly assigned to this program from the population at risk ($n_1 = 185$).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 (sample size $n_0 = 260$). The second one comes from observational data: it is a sample from the Panel Study of Income Dynamics (PSID) (sample size $n_0 = 2490$). The presence of the experimental sample allows to obtain an experimental ATET which gives a benchmark for ATET obtained with observational data. We use these datasets only to compare the BEAST estimator with other competitors and defer discussion of the NSW program and the controversy regarding econometric estimates of nonexperimental causal effects to the paper by LaLonde (1986) and subsequent contributions by Dehejia and Wahba (1999, 2002); Smith and Todd (2005).

To allow for a flexible specification, 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. Continuous variables are linearly rescaled to $[0, 1]$. All in all, we end up with 172 variables to select from. This is the setting of Farrell (2015). The experimental benchmark for the ATT estimate is given by \$1,794 (671). We compare several estimators: the naive plug-in estimator, the immunized plug-in estimator (BEAST estimator), the doubly-robust estimator of Farrell (2015), the double-post-selection linear estimator of Belloni *et al.* (2014b), and a simple OLS estimator where all the covariates are included. Because the size of the control group is quite large, the synthetic control method could not be used.

Table 5 displays the results. Columns (3)-(5) show estimators that give a credible value for the ATT with respect to the experimental benchmark. However, they differ in their variances as one can easily see. Farrell (2015) in its Lasso version and the BEAST estimator achieve the lowest standard-error. Notably, Farrell (2015) in its Lasso version and the BEAST estimator are the only ones out of six estimators which display a significant, positive impact similarly

to the experimental benchmark. The BEAST estimator offers a large improvement on bias and standard error over the naive plug-in estimator, which augments the evidence given by the Monte Carlo experiment. The estimate obtained using Farrell (2015) shown in the table differ from the on displayed in the original paper because we have not automatically included the variables *education, 1974 income* and *nodegree* in the set of theory pre-selected covariates as it is done in the original paper. When doing so, the results are slightly better but not qualitatively different for this estimator, but we thought it would bias the comparison as other estimators do not include a set of pre-selected variables. For estimators from columns (2) to (6), the penalty parameters can potentially be tuned to obtain a better bias-variance trade-off. The OLS estimator in column (7) presents a benchmark of a very simple model that does not use any selection at all.

**Table 5:** Average Treatment Effect on the Treated (ATT) for several estimators

| | *Estimator:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | *Experimental benchmark* | *Plug-In Naive* | *BEAST Immunized* | *Farrell (2015) Lasso* | *Farrell (2015) Post-Lasso* | *BCH (2014)* - | *OLS* - |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Estimate | 1,794.34 | 214.72 | 1,495.91 | 1,537.80 | 1,340.24 | 382.28 | 83.17 |
| Standard error (Asy.) | (671.00) | (873.88) | (705.32) | (675.16) | (778.38) | (852.40) | (1184.48) |
| .95 confidence interval (Asy) | [519;3046] | [-1498;1928] | [114;2878] | [214;2861] | [-185;2866] | [-1288;2053] | [-2238;2405] |
| # variables in Propensity Score | none | 8 | 8 | 3 | 3 | 8 | none |
| # variables in Outcome function | none | none | 11 | 16 | 16 | 10 | 172 |

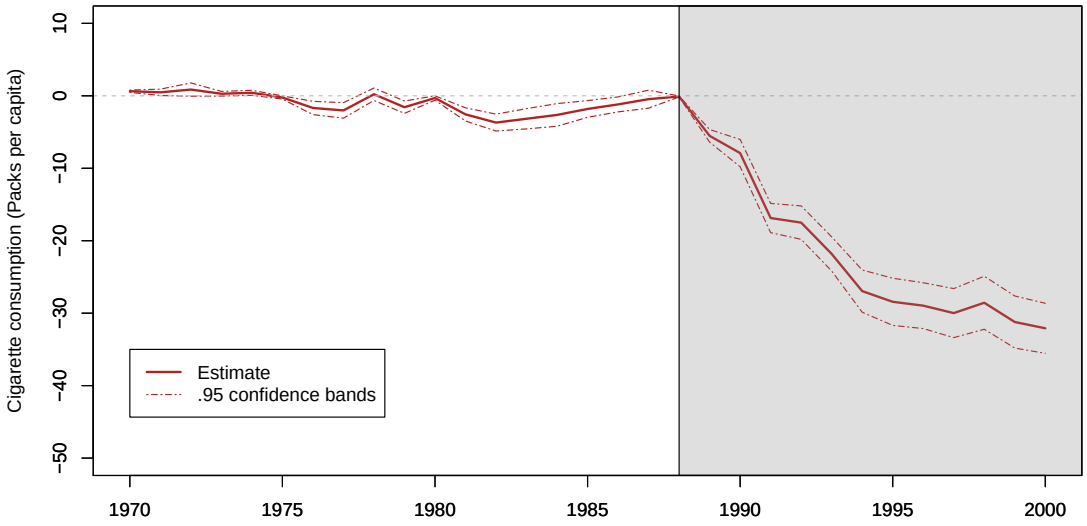
*Note:* The experimental estimate is computed on experimental data, column (1). (Asy.) signals the asymptotic approximation estimator of the quantity is used.

## 6.2 Tobacco Control Program in California

Proposition 99 is one of the first and most ambitious large-scale tobacco control program, implemented in 1989 in California. It includes a vast array of measures, including an increase in cigarette taxation of 25 cents per pack, and a significant effort in prevention and education. In particular, the tax revenues generated by Proposition 99 were used to fund anti-smoking campaigns. Abadie *et al.* (2010) aims at analyzing the impact of the law on actual tobacco consumption in California. Since this program was only enforced in California, it is a classic example where the Synthetic Control method applies, and more standard public policy evaluation tools cannot be used. It is possible to reproduce a synthetic California by reweighting other states so as to imitate California.

For this purpose, Abadie *et al.* (2010) consider the following covariates: retail price of cigarettes, state log income per capita, percentage of population between 15-24, per capita beer consumption (all 1980-1988 averages). 1970 to 1975, 1980 and 1988 cigarette consumptions are also included. Using the same variables, we conduct the same analysis with our estimator. Figure 3 displays the estimated effect of Proposition 99 using the immunized estimator.

**Figure 3:** Proposition 99 effect: Immunized Estimation




Note: The shaded area represents the post-treatment period. The confidence bands are based on the asymptotic variance.

Our estimation finds almost no effect of the policy over the pre-treatment period, which is reassuring in making us think that we have a plausible counter-factual. A steady decline takes

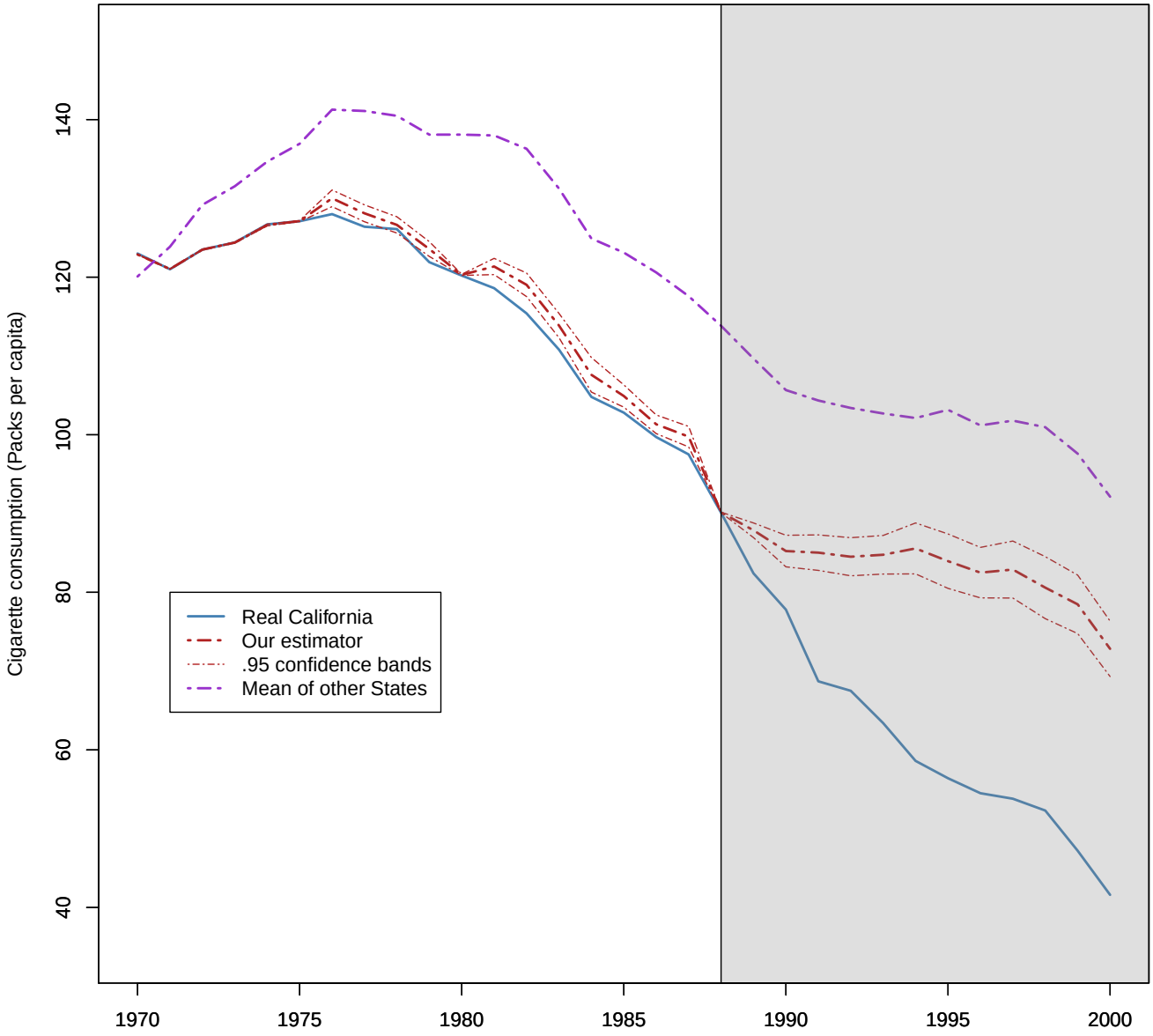
place right after 1988, and in the long-run, the policy is estimated have decreased the tobacco consumption of about 30 packs per capita per year in California. As expected, the variance is larger towards the end of the period, because the data used in the covariates becomes less and less relevant to predict the future. It is also to be noted that by construction, including 1970 to 1975, 1980 and 1988 cigarette consumptions among the covariates yields an almost perfect fit at theses dates because of the immunization procedure (up to the amount of shrinkage induced by the Lasso).

Figure 4 shows how our immunized estimator allows to build a counterfactual which is more credible than both a simple average of per capita tobacco consumption in all other states. The simple unweighted mean of other states over the period displays a higher tobacco consumption, even though it is roughly the same in 1970. California thus appears to have a significantly different pattern of tobacco consumption compared with other USA states even before the adoption of Proposition 99. This highlights that a simple mean of other states, as in a classic difference-in-difference approach, would likely yield a biased estimation of the impact of Proposition 99. In contrast, the Immunized lasso estimator is able to replicate the changes that happened in California over the all pre-treatment period.

Finally, Figure 5 allows a comparison between the immunized estimator and the synthetic control method. The green line is the synthetic control counterfactual. Notice that they do not exactly match the plots of Abadie *et al.* (2010), in which the weights $V$ given to each equation are optimized to best fit the outcome over the whole pre-treatment period. Instead, the green curve optimizes the weights $V$ so as to give a best fit on dates 1970 through 1975, 1980 and 1988. This strategy allows a fairer comparison with our estimator that does not use California's per capita tobacco consumption outside those dates to optimize the fit over whole the pre-treatment period. In other words, one can think of the period 1975-1988 as a semi-placebo test.

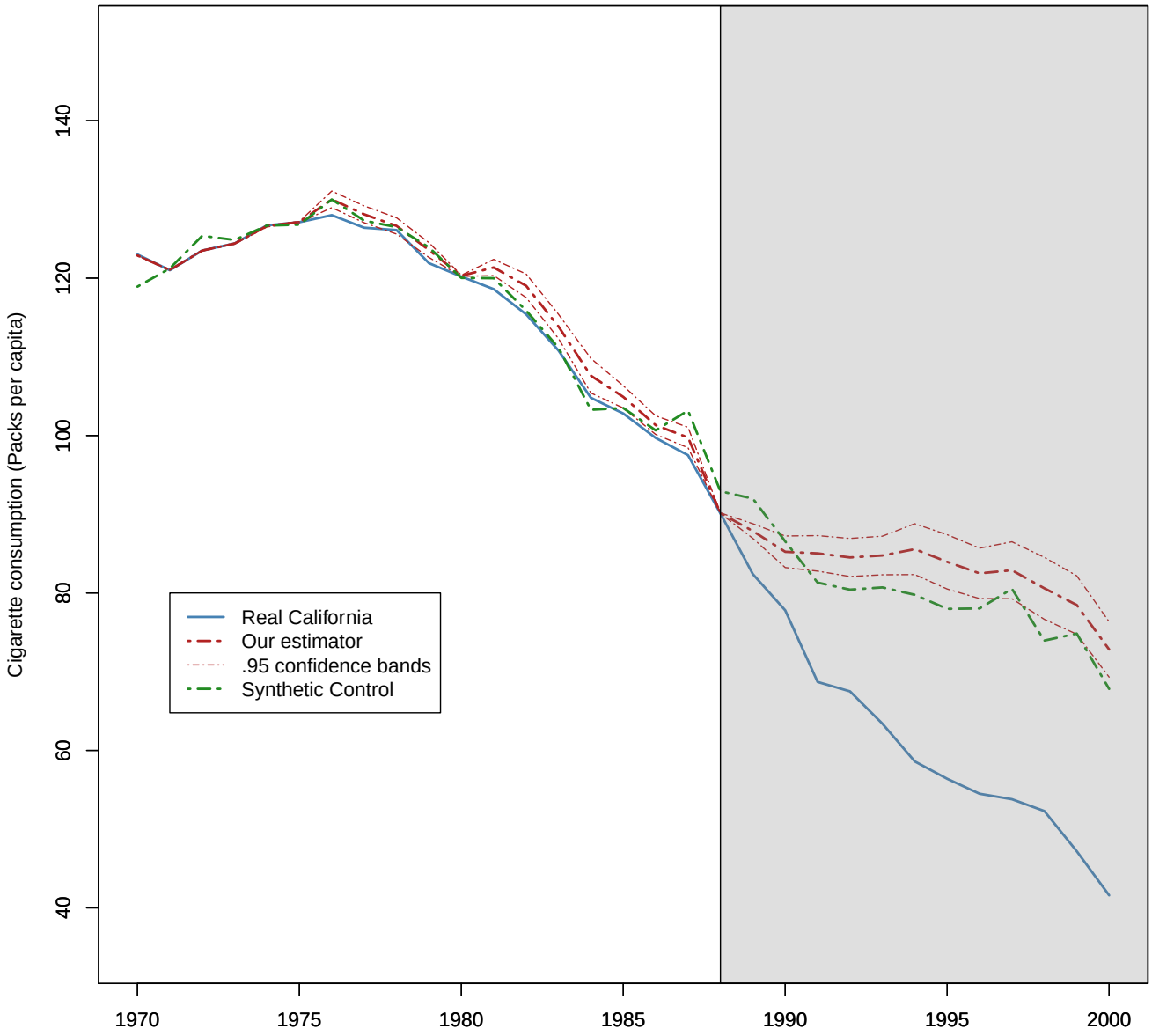
This figure highlights the key role that weights $v$ play in the synthetic control procedure. Indeed, both our estimator and the weighted synthetic control are more credible counterfactuals. Both are able to closely match California tobacco consumption before the policy change. Furthermore, even if our estimator appears to be relatively similar to the weighted synthetic control estimate, it displays a smoother pattern especially towards the end of the 1980s. The estimated effect of California tobacco consumption appears to be bigger with the immunized estimate than with the synthetic control. However, it is hard to conclude that this difference is significant because one cannot easily compute confidence intervals for the synthetic control estimates, without making very stringent assumptions about the program effect.

**Figure 4:** Cigarette Consumption in California: Immunized Estimator




Note: The solid blue line is California tobacco consumption as in the data. The dashed purple line is the mean tobacco consumption of all the other states in the sample. The dashed red line is the immunized Lasso estimator. Only 1970-1974 California tobacco consumption is used for these estimations, meaning that the period 1975-1988 constitutes a placebo test.

**Figure 5:** Cigarette Consumption in California: Immunized Estimation and Synthetic Control




Note: The solid blue line is California tobacco consumption as in the data. The dashed red line is the immunized estimator. The dashed green and yellow lines are the Synthetic Control counterfactuals. For the yellow line, all equations are given the same weight $v$ in the Synthetic Control program, while for the green line, each equation is weighted so as to optimize the fit on dates 1970 through 1975, 1980 and 1988. Only 1970-1974 California tobacco consumption is used for these estimations, meaning that the period 1975-1988 constitutes a placebo test.

# 7 Bibliography


Abadie, A., Diamond, A., and Hainmueller, J. (2010): "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program". *Journal of the American Statistical Association*, 105(490):493–505.

——— (2015): "Comparative Politics and the Synthetic Control Method". *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003): "The Economic Costs of Conflict: A Case Study of the Basque Country". *American Economic Review*, 93(1):113–132.

Acemoglu, D., Hassan, T., and Tahoun, A. (2014): "The Power of the Street: Evidence from Egypt's Arab Spring". CEPR Discussion Papers 10262, C.E.P.R. Discussion Papers.

Addison, J. T., Blackburn, M. L., and Cotti, C. D. (2014): "On the Robustness of Minimum Wage Effects: Geographically-Disparate Trends and Job Growth Equations". Working Paper Series in Economics 330, University of Lüneburg, Institute of Economics.

Allegretto, S., Dube, A., Reich, M., and Zipperer, B. (2013): "Credible Research Designs for Minimum Wage Studies". IZA Discussion Papers 7638, Institute for the Study of Labor (IZA).

Bang, H. and Robins, J. M. (2005): "Doubly Robust Estimation in Missing Data and Causal Inference Models". *Biometrics*, 61(4):962–973.

Belasen, A. R. and Polachek, S. W. (2008): "How Hurricanes Affect Wages and Employment in Local Labor Markets". *American Economic Review*, 98(2):49–53.

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. (2011): "$\ell_1$-penalized quantile regression in high-dimensional sparse models". *Ann. Statist.*, 39(1):82–130.

——— (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.

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

Bilgel, F. and Galle, B. (2015): "Financial incentives for kidney donation: A comparative case study using synthetic controls". *Journal of Health Economics*, 43(C):103–117.

Bohn, S., Lofstrom, M., and Raphael, S. (2011): "Did the 2007 Legal Arizona Workers Act Reduce the State's Unauthorized Immigrant Population?" IZA Discussion Papers 5682, Institute for the Study of Labor (IZA).

Boucheron, S., Lugosi, G., and Massart, P. (2013): *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

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

Card, D. (1990): "The impact of the Mariel boatlift on the Miami labor market". *Industrial and Labor Relations Review*, 43(2):245–257.

Card, D. and Krueger, A. B. (1994): "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania". *American Economic Review*, 84(4):772–93.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015a): "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments". *American Economic Review*, 105(5):486–90.

——— (2015b): "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach". *ArXiv e-prints*.

Claeskens, G. and Hjort, N. L. (2003): "The Focused Information Criterion". *Journal of the American Statistical Association*, 98(464):900–916.

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.

Deville, J.-C., Sarndal, C.-E., and Sautory, O. (1993): "Generalized Raking Procedures in Survey Sampling". *Journal of the American Statistical Association*, 88(423):1013–1020.

Dietrichson, J. and Ellegård, L. M. (2015): "Assist or desist? Conditional bailouts and fiscal discipline in local governments". *European Journal of Political Economy*, 38(C):153–168.

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

Fong, C., Hazlett, C., and Imai, K. (2015): "Parametric and Nonparametric Covariate Balancing Propensity Score for General Treatment Regimes". Unpublished manuscript.

van de Geer, S. A. (2008): "High-dimensional Generalized Linear Models and the Lasso". *Ann. Statist.*, 36(2):614–645.

Gobillon, L. and Magnac, T. (2014): "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Control". Technical report.

Graham, B. S., Pinto, C. C. D. X., and Egel, D. (2012): "Inverse Probability Tilting for Moment Condition Models with Missing Data". *Review of Economic Studies*, 79(3):1053–1079.

Hainmueller, J. (2012): "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies". *Political Analysis*, 20(1):25–46.

Imai, K. and Ratkovic, M. (2014): "Covariate balancing propensity score". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Imbens, G. W. and Rubin, D. B. (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, New York, NY, USA.

Kitagawa, T. and Muris, C. (2015): "Model Averaging in Semiparametric Estimation of Treatment Effects". Technical report, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Kline, P. (2011): "Oaxaca-Blinder as a Reweighting Estimator". *American Economic Review*, 101(3):532–37.

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.

Leeb, H. and Pötscher, B. M. (2005): "Model Selection And Inference: Facts And Fiction". *Econometric Theory*, 21(01):21–59.

——— (2008a): "Recent developments in model selection and related areas". *Econometric Theory*, 24:319–322.

——— (2008b): "Sparse estimators and the oracle property, or the return of Hodges' estimator". *Journal of Econometrics*, 142(1):201–211.

Nannicini, T. and Billmeier, A. (2011): "Economies in Transition: How Important Is Trade Openness for Growth?" *Oxford Bulletin of Economics and Statistics*, 73(3):287–314.

Newey, W. K. and McFadden, D. (1994): "Chapter 36 Large sample estimation and hypothesis testing". *Handbook of Econometrics*, volume 4, pp. 2111 – 2245. Elsevier.

Oliveira, R. (2013): "The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties". *ArXiv e-prints*.

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. (1994): "Regression Shrinkage and Selection Via the Lasso". *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Wassmann, P. (2015): "The Economic Effect of the EU Eastern Enlargement for Border Regions in the Old Member States". Technical report.

# A Algorithm for Feasible Penalty Loadings

The ideal penalty loadings for estimation of $\beta_0$ are given by:

$$\lambda_d := c\Phi^{-1}(1-\gamma/2p)/\sqrt{n}$$
$$\psi_j^d := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(1-D_i)h(X_i^T\beta_0)-D_i\right]^2 X_{i,j}^2}, \text{ for } j=1,...,p$$

The ideal penalty loadings for estimation of $\mu_0$ are given by:

$$\lambda_y := c\Phi^{-1}(1-\gamma/2p)/\sqrt{n}$$
$$\psi_j^y := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(1-D_i)h'(X_i^T\beta_0)^2\left[Y_i-X_i^T\mu_0\right]^2 X_{i,j}^2}, \text{ for } j=1,...,p$$

where $c>1$ is an absolute constant, $\gamma \lesssim \log(p\vee n)$ and $\beta_0$ and $\mu_0$ are the true coefficients. $c$ and $\gamma$ are arbitrary choices and Belloni *et al.* (2012) set $\gamma := .1/\log(p\vee n)$ and $c := 1.1$. However, both $\beta_0$ and $\mu_0$ are unobserved, which prevent from using the ideal penalty loadings. However they can be estimated.

For estimating the penalty loadings $\left\{\psi_j^d\right\}_{j=1}^d$ in the calibration part, we use the following algorithm:

Set a small constant $v>0$ and a maximal number of iterations $K$.

(1) Start by using a preliminary estimate $\beta^{(0)}$ of $\beta_0$ for example using only a constant,[2] then set $\tilde{\psi}_j^{(0)} = \sqrt{\mathbb{E}_n\left[(1-D_i)h(X_i^T\beta^{(0)})-D_i\right]^2 X_{i,j}^2}$, $j=1,...,p$. At step $k$, set $\tilde{\psi}_j^{(k)} = \sqrt{\mathbb{E}_n\left[(1-D_i)h(X_i^T\beta^{(k)})-D_i\right]^2 X_{i,j}^2}$, $j=1,...,p$.

(2) Estimate the model by the Calibration Lasso of equation 24 using the overall penalty level $\lambda$ and penalty loadings found previously, to obtain $\hat{\beta}^{(k)}$.

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

Asymptotic validity of this approach is established in Belloni *et al.* (2012, Lemma 11). The penalty loadings estimation of the immunization step follows a similar procedure. In this specific case, replace $\beta_0$ by $\hat{\beta}$ obtained in the calibration step.

[2] We set $\beta^{(0)}$ with its first element equal to $\log(n_1/n_0)$.

# B Verification of the adaptivity condition

Among the three high-level conditions, the adaptivity is the less obvious one. Achieving adaptivity requires putting some structure on the problem and characterizing the class of DGPs we are dealing with. We consider for that purpose the following conditions.

**Assumption B.1** Approximate Sparsity and Dimension Restrictions
*(i)* $\log(p) = o(n^{1/3})$
*(ii) The nuisance parameter* $\eta_0$ *can be decomposed between two components: a component which is sparse and a component which is small in the following sense:*

$$\eta_0 = \eta_0^a + \eta_0^b, \text{ with } support(\eta_0^a) \cap support(\eta_0^b) = \varnothing.$$
$$\|\eta_0^a\|_0 \leq s, \ \|\eta_0^a\|_1 := C_s < \infty, \|\eta_0^b\|_1 \leq c_1\sqrt{s^2/n}, \ \|\eta_0^b\|_2 \leq c_1\sqrt{s/n}.$$

*At most* $s$ *elements of the sparse components are non-zero and* $\ell_1$ *and* $\ell_2$ *norms are bounded and decreasing to zero for the small component.*

**Assumption B.2** Penalty Loadings

$$\lambda_d := c\Phi^{-1}(1-\gamma/2p)/\sqrt{n}$$
$$\psi_j^d := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(1-D_i)h(X_i^T\beta_0) - D_i\right]^2 X_{i,j}^2}, \text{ for } j = 1, ..., p,$$

*where* $c > 1$ *is an absolute constant,* $\gamma \lesssim \log(p \vee n)$ *and* $\beta_0$ *is the true coefficient. Moreover,*

$$\bar{\psi} := \max_{1\leq j\leq p}\psi_j < \infty, \ \underline{\psi} := \min_{1\leq j\leq p}\psi_j < \infty. \tag{17}$$

$c$ and $\gamma$ are arbitrary choices and Belloni *et al.* (2012) set $\gamma := .1/\log(p \vee n)$ and $c := 1.1$. We let $c_\psi := \frac{\bar{\psi}}{\underline{\psi}}$.

**Assumption B.3** Bounded Covariates
*For every* $i$*:*

$$\max_{1\leq j\leq p}|X_{ij}| := K_n < \infty \tag{18}$$

Define $\Sigma := \mathbb{E}((1-D_1)X_1X_1^T)$, the theoretical Gram matrix on the control group. For a non-empty subset $S \subset \{1, ..., p\}$ and $\alpha > 0$, we also define the set:

$$\mathcal{C}[S, \alpha] := \{v \in \mathbb{R}^p : \|v_{S^C}\|_1 \leq \alpha\|v_S\|_1, v \neq 0\} \tag{19}$$

Finally, let $c_0 = (c+1)/(c-1)$.

**Assumption B.4** Restricted Eigenvalue on Gram matrix
*For* $\alpha \in \{c_0 c_\psi, 2c_0 c_\psi\}$,

$$\kappa_\alpha^2(\Sigma) := \min_{J \subset \{1,\dots,p\}, |J|_0 \leq s} \min_{\delta \in \mathcal{C}[J,\alpha]} \frac{\delta^T \Sigma \delta}{\|\delta_J\|_2^2} > 0.$$

*Moreover, there exists* $h \in ]1, +\infty[$ *such that for all* $v \in \mathbb{R}^p$,

$$\sqrt{\mathbb{E}[(v^T(1-D_1)X_1 X_1^T v)^2]} \leq h v^T \Sigma v.$$

The following theorem shows that the adaptivity condition is satisfied in this context.

**Theorem B.1**
*Suppose that Assumptions 3.1, B.1-B.4 hold. Then the adaptivity condition (1) in Assumption 4.1 holds.*

# C Proofs

Hereafter, we use the notations $a \wedge b = \min(a,b)$, $a \vee b = \max(a,b)$. $a \lesssim b$ means that $a \leq cb$ for some constant $c > 0$ independent of the sample size $n$.

## C.1 Proof of Theorem 3.1

First, note that $\beta_0$ satisfies

$$\mathbb{E}\left[(1-D)h(X^T\beta_0)X\right] = \mathbb{E}\left[DX\right].$$

As a result, for any $\mu \in \mathbb{R}^p$,

$$\mathbb{E}\left[\left(D - (1-D)h(X^T\beta_0)\right)(Y - X^T\mu)\right] = \mathbb{E}\left[\left(D - (1-D)h(X^T\beta_0)\right)Y\right].$$

Now, because $(1-D)Y = (1-D)Y_0$ and $DY = DY_1$, $\theta_0$ verifies the moment condition if and only if:

$$\mathbb{E}\left[(1-D)h(X^T\beta_0)Y\right] = \mathbb{E}\left[DY_0\right].$$

The mean independence assumption allows to write:

$$\mathbb{E}\left[h(X^T\beta_0)Y(1-D)\right] = \mathbb{E}\left[h(X^T\beta_0)\mathbb{E}(1-D|X)\mathbb{E}(Y_0|X)\right]$$

We consider each of the two cases.
1) The linear case: $\mathbb{E}(Y_0|X) = X^T\gamma$.

$$\begin{aligned}
\mathbb{E}\left[h(X^T\beta_0)Y(1-D)\right] &= \mathbb{E}\left[h(X^T\beta_0)(1-D)X^T\gamma\right] \\
&= \mathbb{E}\left[DX^T\gamma\right] \\
&= \mathbb{E}\left[D\mathbb{E}(Y_0|X)\right] \\
&= \mathbb{E}\left[DY_0\right]
\end{aligned}$$

The first equality uses the Mean Independence assumption. The second line follows from the fact that $\beta_0$ is such that $\mathbb{E}[(1-D)h(X^T\beta_0)X] = \mathbb{E}[DX]$.

2) Propensity score given by $P(D=1|X) = G(X^T\beta_0)$.

$$\begin{aligned}\mathbb{E}\left[h(X^T\beta_0)Y(1-D)\right] &= \mathbb{E}\left[h(X^T\beta_0)(1-G(X^T\beta_0))\mathbb{E}(Y_0|X)\right] \\ &= \mathbb{E}\left[G(X^T\beta_0)\mathbb{E}(Y_0|X)\right] \\ &= \mathbb{E}\left[\mathbb{E}(D|X)\mathbb{E}(Y_0|X)\right] \\ &= \mathbb{E}\left[DY_0\right]\end{aligned}$$

## C.2 Proof of Theorem 4.1

For the sake of brevity, let us denote the observed data by $Z_i := (Y_i, D_i, X_i)$ and let $\eta_0 := (\beta_0^T, \mu_0^T)^T$ denote the parameter gathering the two nuisance vectors. Also denote the estimating moment for $\theta_0$ by $g(Z, \theta, \eta) := [D - (1-D)h(X^T\beta)][Y - X^T\mu] - D\theta$ and $\pi$ the probability of being treated: $\pi = \mathbb{P}(D=1)$. Recall that the true values satisfy:

$$\mathbb{E}g(Z, \theta_0, \eta_0) = 0. \tag{20}$$

The following paragraphs gather the main assumptions required to prove the theorem. For sake of simplicity dependence of the data and parameters in $i$ and $n$ is omitted when obvious.

**Assumption C.1**

*Consider a sequence $\{\boldsymbol{P_n}\}$ of sets of probability measures such that for each sequence $\{\mathbb{P}_n\} \in \{\boldsymbol{P_n}\}$:*

*(1) the sequence of random vectors $(Z_i)_{i=1}^n := (Y_i, D_i, X_i)_{i=1}^n$ is independent and identically distributed across $i$,*

*(2)* $\liminf_{n\to\infty} \mathbb{E}\left(D(Y - X^T\mu_0 - \theta_0)^2\right) > 0$,

*(3) There exists $\delta > 0$ such that :* $\mathbb{E}\left(h(X^T\beta_0)^{4+2\delta}\right)/n^{\delta/2} \to 0$,

*(4) and* $\limsup_{n\to\infty} \mathbb{E}\left((Y - X^T\mu_0)^{4+2\delta}\right) < +\infty$,

*(5) (A reecrire correctement)* $\limsup Supp(X'\beta_0) \subset K$ *with $K$ compact of $\mathbb{R}$.*

**Remark C.1**

*Condition (2) can for example stem from lower-level assumptions:* $\liminf_{n\to\infty} Var\left(Y - X^T\mu_0 | D=1\right) > 0$ *and* $\liminf_{n\to\infty} \pi > 0$. *Condition (3) could be linked to lower-level assumptions regarding the propensity score such as* $\mathbb{P}(D=1|X) < 1 - \varepsilon$.

For simplicity denote $\psi_{i,n} := g_n(Z_{i,n}, \theta_{0,n}, \eta_{0,n})$, and recall that $\mathbb{E}\psi_{i,n} = 0$.

**Step 1** Taylor expansion

*By linearity of the estimating moment in $\theta$ and using the mean-value theorem, for a $\tilde{\eta} = t\eta_0 + (1-t)\hat{\eta}$ with $t \in (0,1)$:*

$$\begin{aligned}\mathbb{E}_n g(Z, \hat{\theta}, \hat{\eta}) &= \mathbb{E}_n g(Z, \theta_0, \hat{\eta}) + \hat{\pi}(\theta_0 - \hat{\theta}) \\ &= \hat{\pi}(\theta_0 - \hat{\theta}) + \mathbb{E}_n g(Z, \theta_0, \eta_0) + (\hat{\eta} - \eta_0)^T \mathbb{E}_n \partial_\eta g(Z, \theta_0, \eta_0) \\ &+ \frac{1}{2}(\hat{\eta} - \eta_0)^T \mathbb{E}_n \partial_\eta \partial_{\eta^T} g(Z, \theta_0, \tilde{\eta})(\hat{\eta} - \eta_0)\end{aligned}$$

*By definition of the immunized estimator: $\mathbb{E}_n g(Z, \hat{\theta}, \hat{\eta}) = 0$ so we obtain:*

$$\begin{aligned}\hat{\pi}\sqrt{n}(\hat{\theta} - \theta_0) = &\underbrace{\sqrt{n}\mathbb{E}_n g(Z_i, \theta_0, \eta_0)}_{:=I_1} + \underbrace{\sqrt{n}(\hat{\eta} - \eta_0)^T \mathbb{E}_n \partial_\eta g(Z_i, \theta_0, \eta_0)}_{:=I_2} \\ &+ \underbrace{\frac{\sqrt{n}}{2}(\hat{\eta} - \eta_0)^T \mathbb{E}_n \partial_\eta \partial_{\eta^T} g(Z_i, \theta_0, \tilde{\eta})(\hat{\eta} - \eta_0)}_{:=I_3}\end{aligned}$$

*Now the goal is to show that $I_1$ is asymptotically normal while $I_2$ and $I_3$ tend to zero in probability. We will deal with each term separately over the next steps.*

**Step 2** Normality of $I_1$

*We want to prove:*

$$Var\left(g(Z, \theta_0, \eta_0)\right)^{-1/2} \sqrt{n}\mathbb{E}_n g(Z_i, \theta_0, \eta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

*A Lyapunov condition is necessary to apply the Lindeberg-Feller Central Limit Theorem for triangular arrays. Because we are in a iid case, we need to control the quantity:*

$$\limsup \frac{\mathbb{E}(\psi_{1,n}^{2+\delta})}{\mathbb{E}(\psi_{1,n}^2)^{1+\delta/2}} < +\infty$$

*Consequently, we want to bound the numerator from above. Write:*

$$\psi_{1,n}^{2+\delta} \leq 2^{1+\delta}\left(\underbrace{[D_i - (1-D_i)h(X_i^T\beta_0)]^{2+\delta}(Y_i - X_i^T\mu_0)^{2+\delta}}_{:=A^{2+\delta}} + \underbrace{D_i\theta_0^{2+\delta}}_{:=B^{2+\delta}}\right)$$

*If we assume that $|\theta_0|$ is bounded then $B$ is also bounded. Moreover we'll have:*

$$|A|^{2+\delta} \lesssim \left(1 + h(MC_s)^{2+\delta}\right)\left(Y^{2+\delta} + (MC_s)^{2+\delta}\right)$$

*Where $C_s := \max(\|\mu_0^a\|_1, \|\beta_0^a\|_1)$. For example $C_s := s \times C$ and $h = \exp$.*

*Then we need to bound* $\mathbb{E}(\psi_{1,n}^2)$ *from below.*

$$\mathbb{E}(\psi_{1,n}^2) \geq \mathbb{E}\left((Y - X^T\mu_0 - \theta_0)^2 | D = 1\right)\pi \geq Var\left(Y - X^T\mu_0 | D = 1\right)\pi$$

**Step 3** $I_2$

*The first derivatives of the estimating moment with respect to the nuisance parameters write:*

$$\frac{\partial}{\partial\beta}g(Z,\theta,\eta) = -(1-D)h'(X^T\beta)\left[Y - X^T\mu\right]X$$
$$\frac{\partial}{\partial\mu}g(Z,\theta,\eta) = -\left[D - (1-D)h(X^T\beta)\right]X$$

*Define the random vector* $U_i$ *of size* $2p$ *with each element given by:*

$$U_{ij} := \begin{cases} -(1-D_i)h'(X_i^T\beta_0)\left[Y_i - X_i^T\mu_0\right]X_{ij} & \text{if } 1 \leq j \leq p \\ -\left[D_i - (1-D_i)h(X_i^T\beta_0)\right]X_{ij} & \text{if } p+1 \leq j \leq 2p \end{cases}$$

*Recall that* $\Psi$ *is a square diagonal matrix of dimension* $2p$ *with elements* $\sqrt{\sum_{i=1}^n U_{ij}^2/n}$ *on, its diagonal. Notice that:*

$$\|I_2\|_1 \leq \|\Psi(\hat{\eta} - \eta_0)\|_1 \|\Psi^{-1}\sqrt{n}\mathbb{E}_n\partial_\eta g(Z_i,\theta_0,\eta_0)\|_\infty$$

*From the orthogonality conditions we have that for any* $i$ *and any* $j$ $\mathbb{E}U_{ij} = 0$ *and further assume that* $\mathbb{E}|U_{ij}|^3 < \infty$*. By Lemma D.2 and using also Lemma D.1, with a probability* $o(1)$*:*

$$\|\Psi^{-1}\sqrt{n}\mathbb{E}_n\partial_\eta g(Z_i,\theta_0,\eta_0)\|_\infty \leq \Phi^{-1}(1-\gamma/4p) \leq \sqrt{2\log(2p/\gamma)}$$

*We can also establish that with a probability* $o(1)$*:*

$$\|\Psi(\hat{\eta} - \eta_0)\|_1 \lesssim s\sqrt{\log(2pn)/n}$$

**Step 4** $I_3$

*The second derivatives of the estimating moment with respect to the nuisance parameters write:*

$$\frac{\partial^2}{\partial\beta\partial\beta^T}g(Z,\theta,\eta) = -(1-D)h''(X^T\beta)\left[Y - X^T\mu\right]XX^T$$
$$\frac{\partial^2}{\partial\mu\partial\beta^T}g(Z,\theta,\eta) = \frac{\partial^2}{\partial\beta\partial\mu^T}g(Z,\theta,\eta) = (1-D)h'(X^T\beta)XX^T$$
$$\frac{\partial^2}{\partial\mu\partial\mu^T}g(Z,\theta,\eta) = 0$$

$$\|I_3\|_1 \leq \frac{\sqrt{n}}{2}\|\hat{\eta} - \eta_0\|_2^2 \|\mathbb{E}_n\partial_\eta\partial_{\eta^T}g(Z_i,\theta_0,\tilde{\eta})\|_{sp(m)}$$

*Let us focus on the sparse norm of the matrix of second derivatives. Take any vector $a$ such that $\|a\|_0 \leq m$ and decompose it into its first $p$ elements $a_1$ and the last $p$ elements $a_2$. We can write:*

$$\frac{a^T \mathbb{E}_n \left[\partial_\eta \partial_{\eta^T} g(Z_i, \theta_0, \tilde{\eta})\right] a}{\|a\|^2} \leq \frac{1}{n} \frac{1}{\|a_1\|^2} \sum_{i=1}^n |a_1^T h^{''}(X_i^T \tilde{\beta})(Y_i - X_i \tilde{\mu})(1 - D_i) X_i X_i^T a_1| + \frac{2}{n} \frac{1}{\|a\|^2} \sum_{i=1}^n |a_2^T h^{'}(X_i^T \tilde{\beta})(1 - D_i) X_i X_i^T a_1|$$

*The first term on the left-hand-side of the equation can be bounded:*

$$\frac{1}{n} \frac{1}{\|a_1\|^2} \sum_{i=1}^n |a_1^T h^{''}(X_i^T \tilde{\beta})(Y_i - X_i \tilde{\mu})(1 - D_i) X_i X_i^T a_1| \leq \max_{i=1,\ldots,n} |h^{''}(X_i^T \tilde{\beta})(Y_i - X_i \tilde{\mu})| \frac{1}{n} \frac{\sum_{i=1}^n a_1^T (1 - D_i) X_i X_i^T}{\|a_1\|^2}$$
$$\leq \max_{i=1,\ldots,n} |h^{''}(X_i^T \tilde{\beta})(Y_i - X_i \tilde{\mu})| \|\mathbb{E}_n (1 - D_i) X_i X_i^T\|_{sp(m)}$$

*Using Cauchy-Schwarz inequality within the sum, the second term can also be bounded:*

$$\frac{2}{n} \frac{1}{\|a\|^2} \sum_{i=1}^n |a_2^T h^{'}(X_i^T \tilde{\beta}) X_i X_i^T a_1| \leq \max_{i=1,\ldots,n} |h^{'}(X_i^T \tilde{\beta})| \frac{1}{n} \frac{2}{\|a\|^2} \sum_{i=1}^n \frac{1}{2} \left(\|X_i^T a_1\|_2 + \|X_i^T a_2\|_2\right)$$
$$\leq \max_{i=1,\ldots,n} |h^{'}(X_i^T \tilde{\beta})| \left( \frac{1}{n} \frac{1}{\|a_1\|^2} \sum_{i=1}^n a_1^T (1 - D_i) X_i X_i^T a_1 + \frac{1}{n} \frac{1}{\|a_2\|^2} \sum_{i=1}^n a_2^T (1 - D_i) X_i X_i^T a_2 \right)$$
$$\leq \max_{i=1,\ldots,n} |h^{'}(X_i^T \tilde{\beta})| 2 \|\mathbb{E}_n (1 - D_i) X_i X_i^T\|_{sp(m)}$$

*Collecting the terms, we obtain that:*

$$\sup_{\|a\|_0 \leq m} \frac{a^T \mathbb{E}_n \left[\partial_\eta \partial_{\eta^T} g(Z_i, \theta_0, \tilde{\eta})\right] a}{\|a\|^2} \leq \left( \max_{i=1,\ldots,n} |h^{''}(X_i^T \tilde{\beta})(Y_i - X_i^T \tilde{\mu})| + 2 \max_{i=1,\ldots,n} |h^{'}(X_i^T \tilde{\beta})| \right) \|\mathbb{E}_n (1 - D_i) X_i X_i^T\|_s$$

*We need to bind the term in front of the sparse norm of the empirical Gram matrix. For that notice that:*

$$|h^{''}(X_i^T \tilde{\beta})(Y_i - X_i \tilde{\mu})| \leq |(h^{''}(X_i^T \tilde{\beta}) - h^{''}(X_i^T \beta_0))(Y_i - X_i \tilde{\mu})| + |h^{''}(X_i^T \beta_0) X_i^T (\tilde{\mu} - \mu_0)|$$
$$\leq |\left(h^{''}(X_i^T \tilde{\beta}) - h^{''}(X_i^T \beta_0)\right)\left(Y_i - X_i^T \tilde{\mu}\right)| + |h^{''}(X_i^T \beta_0) X_i^T (\tilde{\mu} - \mu_0)|$$
$$\leq |h^{''}(X_i^T \tilde{\beta}) - h^{''}(X_i^T \beta_0)| \left(|Y_i - X_i^T \mu_0| + |X_i^T (\tilde{\mu} - \mu_0)|\right) + |h^{''}(X_i^T \beta_0)| |X_i^T (\tilde{\mu} - \mu_0)|$$
$$\leq |X_i^T (\hat{\beta} - \beta_0)| \sup_{x \in K} |h^{'''}(x)| \left(|Y_i - X_i^T \mu_0| + |X_i^T (\tilde{\mu} - \mu_0)|\right) + \sup_{x \in K} |h^{''}(x)| |X_i^T (\tilde{\mu} - \mu_0)|$$

*We also have:*

$$|h^{'}(X_i^T\tilde{\beta}) - h^{'}(X_i^T\beta_0)| \leq \sup_{x\in K}|h^{''}(x)||X_i^T(\hat{\beta}-\beta_0)|$$

*Then by assumptions B.3 and B.4 the result follows.*

## C.3 Proof of Theorem B.1

By Lemma 5 in Chernozhukov *et al.* (2015b), the result holds provided that the first-step penalized estimators verify, with probability tending to one,

$$\|\hat{\eta}\|_0 \lesssim s, \tag{21}$$
$$\|\hat{\eta} - \eta_0^a\|_1 \lesssim \sqrt{(s^2/n)\log(pn)}, \tag{22}$$
$$\|\hat{\eta} - \eta_0^a\|_2 \lesssim \sqrt{(s/n)\log(pn)}. \tag{23}$$

### C.3.1 Verification of (21)-(23) for $\widehat{\beta}$

Recall that the estimator is $\hat{\beta}$ such that:

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}(1-D_i)H(X_i^T\beta) - D_i X_i^T\beta + \lambda_d \sum_{j=1}^{p}\psi_j^d|\beta_j|, \tag{24}$$

with ideal penalty loadings satisfying Assumption B.2. Define $\Psi$ the diagonal matrix of dimension $p$ which diagonal is $(\psi_1^d, ..., \psi_p^d)$. The *active set* or *sparsity set* of $\beta_0$ is denoted $S_0$ and defined by $S_0 := \left\{j : \beta_{0j}^a \neq 0\right\}$. The cardinal of $S_0$ is such that $\|S_0\|_0 \leq s$.

**Step 5** Concentration Inequality

We first bound the sup-norm of the gradient of the objective function. Let $U_{i,j} := \left[(1-D_i)h(X_i^T\beta_0) - D_i\right]X_{i,j}$, $\mathcal{S}_j := \sum_{i=1}^n U_{ij}/\sqrt{\sum_{i=1}^n U_{i,j}^2}$ and define the following event

$$\mathcal{B}_\lambda := \left\{\frac{1}{n}\max_{1\leq j\leq p}\left|\sum_{i=1}^{n}\frac{U_{ij}}{\psi_j}\right| \leq \frac{\lambda}{c}\right\}$$

$(X_i, D_i)_{i=1}^n$ is a sequence of i.i.d. random vectors. By construction, $\mathbb{E}(U_{ij}) = 0$. Moreover, by Assumptions B.1 and B.3, the variables $U_{ij}$ have finite third-order moments, $\mathbb{E}(|U_{ij}|^3) \leq +\infty$. Then, by Assumptions B.1 and B.2, we have, by Lemma D.2 :

$$\begin{aligned}
\mathbb{P}\left(\mathcal{B}_\lambda^C\right) &= \mathbb{P}\left(\frac{c}{\sqrt{n}}\max_{1\leq j\leq p}|\mathcal{S}_j| > c\Phi^{-1}(1-\gamma/2p)/\sqrt{n}\right) \\
&= \mathbb{P}\left(\max_{1\leq j\leq p}|\mathcal{S}_j| > \Phi^{-1}(1-\gamma/2p)\right) \\
&\to 0.
\end{aligned}$$

**Step 6** Weighted Restricted Eigenvalue on Empirical Gram matrix

We show that for $\alpha \in \{c_0, 2c_0\}$ and with probability tending to one,

$$\bar{\kappa}_\alpha(\hat{\Sigma}) := \min_{\Psi\delta \in \mathcal{C}[J,\alpha], |J|_0 \leq s} \frac{\sqrt{\delta^T \hat{\Sigma} \delta}}{\|\Psi \delta_J\|_2} > 0,$$

$$\tilde{\kappa}_\alpha(\hat{\Sigma}) := \min_{\Psi\delta \in \mathcal{C}[J,\alpha], |J|_0 \leq s} \sqrt{s} \frac{\sqrt{\delta^T \hat{\Sigma} \delta}}{\|\Psi \delta_J\|_1} > 0.$$

First, we show that Assumption B.4 for $\Sigma$ implies a restricted eigenvalue condition for $\hat{\Sigma}$. Set $\varepsilon := (1 - 7h\sqrt{(p + 2\log(2/\mu))/n})$ for some $\mu \to 0$, $\log(\mu) = o(n)$. By Lemma D.3, we have, for any $\delta \in \mathbb{R}^p$, that $\delta^T \hat{\Sigma} \delta \geq (1 - \varepsilon)\delta^T \Sigma \delta$ with probability tending to one. Then, by Assumption B.4, with probability tending to one,

$$\min_{J \subset \{1,\ldots,p\}, |J|_0 \leq s} \min_{\delta \in \mathcal{C}[J,\alpha]} \frac{\delta^T \hat{\Sigma} \delta}{\|\delta_J\|_2^2} \geq (1 - \varepsilon) \min_{J \subset \{1,\ldots,p\}, |J|_0 \leq s} \min_{\delta \in \mathcal{C}[J,\alpha]} \frac{\delta^T \Sigma \delta}{\|\delta_J\|_2^2} \geq (1 - \varepsilon)\kappa_\alpha^2(\Sigma).$$

Secondly, notice that $\|\Psi \delta_J\|_2 \leq \bar{\psi} \|\delta_J\|_2$. Consequently:

$$\frac{\sqrt{\delta^T \hat{\Sigma} \delta}}{\|\Psi \delta_J\|_2} \geq \frac{1}{\bar{\psi}} \frac{\sqrt{\delta^T \hat{\Sigma} \delta}}{\|\delta_J\|_2}.$$

Moreover, $\Psi\delta \in \mathcal{C}[J, \alpha]$ implies by Assumption B.1 that $\delta \in \mathcal{C}[J, c_\psi \alpha]$. Then with a probability $1 - o(1)$:

$$\kappa_\alpha(\hat{\Sigma}) \geq \frac{\sqrt{1 - \varepsilon}}{\bar{\psi}} \kappa_{c_\psi \alpha}(\Sigma) > 0.$$

The restricted eigenvalue with the $\ell_1$-norm at the denominator holds by the Cauchy-Schwarz inequality $\|\delta\|_1 \leq \sqrt{\|\delta\|_0} \|\delta\|_2$.

**Step 7** Basic Inequality

We prove that with probability tending to one, the estimator $\hat{\beta}$ satisfies:

$$h'\left(-K_n C_n\right)(\hat{\beta} - \beta_0)^T \hat{\Sigma} (\hat{\beta} - \beta_0) \leq 2\lambda \left( \|\Psi \beta_0\|_1 - \|\Psi \hat{\beta}\|_1 \right) + \frac{2\lambda}{c} \|\Psi(\hat{\beta} - \beta_0)\|_1$$

where $C_n := c_\psi (C_s \vee c_1) \left(1 + \sqrt{s^2/n}\right) (c + 1)/(c - 1)$.

By optimality of $\hat{\beta}$:

$$\frac{1}{n} \sum_{i=1}^n \gamma_{\hat{\beta}}^D(X_i, D_i) - \gamma_{\beta_0}^D(X_i, D_i) \leq \lambda \left( \|\Psi \beta_0\|_1 - \|\Psi \hat{\beta}\|_1 \right),$$

where $\gamma_\beta^D(X, D) := (1 - D)H(X^T \beta) - X^T \beta$. Subtract the inner product of the gradient $\bigtriangledown_\beta \gamma_{\beta_0}^D(X_i, D_i)$ and $\hat{\beta} - \beta_0$ on each side:

$$\frac{1}{n}\sum_{i=1}^{n}\gamma^D_{\hat{\beta}}(X_i,D_i)-\gamma^D_{\beta_0}(X_i,D_i)-\left((1-D_i)h(X_i^T\beta_0)-D_i\right)(\hat{\beta}-\beta_0)^TX_i\leq$$
$$\lambda\left(\|\Psi\beta_0\|_1-\|\Psi\hat{\beta}\|_1\right)-\frac{1}{n}\sum_{i=1}^{n}\left((1-D_i)h(X_i^T\beta_0)-D_i\right)(\hat{\beta}-\beta_0)^TX_i$$

Now focus on the left-hand side of the equation. By the mean value theorem there exist $0\leq t\leq 1$ such that if we denote $\tilde{\beta}=t\hat{\beta}+(1-t)\beta_0$:

$$\frac{1}{n}\sum_{i=1}^{n}\gamma^D_{\hat{\beta}}(X_i,D_i)-\gamma^D_{\beta_0}(X_i,D_i)-\left((1-D_i)h(X_i^T\beta_0)-D_i\right)(\hat{\beta}-\beta_0)^TX_i=$$
$$\frac{1}{2}(\hat{\beta}-\beta_0)^T\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i^Th'(X_i^T\tilde{\beta})\right](\hat{\beta}-\beta_0)$$

Plug this into the equation at the beginning of this paragraph and use $|\sum_i a_ib_i|\leq \max_i|b_i|\sum_i|a_i|$ on the right-hand side. Recall the definition of $\mathcal{B}_\lambda$, with probability $1-o(1)$:

$$\frac{1}{2}(\hat{\beta}-\beta_0)^T\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i^Th'(X_i^T\tilde{\beta})\right](\hat{\beta}-\beta_0)\leq \tag{25}$$
$$\lambda\left(\|\Psi\beta_0\|_1-\|\Psi\hat{\beta}\|_1\right)-\frac{1}{n}\sum_{i=1}^{n}\left((1-D_i)h(X_i^T\beta_0)-D_i\right)(\hat{\beta}-\beta_0)^TX_i\leq \tag{26}$$
$$\lambda\left(\|\Psi\beta_0\|_1-\|\Psi\hat{\beta}\|_1\right)+\|\Psi(\hat{\beta}-\beta_0)\|_1\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{X_{i,j}}{\psi_j}\left((1-D_i)h(X_i^T\beta_0)-D_i\right)\right|\leq \tag{27}$$
$$\lambda\left(\|\Psi\beta_0\|_1-\|\Psi\hat{\beta}\|_1\right)+\frac{\lambda}{c}\|\Psi(\hat{\beta}-\beta_0)\|_1 \tag{28}$$

We are now going to show that on $\mathcal{B}_\lambda$ we have $\|\hat{\beta}\|_1\leq c_\psi(C_s\vee c_1)\left(1+\sqrt{s^2/n}\right)(c+1)/(c-1):=C_n$. The expression 25 is non-negative. Moreover, using the definition of event $\mathcal{B}_\lambda$ and provided that $c>1$, we can write:

$$\begin{aligned}
0&\leq\lambda\left(\|\Psi\beta_0\|_1-\|\Psi\hat{\beta}\|_1\right)+\lambda/c\|\Psi(\hat{\beta}-\beta_0)\|_1\\
\lambda\|\Psi\hat{\beta}\|_1&\leq\lambda\|\Psi\beta_0\|_1+\lambda/c\left(\|\Psi\hat{\beta}\|_1+\|\Psi\beta_0\|_1\right)\\
\|\Psi\hat{\beta}\|_1&\leq\frac{c+1}{c-1}\|\Psi\beta_0\|_1\\
\|\hat{\beta}\|_1&\leq\frac{c+1}{c-1}\frac{\bar{\psi}}{\underline{\psi}}\|\beta_0\|_1
\end{aligned}$$

Assumptions B.1 and B.3 are then used. The proof follows with:

$$\frac{h'(-K_nC_n)}{2}(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0) \leq \frac{1}{2}(\hat{\beta}-\beta_0)^T\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i^Th'(X_i^T\tilde{\beta})\right](\hat{\beta}-\beta_0),$$

which gives a lower bound for inequality 25 and gives us the desired result.

**Step 8** Bound on $(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0)$

We prove that with probability tending to one,

$$\sqrt{(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0)} \leq \tilde{b}_n, \text{with } \tilde{b}_n := \frac{4}{h'(-K_nC_n)}\lambda\frac{\sqrt{s}}{\tilde{\kappa}_{c_0}(\hat{\Sigma})} + \frac{2}{\sqrt{h'(-K_nC_n)}}\sqrt{\lambda c_1\bar{\psi}}\frac{\sqrt{s}}{n^{1/4}}.$$

The probability of $\mathcal{B}_\lambda$ is computed in Step 5. The first step of the proof seeks to bound $\|\Psi\beta_0\|_1 - \lambda\|\Psi\hat{\beta}\|_1$. By the triangular inequality:

$$\|\Psi\beta_{0,S_0}\|_1 - \|\Psi\hat{\beta}_{S_0}\|_1 \leq \|\Psi(\beta_{0,S_0}-\hat{\beta}_{S_0})\|_1$$

Focusing on the other part and using $|a-b| \leq |a|+|b|$ :

$$\begin{aligned}\|\Psi\beta_{0,S_0^C}\|_1 - \|\Psi\hat{\beta}_{S_0^C}\|_1 &= 2\|\Psi\beta_{0,S_0^C}\|_1 - \|\Psi\beta_{0,S_0^C}\|_1 - \|\Psi\hat{\beta}_{S_0^C}\|_1 \\ &\leq 2\|\Psi\beta_{0,S_0^C}\|_1 - \|\Psi(\beta_{0,S_0^C}-\hat{\beta}_{S_0^C})\|_1 \\ &\leq 2c_1\bar{\psi}\sqrt{\frac{s^2}{n}} - \|\Psi(\beta_{0,S_0^C}-\hat{\beta}_{S_0^C})\|_1\end{aligned}$$

The last inequality comes from $\|\beta_{0,S_0^C}\|_1 = \|\beta_0^b\|_1$ and assumptions B.1. Consequently:

$$\begin{aligned}&\lambda\|\Psi\beta_0\|_1 - \lambda\|\Psi\hat{\beta}\|_1 + \frac{\lambda}{c}\|\Psi(\hat{\beta}-\beta_0)\|_1 \leq \\ &\lambda\left(1+\frac{1}{c}\right)\|\Psi(\hat{\beta}_{S_0}-\beta_{0,S_0})\|_1 - \lambda\left(1-\frac{1}{c}\right)\|\Psi(\hat{\beta}_{S_0^C}-\beta_{0,S_0^C})\|_1 + 2\lambda c_1\bar{\psi}\sqrt{\frac{s^2}{n}}\end{aligned}$$

On $\mathcal{B}_\lambda$, by the basic inequality:

$$h'(-K_nC_n)(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0) \leq 2\left(\lambda\|\Psi\beta_0\|_1 - \lambda\|\Psi\hat{\beta}\|_1 + \frac{\lambda}{c}\|\Psi(\hat{\beta}-\beta_0)\|_1\right)$$

So :

$$(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0) \leq \tag{29}$$

$$\frac{2}{h'(-K_nC_n)}\lambda\left[\left(1+\frac{1}{c}\right)\|\Psi(\hat{\beta}_{S_0}-\beta_{0,S_0})\|_1 - \left(1-\frac{1}{c}\right)\|\Psi(\hat{\beta}_{S_0^C}-\beta_{0,S_0^C})\|_1 + 2c_1\bar{\psi}\sqrt{\frac{s^2}{n}}\right] \tag{30}$$

For this proof, denote $c^* := \frac{4}{h'(-K_nC_n)}\lambda c_1\bar{\psi}\sqrt{\frac{s^2}{n}}$. Two cases must be distinguished.

(1) If $(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0) \leq c^*$, there is no need to go further since we get a bound for $(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0)$.

(2) If $(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0) > c^*$ using the non-negativity of the left-hand side of the basic inequality, there is a cone condition:

$$\|\Psi(\hat{\beta}_{S_0^C}-\beta_{0,S_0^C})\|_1 \leq \frac{c+1}{c-1}\|\Psi(\hat{\beta}_{S_0}-\beta_{0,S_0})\|_1$$

in other words: $\Psi(\hat{\beta}-\beta_0) \in \mathcal{C}[S_0, c_0]$ with $c_0 := (c+1)/(c-1)$. For clarity of the proof, denote $b^* := \frac{2}{h'(-K_nC_n)}\lambda(1+1/c)\frac{\sqrt{s}}{\tilde{\kappa}_{c_0}(\hat{\Sigma})}$. Furthermore, using Step 6 on the weighted restricted eigenvalue: $\|\Psi(\hat{\beta}_{S_0}-\beta_{0,S_0})\|_1 \leq \sqrt{s}\sqrt{(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0)}/\tilde{\kappa}_{c_0}(\hat{\Sigma})$. Consequently using (29) and dropping the negative term in $\|\Psi(\hat{\beta}_{S_0^C}-\beta_{0,S_0^C})\|_1$:

$$(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0) \leq b^*\sqrt{(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0)} + c^*$$

Because $(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0) > c^*$:

$$(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0) \leq \left(b^* + \sqrt{c^*}\right)\sqrt{(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0)}$$

All in all, in both case, since $b^* > 0$ the bound is:

$$\begin{aligned}\sqrt{(\hat{\beta}-\beta_0)^T\hat{\Sigma}(\hat{\beta}-\beta_0)} &\leq \frac{2}{h'(-K_nC_n)}\lambda(1+1/c)\frac{\sqrt{s}}{\tilde{\kappa}_{c_0}(\hat{\Sigma})} + \frac{2}{\sqrt{h'(-K_nC_n)}}\sqrt{\lambda c_1\bar{\psi}}\frac{\sqrt{s}}{n^{1/4}} \\ &\leq \frac{4}{h'(-K_nC_n)}\lambda\frac{\sqrt{s}}{\tilde{\kappa}_{c_0}(\hat{\Sigma})} + \frac{2}{\sqrt{h'(-K_nC_n)}}\sqrt{\lambda c_1\bar{\psi}}\frac{\sqrt{s}}{n^{1/4}}\end{aligned}$$

since $c > 1$.

**Step 9** Control of the $\ell_1$-error

We prove here that with probability tending to one,

$$\|\hat{\beta}-\beta_0\|_1 \leq 3\frac{c_0}{\underline{\psi}}\frac{\sqrt{s}}{\tilde{\kappa}_{2c_0}(\hat{\Sigma})}\tilde{b}_n \vee 6c_0c_1c_\psi\sqrt{\frac{s^2}{n}}$$

Denote $\delta := \hat{\beta}-\beta_0$. Consider two cases.

(1) Under the cone condition $\Psi\delta \in \mathcal{C}[S_0, 2c_0]$. $c_0 := c + 1/c - 1$. It is obvious that: $\|\Psi\delta_{S_0^C}\|_1 \leq 2c_0\|\Psi\delta_{S_0}\|_1$. Using the restricted eigenvalue assumption, the following

holds:

$$\begin{aligned}\|\Psi\delta\|_1 &\leq (1+2c_0)\|\Psi\delta_{S_0}\|_1 \\ &\leq (1+2c_0)\sqrt{s}\frac{\sqrt{\delta^T\hat{\Sigma}\delta}}{\tilde{\kappa}_{2c_0}(\hat{\Sigma})} \\ &\leq 3c_0\frac{\sqrt{s}}{\tilde{\kappa}_{2c_0}(\hat{\Sigma})}\tilde{b}_n\end{aligned}$$

Recall that $c_0 > 1$ and use bound provided in Step 8.

(2) Assume the other case case: $\|\Psi\delta_{S_0^C}\|_1 > 2c_0\|\Psi\delta_{S_0}\|_1$. From Equation (29) by dropping the term $\delta^T\hat{\Sigma}\delta > 0$ from the left-hand side:

$$\begin{aligned}\|\Psi\delta_{S_0^C}\|_1 &\leq c_0\|\Psi\delta_{S_0}\|_1 + \frac{c}{c-1}2c_1\bar{\psi}\sqrt{\frac{s^2}{n}} \\ &\leq 4\frac{c}{c-1}c_1\bar{\psi}\sqrt{\frac{s^2}{n}}\end{aligned}$$

where the last inequality comes from $\|\Psi\delta_{S_0^C}\|_1 > 2c_0\|\Psi\delta_{S_0}\|_1$. In consequence:

$$\begin{aligned}\|\Psi\delta\|_1 &\leq \left(1+\frac{1}{2c_0}\right)\|\Psi\delta_{S_0^C}\|_1 \\ &\leq \left(1+\frac{1}{2c_0}\right)4c_1\frac{c}{c-1}\bar{\psi}\sqrt{\frac{s^2}{n}} \\ &\leq 6c_0c_1\bar{\psi}\sqrt{\frac{s^2}{n}}\end{aligned}$$

Where the last inequality uses $(1+1/2c_0) \leq 3/2$ and $c/(c-1) \leq c_0$.

**Step 10** Control of the $\ell_2$-error.

This step is devoted to provide a bound for the $\ell_2$-error of the Lasso. We use Bickel *et al.* (2009), end of the proof of Theorem 7.1.

Cone condition: on $\mathcal{B}_\lambda$,

$$\|\Psi(\widehat{\beta}_{S_0^C} - \beta_{0,S_0^C})\|_1 \leq c_0\|\Psi(\widehat{\beta}_{S_0} - \beta_{0,S_0})\|_1$$

with $c_0 = (c+1)/(c-1)$

Then, applying the reasoning of BRT (p.1729, between "It is easy" and B.28):

$$\|\Psi(\widehat{\beta} - \beta_0)\|_2 \leq \left(1 + c_0\sqrt{s/m}\right)\|\Psi(\widehat{\beta}_{J_{01}} - \beta_{0,J_{01}})\|_2, \tag{31}$$

where $J_{01} = S_0 \cup S_1$, where $S_1$ is the set of the $m$ largest elements of $\widehat{\beta} - \beta_0$ ($= \widehat{\beta}$ here) not in $S_0$ (careful, adjust for approximate sparsity here).

Next, using (29), we have

$$\begin{aligned}(\hat{\beta} - \beta_0)^T \hat{\Sigma} (\hat{\beta} - \beta_0) &\leq \frac{2}{h'(-K_n C_n)} \lambda \left[ \left(1 + \frac{1}{c}\right) \|\Psi(\hat{\beta}_{S_0} - \beta_{0,S_0})\|_1 + 2c_1 \bar{\psi} \sqrt{\frac{s^2}{n}} \right] \\ &\leq \frac{2}{h'(-K_n C_n)} \lambda \left[ \left(1 + \frac{1}{c}\right) \sqrt{s} \|\Psi(\hat{\beta}_{S_0} - \beta_{0,S_0})\|_2 + 2c_1 \bar{\psi} \sqrt{\frac{s^2}{n}} \right] \\ &\leq \frac{2}{h'(-K_n C_n)} \lambda \left[ \left(1 + \frac{1}{c}\right) \sqrt{s} \|\Psi(\hat{\beta}_{J_{01}} - \beta_{0,J_{01}})\|_2 + 2c_1 \bar{\psi} \sqrt{\frac{s^2}{n}} \right].\end{aligned}$$

Then, we have to use TO DO a suitable version of the $RE(s, m, c_0)$ in BRT to get:

$$(\hat{\beta} - \beta_0)^T \hat{\Sigma} (\hat{\beta} - \beta_0) \geq C \|\Psi(\hat{\beta}_{J_{01}} - \beta_{0,J_{01}})\|_2^2.$$

Preciser $C$! As a result,

$$\|\Psi(\hat{\beta}_{J_{01}} - \beta_{0,J_{01}})\|_2 \leq \frac{4\lambda c_1}{C h'(-K_n C_n) - 2\lambda(1 + 1/c)\sqrt{s}} \bar{\psi} \sqrt{\frac{s^2}{n}}.$$

Finally, combined with (31), we obtain

$$\|\Psi(\widehat{\beta} - \beta_0)\|_2 \leq \left(1 + c_0 \sqrt{s/m}\right) \frac{4\lambda c_1}{C h'(-K_n C_n) - 2\lambda(1 + 1/c)\sqrt{s}} \bar{\psi} \sqrt{\frac{s^2}{n}}$$

### C.3.2 Verification of (21)-(23) for $\widehat{\mu}$

Recall that $\widehat{\mu}$ is such that:

$$\hat{\mu} = \arg\min_{\mu} \frac{1}{n} \sum_{i=1}^{n} (1 - D_i) h'(X_i^T \hat{\beta}) \left(Y_i - X_i^T \mu\right)^2 + \lambda_y \sum_{j=1}^{p} \psi_j^y |\mu_j|,$$

This estimator is a weighted version of the usual Lasso. The steps needed to achieve (21)-(23) for $\widehat{\mu}$ closely follow the ones from the previous subsection or can be found in Belloni *et al.* (2012) for example. The main concern is that the weights are estimated and thus should be taken into account. For the sake of clarity we will state the points where the proof differs from the one before.

**Step 11** Concentration Inequality

*Another concentration inequality is needed to bound the sup-norm of the gradient of the objective function. Let* $V_{i,j} := (1 - D_i) h'(X_i^T \beta_0) \left[Y_i - X_i^T \mu_0\right] X_{i,j}$, $\mathcal{S}_j := \sum_{i=1}^n V_{ij} / \sqrt{\sum_{i=1}^n V_{i,j}^2}$ *and define the following event*

$$\mathcal{B}'_\lambda := \left\{ \frac{1}{n} \max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \frac{V_{ij}}{\psi_j} \right| \leq \frac{\lambda^y}{c} \right\}$$

$(Y_i, X_i, D_i)_{i=1}^n$ *is a sequence of i.i.d. random vectors. By construction,* $\mathbb{E}(V_{ij}) = 0$. *Moreover, by Assumptions B.1 and B.3, the variables* $V_{ij}$ *have finite third-order moments,* $\mathbb{E}(|V_{ij}|^3) \leq +\infty$. *Then, by Assumptions B.1 and B.2, we have, by Lemma D.2 :*

$$\begin{aligned}\mathbb{P}\left(\mathcal{B}_\lambda'^C\right) &= \mathbb{P}\left(\frac{c}{\sqrt{n}}\max_{1\leq j\leq p}|\mathcal{S}_j| > c\Phi^{-1}(1-\gamma/2p)/\sqrt{n}\right)\\ &= \mathbb{P}\left(\max_{1\leq j\leq p}|\mathcal{S}_j| > \Phi^{-1}(1-\gamma/2p)\right)\\ &\to 0.\end{aligned}$$

Finally, using this concentration inequality and previous results on the consistency rate of $\hat{\beta}$ the gradient of the objective function for $\hat{\mu}$ can be bounded and the result follows.

**Step 12** Basic Inequality for $\hat{\mu}$

*This step is complicated because the empirical loss function for* $\hat{\mu}$ *depends on* $\hat{\beta}$ *rather than* $\beta_0$. *Denote* $\gamma^Y_{\beta,\mu}(Z_i) := (1-D_i)h'(X_i^T\beta)\left(Y_i - X_i^T\mu\right)^2$. *Use the decomposition:*

$$\frac{1}{n}\sum_{i=1}^n \gamma^Y_{\beta_0,\hat{\mu}}(Z_i) - \gamma^Y_{\beta_0,\mu_0}(Z_i) = R_n + \frac{1}{n}\sum_{i=1}^n \gamma^Y_{\hat{\beta},\hat{\mu}}(Z_i) - \gamma^Y_{\hat{\beta},\mu_0}(Z_i),$$

*with* $R_n := (1/n)\sum_{i=1}^n \gamma^Y_{\beta_0,\hat{\mu}}(Z_i) - \gamma^Y_{\hat{\beta},\hat{\mu}}(Z_i) - \gamma^Y_{\beta_0,\mu_0}(Z_i) + \gamma^Y_{\hat{\beta},\mu_0}(Z_i)$. *Rewrite this remainder term as:*

$$R_n = \frac{1}{n}\sum_{i=1}^n (1-D_i)\left[h'(X_i^T\beta_0) - h'(X_i^T\hat{\beta})\right]X_i^T(\mu_0-\hat{\mu})\left[2\left(Y_i - X_i^T\mu_0\right) - X_i^T(\hat{\mu}-\mu_0)\right].$$

*With probability tending to one,* $[\min_i X_i'\widehat{\beta}, \max_i X_i'\widehat{\beta}] \subset K$ *compact and if* $h'$ *is Lipschitz on this compact,*

$$|h'(X_i'\widehat{\beta}) - h'(X_i'\beta_0)| \leq C_{h'}|X_i'(\widehat{\beta}-\beta_0)|,$$

*we can state:*

$$\begin{aligned}|R_n| &\leq C_{h'}K_n\|\hat{\beta}-\beta_0\|_1\frac{1}{n}\sum_{i=1}^n(1-D_i)|X_i^T(\mu_0-\hat{\mu})\left[2\left(Y_i - X_i^T\mu_0\right) - X_i^T(\hat{\mu}-\mu_0)\right]|\\ &\leq C_{h'}K_n\|\hat{\beta}-\beta_0\|_1\frac{1}{n}\sum_{i=1}^n 2(1-D_i)|X_i^T(\mu_0-\hat{\mu})\left(Y_i - X_i^T\mu_0\right)| + (1-D_i)\left(X_i^T(\hat{\mu}-\mu_0)\right)^2\\ &\leq C_{h'}K_n\|\hat{\beta}-\beta_0\|_1\|\Psi^y(\hat{\mu}-\mu_0)\|_1\left(\frac{1}{n}\sum_{i=1}^n 2(1-D_i)\max_{j=1,\ldots,p}\frac{|X_{i,j}|}{\psi_j^y}|(Y_i - X_i^T\mu_0)|\right)\\ &+ C_{h'}K_n\|\hat{\beta}-\beta_0\|_1(\hat{\mu}-\mu_0)^T\left(\frac{1}{n}\sum_{i=1}^n(1-D_i)X_iX_i^T\right)(\hat{\mu}-\mu_0).\end{aligned}$$

*Because $\hat{\mu}$ is the minimizer of the empirical loss function we obtain:*

$$\frac{1}{n}\sum_{i=1}^{n}\gamma^{Y}_{\hat{\beta},\hat{\mu}}(Z_i)-\gamma^{Y}_{\hat{\beta},\mu_0}(Z_i)\leq\lambda^{y}\left(\|\Psi^{y}\mu_0\|_1\ -\|\Psi^{y}\hat{\mu}\|_1\right)$$

*On the other hand, we have:*

$$\frac{1}{n}\sum_{i=1}^{n}\gamma^{Y}_{\beta_0,\hat{\mu}}(Z_i)-\gamma^{Y}_{\beta_0,\mu_0}(Z_i)+2(1-D_i)h'(X_i^T\beta_0)X_i^T(\hat{\mu}-\mu_0)(Y_i-X_i^T\mu_0)=$$
$$(\hat{\mu}-\mu_0)^T\left(\frac{1}{n}\sum_{i=1}^{n}(1-D_i)h'(X_i^T\beta_0)X_iX_i^T\right)(\hat{\mu}-\mu_0)$$

*Combining the previous equality with the two previous inequalities we get:*

$$\begin{aligned}(\hat{\mu}-\mu_0)^T\left(\frac{1}{n}\sum_{i=1}^{n}(1-D_i)h'(X_i^T\beta_0)X_iX_i^T\right)(\hat{\mu}-\mu_0)&\leq\lambda^{y}\left(\|\Psi^{y}\mu_0\|_1\ -\|\Psi^{y}\hat{\mu}\|_1\right)\\&+C_{h'}K_n\|\hat{\beta}-\beta_0\|_1\|\Psi^{y}(\hat{\mu}-\mu_0)\|_1\left(\frac{2}{n}\sum_{i=1}^{n}(1-D_i)\max_{j=1,\dots,p}\frac{|X_{i,j}|}{\psi_j}|(Y_i-X_i^T\mu_0)|\right)\\&+C_{h'}K_n\|\hat{\beta}-\beta_0\|_1(\hat{\mu}-\mu_0)^T\left(\frac{1}{n}\sum_{i=1}^{n}(1-D_i)X_iX_i^T\right)(\hat{\mu}-\mu_0)\\&+\frac{2}{n}\sum_{i=1}^{2}(1-D_i)h'(X_i^T\beta_0)X_i^T(\hat{\mu}-\mu_0)(Y_i-X_i^T\mu_0).\end{aligned}$$

*Using similar arguments as in equation 25, one can show that on $\mathcal{B}'_\lambda$:*

$$\begin{aligned}(1-C_{h'}K_n\|\hat{\beta}-\beta_0\|_1)(\hat{\mu}-\mu_0)^T\left(\frac{1}{n}\sum_{i=1}^{n}(1-D_i)h'(X_i^T\beta_0)X_iX_i^T\right)(\hat{\mu}-\mu_0)&\leq\lambda^{y}\left(\|\Psi^{y}\mu_0\|_1\ -\|\Psi^{y}\hat{\mu}\|_1\right)\\&+\left(\frac{\lambda^{y}}{c}+C_{h'}K_n\|\hat{\beta}-\beta_0\|_1\left(\frac{2}{n}\sum_{i=1}^{n}(1-D_i)\max_{j=1,\dots,p}\frac{|X_{i,j}|}{\psi_j}|(Y_i-X_i^T\mu_0)|\right)\right)\|\Psi^{y}(\hat{\mu}-\mu_0)\end{aligned}$$

# D Useful Lemmas

**Lemma D.1** Upper tail bound of the Normal distribution
$\forall a\in]0,1/2[$*:*

$$\Phi^{-1}(1-a)\leq\sqrt{-2\log(2a)}\tag{32}$$

*Where $\Phi$ and $\Phi^{-1}$ are the distribution and quantile functions of a standard Normal random variable.*

*Proof.* By Chernoff's bound (Boucheron *et al.*, 2013, p. 22):

$$\sup_{x>0}\left[(1-\Phi(x))e^{x^2/2}\right]=1/2$$

Provided $u\in]1/2,1[$, set $x=\sqrt{-2\log(2(1-u))}$. Because $\Phi^{-1}$ is increasing:

$$\Phi^{-1}(u)=\sqrt{-2\log(2(1-u))}$$

□

**Lemma D.2** Moderate Deviation Inequality for Maximum of a Vector
*Suppose that* $\mathcal{S}_j:=\sum_{i=1}^n U_{ij}/\sqrt{\sum_{i=1}^n U_{i,j}^2}$, *where* $U_{ij}$ *are independent random variables across* $i$ *with mean zero and finite third-order moments. Then:*

$$\mathbb{P}\left(\max_{1\leq j\leq p}|\mathcal{S}_j|>\Phi^{-1}(1-\gamma/2p)\right)\leq\gamma\left(1+\frac{A}{\ell_n^3}\right)$$

*where* $A$ *is an absolute constant, provided for* $\ell_n>0$*:*

$$0\leq\Phi^{-1}(1-\gamma/2p)\leq\frac{n^{1/6}}{\ell_n}\min_{1\leq j\leq p}M_j-1,\ M_j:=n^{-1/6}\frac{\left(\sum_{i=1}^n\mathbb{E}[U_{ij}^2]\right)^{1/2}}{\left(\sum_{i=1}^n\mathbb{E}[|U_{ij}|^3]\right)^{1/3}}\tag{33}$$

*Proof.* See Belloni *et al.* (2012, p. 2409). □

**Lemma D.3** Lower Tail of Random Quadratic Form
*Assume* $A_1,\ldots,A_n\in\mathbb{R}^{p\times p}$ *are i.i.d. random positive semidefinite matrices whose coordinates have bounded second moments. Define* $\Sigma:=\mathbb{E}(A_1)$ *and* $\hat{\Sigma}:=\sum_{i=1}^n A_i/n$. *Let* $h\in]1,+\infty[$ *be such that* $\sqrt{\mathbb{E}[(v^TA_1v)^2]}\leq hv^T\Sigma v$, *for all* $v\in\mathbb{R}^p$. *Then for any* $\delta\in]0,1[$*:*

$$\mathbb{P}\left(\forall v\in\mathbb{R}^p:v^T\hat{\Sigma}v\geq\left(1-7h\sqrt{\frac{p+2\log(2/\delta)}{n}}\right)v^T\Sigma v\right)\geq1-\delta$$

*Proof.* See Oliveira (2013, Theorem 3.1). □