

Estimation of Average Treatment Effects using Penalization

Ashkan Ertefaie, Masoud Asgharian and David A. Stephens

Department of Mathematics and Statistics,
McGill University

805 Sherbrooke Street West, Montreal, Quebec, Canada.

email: `ashkan@math.mcgill.ca`, `masoud@math.mcgill.ca` and
`dstephens@math.mcgill.ca`

February 29, 2012

Abstract

We address the common problem faced by practitioners of how to select variables for the propensity score model commonly used to correct for confounding in observational data. We construct a double robust regression procedure which incorporates the treatment assignment model as well as the response mean model for estimating the treatment effect. In an analytic example, we show that selecting important covariates can increase efficiency of estimation of the causal parameter while retaining consistency. We demonstrate that treatment-only variable selection techniques may ignore important covariates which are strongly related to the outcome but not to the treatment. We introduce a novel covariate selection technique based on penalized likelihood which considers the response and treatment assignment models simultaneously. The selected covariates via our proposed method can then be used in other causal adjustment techniques as well as our proposed regression estimator. We derive the asymptotic properties of the estimators using a novel method that does not require the second derivative of the joint density function, and illustrate their small-sample behaviour using simulation. We apply the proposed method to analyze the National Supported Work Demonstration (NSWD) data.

KEYWORDS: Causal inference, Average treatment effect, Propensity score, Variable selection, Penalized likelihood, Oracle estimator.

1 Introduction

In the analysis of observational data, when attempting to establish the magnitude of the causal effect of treatment (or exposure) in the presence of confounding, the practitioner is faced with certain modelling decisions that facilitate the estimation. Most commonly, two statistical models must be proposed; (i) the *conditional mean model* that models the expected response as a function of predictors, and (ii) the *treatment allocation model* that describes the mechanism via which treatment is allocated (or, at least, received) by individuals in the study, again as a function of the predictors.

Predictors that appear in both mechanisms (i) and (ii) are termed *confounders*, and their omission from model (ii) is typically regarded as a serious error, as it leads to inconsistent estimators of the treatment effect. Because of this, practitioners usually adopt a conservative approach, and attempt to ensure that they do not omit confounders by fitting a richly parameterized treatment allocation model. Note that if confounders are omitted (so that the treatment allocation mechanism is not ignorable) most causal adjustment mechanisms will lead to inconsistent estimators of the treatment effect.

The conservative approach, however, can lead to predictors of treatment allocation, but not response, being included in the treatment allocation model. The inclusion of such “spurious” variables in model (ii) is usually regarded as harmless. However, the typical reported forfeit for this conservatism is inflation of variance of the effect estimator (Greenland (2008) and Schisterman et al. (2009)). This problem also applies to the conditional mean model, but is in practice less problematic as practitioners seem to be more concerned with adjustment for confounding, and therefore more liable to introduce the spurious variables in model (ii). Little formal guidance as to how the practitioner should act in this setting has been provided.

As has been conjectured and studied in simulation by Brookhart et al. (2006a), it is plausible that variable selection will lead to appreciable efficiency gains. However, confounder selection methods based on either the treatment assignment or the response models may result in failure to account for important confounders which barely predict the treatment or the response, respectively (Crainiceanu et al. (2008)). Vansteelandt et al. (2010) show that confounder selection procedures based on AIC and BIC can be sub-optimal and introduce a method based on the focused information criterion (FIC) which targets the treatment effect by minimizing a prediction mean square error (see also the cross-validation method of Brookhart and van der Laan (2006b)). van der Laan and Gruber (2010) introduces a method to select the sufficient and minimal variables necessary for the propensity score model to estimate an unbiased causal effect by inspecting the efficient influence curve. Their proposed iterative algorithm achieves its goal as long as the conditional mean model is correctly specified (Porter et al. (2011)); this assumption may be quite restrictive when the covariate dimension is large.

In this paper, we demonstrate the variance inflation caused by the inclusion of spurious vari-

ables, and provide a technique based on penalization that provides automatic consistent variable selection for both conditional mean and treatment allocation models. Unlike the standard results in penalized regression literature, we drive the asymptotic properties of our proposed estimator using a method that does not require the second derivative of the joint density function. We restrict attention to the *unmediated* causal effect (where the effect of exposure on outcome is not mediated by an intermediate variable); in the presence of mediation, direct and indirect effects may not in general be identifiable (see Robins and Greenland (1992), Petersen et al. (2006), Robins et al. (2010), and Hafeman and VanderWeele (2010)).

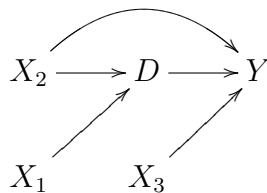
2 Preliminaries

We now present some preliminary concepts and establish notation. We let $Y(d)$ denote the (potential) response to treatment d , and D denote the treatment received. We will assume three types of predictors:

- (I) *treatment predictors* (X_1), which are related to treatment and not to outcome.
- (II) *confounders* (X_2), which are related to both outcome and treatment.
- (III) *outcome modifiers* (X_3), which are related to outcome and not to treatment.

See the directed acyclic graph (DAG) in Figure 1.

Figure 1: Covariate types: Type-I: X_1 , Type-II: X_2 and Type-III: X_3 .



We restrict attention here to the (artificially simple) situation where each predictor can be classified into one of these three types, and to single time-point studies. In addition, as is usual, we will make the assumption of no unmeasured confounders, that is, that treatment received D and potential response to treatment d , $Y(d)$ are independent, given the measured predictors X . In any practical situation, to facilitate causal inference, the analyst must make an assessment as to the structural nature of the relationships between the variables encoded by the DAG in Figure 1.

2.1 The Propensity Score for binary treatments

The propensity score, π , for binary treatment D is defined as $\pi(x) = \Pr(D = 1|x)$, where x is a p -dimensional vector of (all) covariates. Rosenbaum and Rubin (1983) show that π is the coarsest function of covariates that exhibits the balancing property, that is, that treatment assignment is independent of covariates, given the propensity score: $D \perp X|\pi$. As a consequence, the causal effect $\mu = \mathbb{E}[Y(1) - Y(0)]$ can be computed by iterated expectation

$$\mu = \mathbb{E}_X[\mathbb{E}[Y(1)|X] - \mathbb{E}[Y(0)|X]] = \mathbb{E}_\pi[\mathbb{E}[Y(1)|\pi] - \mathbb{E}[Y(0)|\pi]] \quad (1)$$

where $Y(1)$ and $Y(0)$ are counterfactual outcomes if treated and untreated, respectively, and $\mathbb{E}_\pi$ denotes the expectation with respect to the distribution of π . For more details see Rubin (2008) and Rosenbaum (2010).

The assumption of no unmeasured confounders permits us to write, for $d = 0, 1$, that $\mathbb{E}[Y(d)|X] \equiv \mathbb{E}[Y|D = d, X]$, so that the latter expectation can be estimated using, say, a regression of the observed Y on observed X and D . The crucial implication of the balancing property of π is that, after conditioning on π , we again have independence of $Y(d)$ and D ; for $d = 0, 1$, $\mathbb{E}[Y(d)|\pi] \equiv \mathbb{E}[Y|D = d, \pi]$, and again a regression approach can be used.

Remark 1: In the standard formulation of the propensity score, no distinction is made between our three types of covariates. Note that, however, for consistent estimation of μ , *it is not necessary to balance on covariates that are not confounders*. Covariates X_1 that predict D but not Y may be unbalanced in treated and untreated groups, but will not affect the estimation of the effect of D on Y , as D will be conditioned upon, thereby blocking any effect of X_1 . Covariates X_3 are unrelated to D , so will by assumption be in balance in treated and untreated groups in the population. Therefore, the propensity score need only be constructed from confounding variables X_2 ; in this case, it is easy to see that the propensity score, π_2 , say, is a balancing score in the sense that $D \perp X_2|\pi_2$: we have $\Pr(D = 1|\pi_2(x_2) = t, X_2 = x_2) = t = \Pr(D = 1|\pi_2(x_2) = t)$, independent of x_2 , in the usual way. Then, in the presence of outcome modifiers X_3 of Y , the sequel to equation (1) takes the form

$$\begin{aligned} \mu = \mathbb{E}[Y(1) - Y(0)] &= \mathbb{E}_{X_2, X_3}[\mathbb{E}[Y(1)|X_2, X_3] - \mathbb{E}[Y(0)|X_2, X_3]] \\ &= \mathbb{E}_{\pi_2, X_3}[\mathbb{E}[Y(1)|\pi_2, X_3] - \mathbb{E}[Y(0)|\pi_2, X_3]]. \end{aligned} \quad (2)$$

2.2 Propensity Score Regression

Robins et al. (1992) utilized a semiparametric regression technique (see also Engle et al. (1986); Chamberlain (1987)) to adjust for confounders by using a function of covariates as a single scalar covariate which can be estimated by imposing a parametric model; in the binary exposure case,

this function is equivalent to the propensity score. We refer to this approach as Propensity Score Regression (PSR). We adopt a two-stage estimation approach, in which we first fit a propensity score model, then a conditional mean model in the observed exposure D and the fitted values of π . The conditional mean model has the form $Y_i = \mu(D_i, \pi_i; \theta) + \epsilon_i$, where the disturbance ϵ_i is independent of (D, π) and $\mathbb{E}[\epsilon_i] = 0$. In general, $\mu(\cdot)$ is unknown and investigators need to perform model-checking to examine the adequacy of the different possible candidates. Robinson (1988) and Little and An (2004) utilize nonparametric regression in the conditional mean model. Huber and Ronchetti (1981) use a model where the number of the confounders increases with sample size.

2.3 Penalized Estimation

In a given parametric model, if η is a p -dimensional regression coefficient, $p_\lambda(\cdot)$ is a penalty function and $l_m(\eta)$ is the minus log-likelihood, the maximum penalized likelihood (MPL) estimator $\hat{\eta}_{ml}$ is defined as

$$\hat{\eta}_{ml} = \arg \min_{\eta} \left\{ l_m(\eta) + n \sum_{j=1}^p p_\lambda(|\eta_j|) \right\},$$

MPL estimators are shrinkage estimators, and as such, they have more bias, though less variation than unpenalized analogues. Commonly used penalty functions include LASSO (Tibshirani (1996)), SCAD (Fan and Li (2001)), EN (Zou and Hastie (2005)) and HARD (Antoniadis (1997)). A penalized estimator has the oracle properties if it sets the true zero coefficients to zero (sparsity) and behaves like the MLE for large samples. While LASSO satisfies sparsity, it does not behave like the MLE when $n \rightarrow \infty$. Fan and Li (2001) introduced the SCAD penalty function to avoid these deficiencies.

The remainder of this paper is organized as follows. In Section 3, we show analytically how the variable selection can reduce the variance of the estimator. Section 4 introduces a penalized (pseudo-) causal likelihood and defines the penalized estimator corresponding to the derived likelihood, and we study its theoretical properties. The performance of the proposed method is studied via simulation in Section 5. We analyze the NSW data in Section 6, and Section 7 contains concluding remarks. Proofs are given in the Appendix.

3 A Simple Illustrative Example

We now explain the importance of performing model selection in the context of causal adjustment modelling. First, we establish that true outcome modifiers can be advantageously placed in either the response model or the propensity score model to **decrease** the variance of the causal effect

estimator. Secondly, we demonstrate that inclusion of treatment predictors in the propensity score model will lead to **inflation** of the variance of the estimator.

3.1 Including outcome modifiers in the propensity score model

Following the above notation, assume the **true** conditional mean model is

$$\mathbb{E}[Y|D = d, X = x] = \beta_0 + \beta d + \beta_2 x_2 + \beta_2 x_3,$$

with **true** propensity score model $\text{logit}(\pi(x)) = \alpha_0 + \alpha_2 x_2$ where X_2 is a confounder (i.e. there are no pure predictors of exposure) and where β is the causal parameter of interest. For estimation of β , suppose the **fitted** conditional mean model

$$\mathbb{E}[Y|D = d, X = x] = \beta_0 + \beta d + \theta \text{logit}(\pi(x)), \quad (3)$$

and the **fitted** propensity score model $\text{logit}(\pi(x)) = \alpha_0 + \alpha_2 x_2 + \alpha_3 x_3$, that includes the effect modifier X_3 which is not actually a predictor of exposure, are used. The logit transform maps $\pi(x)$ onto the whole real line; this strategy often appears to improve the performance of the propensity score regression for a continuous response.

Inclusion of X_3 into the fitted propensity score model increases the covariance between the fitted π and Y , which, by usual OLS arguments, decreases the variance of the estimated causal effect, in line with the simulations of Brookhart et al. (2006a). Adding X_3 to the propensity score model still yields a balancing score (as confounders X_2 are already included) and the estimator remains unbiased.

3.2 Including treatment predictors in the propensity score model

Consider now the effect of including a spurious covariate in the propensity score model, specifically, a covariate that is **not** related to outcome. Suppose that the **true** conditional mean model is $\mathbb{E}[Y|D = d, X = x] = \beta_0 + \beta d + \beta_2 x_2$, and that the **fitted** conditional mean model is identical to (3). To study the impact of including the spurious (non-confounder, pure predictor of treatment) covariate X_1 , suppose that the **true** propensity score model $\text{logit}(\pi(x)) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$ is actually fitted. Denote by $Z_{1,2} = \text{logit}(\hat{\pi}(x_1, x_2))$ the fitted PS random variable obtained from the first stage fit. Consider also the random variable $Z_2 = \text{logit}(\hat{\pi}(x_2))$ using a specification $\text{logit}(\pi(x)) = \alpha_0 + \alpha_2 x_2$ that omits X_1 . Consider estimation of β using the two models

$$Y = \beta d + \theta Z_{1,2} + \epsilon \quad (4a)$$

$$Y = \beta d + \theta Z_2 + \epsilon \quad (4b)$$

Both models yield consistent estimators of β . By standard OLS arguments, for each model, and for large n , the variance of $\hat{\beta}$ simplifies to

$$\frac{\sigma^2}{n\text{Var}[D](1 - \text{Corr}[D, Z]^2)}$$

which is a monotonic increasing function of $|\text{Corr}[D, Z]|$. As X_1 is a predictor of D , $Z_{1,2}$ is (in expectation) more correlated with D than is Z_2 , and hence the variance of $\hat{\beta}$ is larger when model (4a) is fitted. The extent of the difference in correlation depends on the magnitude of α_1 . This result is essentially the well-known result relating to the inclusion of spurious predictors in a linear regression where it is known that those spurious predictors are correlated with the true predictors.

These simple examples suggest that if our aim is to decrease the variance and mean square error (MSE) of the estimator, we should include in the propensity score model those variables that are related to the response, and exclude those that are merely related to the treatment mechanism. It also suggests that the common strategy of building rich (saturated) models for the propensity score will lead to inefficiency if covariates that are not confounders are included. Excluding confounders leads to bias and inconsistency, but the impact on MSE is difficult to quantify (although see the multivariate normal example of Vansteelandt et al. (2010)). The extension to more general or complex models is straightforward in principle, although the algebra is more involved, and large sample arguments are typically needed.

4 Penalized Estimation for Causal Parameters

In this section, we develop the methodology for the construction of a penalized estimation approach suitable for propensity score regression modelling. We adopt a semi-parametric penalized likelihood approach, and establish the theoretical properties of the resulting MPL estimators.

4.1 An Adjusted PS Regression Model

We propose a slight modification of the conventional propensity score regression approach of Robins et al. (1992), and specify

$$\mathbb{E}[Y_i|S_i = s_i, X_i = x_i] = \beta_0 + \beta_1 s_i + g(x_i; \alpha), \quad (5)$$

where $S_i = D_i - \mathbb{E}[D_i|x_i] = D_i - \pi(x_i)$, $g(x; \alpha)$ is a function of covariates and π is the propensity score. The quantity S_i is used in place of D_i ; if D_i is used the fitted model may

result in a biased estimator for β_1 since $g(x; \alpha)$ may be incorrectly specified. By defining S_i in this way, we restore $\text{cor}[S_i, X_{ij}] = 0$ for $j = 1, 2, \dots, p$ (if $\pi(x_i) = \mathbb{E}[D_i|x_i]$ is correctly specified), as $\pi(x_i)$ is the (fitted) expected value of D_i , and hence $\mathbf{x}'_j(D - \pi(x)) = 0$, where $\mathbf{x}'_j = (x_{1j}, \dots, x_{nj})$. Therefore misspecification of $g(\cdot)$ will not result in an inconsistent estimator.

In general, this model is *doubly robust* (see Davidian et al. (2005), Schafer and Kang (2005) and Bang and Robins (2005)); it yields a consistent estimator of β_1 if *either* the propensity score model or conditional mean model (5) is correctly specified, and is the most efficient estimator (Tsiatis (2006)) when both are correctly specified. For additional details on the related asymptotic and finite sample behaviour, see Kang and Schafer (2007), Neugebauer and van der Laan (2005), van der Laan and Robins (2003) and Robins (1999).

4.2 Likelihood construction

Consider the parametric likelihood $\mathcal{L}(\eta; y, d, x)$ (based on a specific parametric submodel) proportional to

$$\prod_{i=1}^n f(y_i|d_i, \pi_i, g(x_i; \alpha), \beta) p(d = 1|x_i, \alpha)^{d_i} p(d = 0|x_i, \alpha)^{1-d_i}, \quad (6)$$

where β is an r_1 -dimensional vector and α is an r_2 -dimensional vector containing parameters that appear in the model for $Y|X$ and $D|X$ respectively. Thus $\eta = (\beta, \alpha)$ is an r -dimensional vector, where $r = r_1 + r_2$. Since our goal is to select covariates for the propensity score model, we just impose a penalty on the propensity score parameters - there is no penalization of the β parameters. Therefore, using (6) the penalized log likelihood can be written as

$$\tilde{l}_n(\eta) = l_{ny}(\eta) + l_{nd}(\eta) - np_{\lambda_n}(\alpha) \quad (7)$$

where, if $\psi(x_i, \alpha) = \log(p(D = 1|x_i, \alpha)/p(D = 0|x_i, \alpha))$,

$$l_{ny}(\eta) = \sum_{i=1}^n \log f(y_i|d_i, \pi_i, g(x_i; \alpha), \beta) \quad l_{nd}(\eta) = \sum_{i=1}^n \{d_i \psi(x_i, \alpha) + \log p(D = 0|x_i, \alpha)\}.$$

The penalized pseudo-density for $z_i = (y_i, d_i, x_i)$ is $f_p(z_i, \eta) = f(z_i; \eta)f(\alpha)$, where $f(\alpha) = \exp\{-p_{\lambda_n}(\alpha)\}$. The MPL estimator, $\hat{\eta}_n$, is defined by

$$\hat{\eta}_n = \arg \sup_{\eta} \prod_{i=1}^n f_p(z_i; \eta) = \arg \sup_{\eta} \sum_{i=1}^n \log f_p(z_i; \eta).$$

4.3 Omission of confounders

The omission of confounders induces bias in the estimation of causal effect, and any covariate selection technique must take into account both outcome and treatment assignment models simultaneously. By using the joint likelihood, we give the covariates a chance to represent their contribution twice, once in the conditional mean model and once in the treatment assignment model. This strategy then gives an equal chance to Type-II and Type-III covariates for selection as key covariates. To deal with this problem, we introduce the *boosting* parameter ν which boosts covariates Type-III relative to Type-II. The boosting parameter can be defined as $\nu = 1/|\tilde{\alpha}_Y|$, where $\tilde{\alpha}_Y$ is the least-squares estimate of the parameters in the response mean model. Our penalty function is proportional to the boosting parameter, $p_{\lambda_n}(\cdot) = \nu p_{\lambda_n}^*(\cdot)$, where $p_{\lambda_n}^*(\cdot)$ is a conventional penalty function. Therefore, the magnitude of penalty on each parameter is proportional to its contribution to the response and treatment model. A similar argument can be found in the adaptive LASSO (Zou (2006)).

4.4 Main Theorems

The following conditions guarantee a consistent penalized estimating procedure which sets the small coefficients to zero for covariate selection.

- P1. For all n , $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(\alpha)$ is symmetric and nonnegative. Moreover, it is twice differentiable with derivatives $p'_{\lambda_n}(\cdot)$ and $p''_{\lambda_n}(\cdot)$ and nondecreasing for $\alpha \neq 0$ with at most a few exceptions.
- P2. As $n \rightarrow \infty$, $\max_{\alpha \neq 0} \{p''_{\lambda_n}(\alpha)\} \rightarrow 0$ and $\max_{\alpha \neq 0} \{\sqrt{n}p'_{\lambda_n}(\alpha)\} \rightarrow 0$.
- P3. For $N_n \equiv (0, M_n)$, $\liminf_{\alpha \in N_n} \sqrt{n}p'_{\lambda_n}(\alpha) = \infty$, where $M_n \rightarrow 0$ as $n \rightarrow \infty$.

Assumption P1 is used to prove Theorem 1, while P2 prevents the j th element of the penalized likelihood from being dominated by the penalty function since it vanishes when $n \rightarrow \infty$. If $\alpha_j = 0$, condition P3 allows the penalty function to dominate the penalized likelihood which leads to the sparsity property.

Suppose the r -dimensional vector of parameters $\eta = (\eta_1, \eta_2)$ are the coefficients in a fitted regression models corresponding to (6). Let the true parameter values be $\eta_0 = (\eta_{01}, 0)$, such that $\eta_2 = (\eta_j) = 0$ for $j = s+1, \dots, r$; s denotes the true number of predictors present in the model. Note that since there is no penalty on the β s, η_2 consists of those α that should be shrunk to zero ($\alpha_j = 0$ for $j = s', \dots, r_2$). Theorem 1 establishes the existence of the consistent penalized maximum likelihood estimator with respect to the joint likelihood (6) under standard regularity conditions (Ibragimov and Has' Minskii (1981)) given as C1-C4 in the Appendix.

Theorem 1 *Under assumptions C1-C4 and P1-P3, the penalized maximum likelihood estimator $\hat{\eta}_n$ converges to η_0 as $n \rightarrow \infty$ almost surely.*

Lemma 2 in the Appendix shows that the proposed penalized estimation technique is able to shrink to zero coefficients attached to covariates not present in the true model. In the next theorem, we show that the proposed method asymptotically sets the detected zero coefficients to zero.

Theorem 2 *If C1-C4 and P1-P3 hold, then $Pr(\hat{\eta}_2 = 0) \rightarrow 1$ as $n \rightarrow \infty$.*

The next result presents asymptotic normality of the score function and the asymptotic distribution of the MPL estimator. Let $I(\eta)$ be the Fisher information matrix derived from the constructed likelihood.

Theorem 3 *Suppose assumptions C1-C5 and P1-P3 are fulfilled and further $\det[I(\eta)] \neq 0$ for $\eta \in \Xi$. Then*

$$(a) \quad \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\frac{\partial \log f(z, \eta_{01})}{\partial \eta_{01}} \right] - \sqrt{n} p'_{\lambda_n}(\alpha_{01}) \xrightarrow{d} N(0, I(\eta_{01})).$$

$$(b) \quad \sqrt{n}(\hat{\eta}_{01} - \eta_{01}) \xrightarrow{d} N(0, I^{-1}(\eta_{01})).$$

where $\eta_{01} = (\beta, \alpha_{01})$ and α_{01} is the true vector of non-zero coefficients.

The proof of Theorem 3 is omitted since it is similar to the proof of Theorem 2.1.1 in Ibragimov and Has'minskii (1981) and Theorem 2.5.2 in Bickel et al. (1993). Theorem 3 states that asymptotically, under P_2 the penalized estimator is as efficient as the ML estimator when the important covariates are known *a priori*.

Remark 2. By Theorem 3, the estimator obtained by the LASSO penalty function will be a $\sqrt{n}$ consistent estimator for non-zero coefficients if $\sqrt{n}\lambda_n \rightarrow 0$, whereas it achieves the sparsity property if $\sqrt{n}\lambda_n \rightarrow \infty$. Therefore, the LASSO penalty function cannot satisfy both sparsity and consistency simultaneously. However, in adaptive LASSO, the sparsity property will be achieved if $n\lambda_n \rightarrow \infty$; therefore, it exhibits the oracle properties. For the HARD and SCAD penalties, P2 and P3 hold if $\lambda_n \rightarrow 0$ and $\sqrt{n}\lambda_n \rightarrow \infty$.

4.5 Choosing the Tuning Parameter

We select the tuning parameter using the *Generalized Cross Validation* (GCV) method suggested by Tibshirani (1996) and Fan and Li (2001). Let

$$\text{GCV}(\lambda) = \frac{\text{RSS}(\lambda)/n}{\{1 - d(\lambda)/n\}^2},$$

Table 1: Penalized ATE estimators based on the SCAD and LASSO penalty functions.

$\sigma = 2$		$n = 500$			$n = 5000$		
Method	ρ	Bias	S.D.	MSE	Bias	S.D.	MSE
SCAD	0	0.012	0.230	0.053	0.000	0.065	0.004
	0.5	0.003	0.233	0.055	0.005	0.064	0.004
LASSO	0	0.003	0.254	0.065	0.002	0.065	0.004
	0.5	0.023	0.246	0.061	0.005	0.064	0.004
Oracle	0	0.006	0.195	0.038	0.000	0.065	0.004
	0.5	0.010	0.187	0.035	0.001	0.064	0.004
Saturated	0	0.001	0.326	0.106	0.002	0.123	0.015
	0.5	0.031	0.329	0.109	0.003	0.125	0.016
$\sigma = 4$		$n = 500$			$n = 5000$		
SCAD	0	0.025	0.541	0.290	0.006	0.127	0.016
	0.5	0.009	0.537	0.288	0.007	0.124	0.016
LASSO	0	0.055	0.618	0.385	0.003	0.127	0.016
	0.5	0.090	0.562	0.316	0.008	0.124	0.015
Oracle	0	0.021	0.387	0.150	0.008	0.127	0.016
	0.5	0.012	0.372	0.139	0.002	0.121	0.015
Saturated	0	0.039	0.899	0.810	0.008	0.332	0.111
	0.5	0.026	0.902	0.814	0.014	0.343	0.118

where $RSS(\lambda) = \|y - \mathbf{X}\hat{\eta}\|^2$, $d(\lambda) = \text{trace}\{\mathbf{X}(\mathbf{X}'\mathbf{X} + n\Sigma_\lambda(\hat{\eta}))^{-1}\mathbf{X}'\}$ is the effective number of parameters and $\Sigma_\lambda(\eta) = \text{diag}\{p'_\lambda(|\eta_1|)/|\eta_1|, \dots, p'_\lambda(|\eta_p|)/|\eta_p|\}$. The selected tuning parameter $\hat{\lambda}$ is defined by $\hat{\lambda} = \arg \min_\lambda \text{GCV}(\lambda)$.

5 Simulation Studies

5.1 The joint penalized likelihood approach

In this section, we represent simulation study based on 1000 data sets of sizes 5000 and 500 to check the performance of the proposed penalized covariate selection method. We assume the following model:

$$D \sim \text{Bernoulli} \left(\frac{\exp\{0.2x_1 - x_2 + 3x_8 - x_9 + x_{10}\}}{1 + \exp\{0.2x_1 - x_2 + 3x_8 - x_9 + x_{10}\}} \right)$$

with $Y \sim \text{Normal}(d + 2x_1 + 0.5x_2 + 5x_3 + 5x_4, \sigma)$, where X_k for $k = 1, \dots, 10$ are normal random variables each with mean $\mu = 1$, variance σ^2 . Table 1 shows the results based on the LASSO and SCAD penalties for different values of σ and ρ where $\rho = \text{Corr}[X_i, X_j]$ for $i \neq j$.

In Tables 1 and 2, *Oracle* refers to the proposed model by Brookhart et al. (2006a) which includes just covariates related to the response in the propensity score model while *Saturated* refers

Table 2: Penalized ATE estimators based on the SCAD and LASSO penalty functions.

Method	ρ	MRME	Correct	Incorrect	MRME	Correct	Incorrect
$\sigma = 2$		$n = 500$			$n = 5000$		
SCAD	0	0.583	5.5	0.0	0.401	6	0
	0.5	0.596	5.4	0.0	0.413	6	0
LASSO	0	0.622	5.2	0.0	0.394	6	0
	0.5	0.538	5.3	0.0	0.316	6	0
$\sigma = 4$		$n = 500$			$n = 5000$		
SCAD	0	0.355	5.1	0.03	0.240	6	0
	0.5	0.346	5.2	0.03	0.215	6	0
LASSO	0	0.377	5.1	0.02	0.220	6	0
	0.5	0.339	5.2	0.00	0.212	6	0

to the propensity score model including all the covariates $(X_1, \dots, X_{10})$. The fitted propensity score model in Oracle and Saturated depend, respectively, on $x_1, \dots, x_4$, and $x_1, \dots, x_{10}$.

Table 1 confirms that our proposed technique converges to the Oracle model as sample size increases. For moderate sample sizes, the covariate selection technique results in a lower MSE than the Saturated model, decreasing both bias and standard error. In fact, the standard error of the treatment effect obtained by the Saturated model is roughly twice that obtained using the penalized technique. Table 2 summarizes the performance of LASSO and SCAD in terms of the median of relative model errors (MRME). The MRME is defined as the median of the following quantity

$$\text{RME} = \frac{(\hat{\beta} - \beta)' \mathbb{E}[\mathbf{X}'\mathbf{X}](\hat{\beta} - \beta)}{(\hat{\beta}_{\text{sat}} - \beta)' \mathbb{E}[\mathbf{X}'\mathbf{X}](\hat{\beta}_{\text{sat}} - \beta)},$$

where $\hat{\beta}$ is the estimator obtained from the penalized method and $\hat{\beta}_{\text{sat}}$ is the one obtained from the saturated model. The average number of coefficients set to zero correctly or incorrectly are also reported in Table 2. For moderate sample sizes, SCAD outperforms LASSO when covariates are independent, while LASSO performs better when there is correlation between covariates. However, for large sample size, LASSO outperforms SCAD in terms of the MRME for both correlated and uncorrelated covariates and both noise levels. In both cases, increasing the variability from 2 to 4 with moderate sample size reduces the MRME.

5.2 Performance of the double robust estimator

The asymptotic behaviour of the proposed double robust PSR estimator can be established using standard techniques for such results. Here, we focus on the small sample behaviour of the estimator, and conduct a simulation study to investigate the performance of the proposed estimator when the $g(\cdot)$ function in our proposed estimator is misspecified. The data generating process here is similar to the one used in the previous section 5.1. In Table 3, superscript s

refers to the saturated model for the propensity score and PS is the conventional PSR estimator based on $\mathbb{E}[Y] = \beta_0 + \beta_1 D + \beta_2 \pi(x)$. We assume that the $g(\cdot)$ function in the proposed double robust PSR estimator is misspecified and it is set to be $g(\alpha, x) = \alpha_1 x_1$. We assume that the $g(\cdot)$ function in the proposed double robust PSR estimator is misspecified and it is set to be $g(\alpha, x) = \alpha_1 x_1$. We also studied the case when both the propensity score and response models are misspecified.

In this example, the R^2 in the correct response model is %98 while in the misspecified model (i.e. when just (d, x_1) are included in the y model), $R^2 = \%10$.

Table 3: Performance of the proposed double robust estimator compared with other estimators. DR-PSR is the proposed double robust propensity score regression method; IPW is inverse probability weighting; PS is the conventional propensity score regression. The superscript ^s refers to the saturated model, without model selection. The superscript ^m refers to the case when both propensity score and the response model are misspecified.

Method	$n = 500$			$n = 5000$		
	Bias	S.D.	MSE	Bias	S.D.	MSE
DR-PSR	0.017	0.371	0.138	0.002	0.120	0.014
IPW	0.005	0.624	0.389	0.002	0.177	0.031
PS	0.009	0.393	0.155	0.001	0.120	0.015
DR-PSR ^s	0.016	1.028	1.057	0.004	0.309	0.096
IPW ^s	0.005	4.737	22.435	0.073	1.496	2.242
PS ^s	0.002	1.000	1.000	0.003	0.308	0.095
DR-PSR ^m	0.587	1.985	4.283	0.644	0.606	0.782
IPW ^m	2.168	2.134	9.252	2.197	0.655	5.257
PS ^m	2.173	2.093	9.100	2.191	0.646	5.216

Our study shows that although the predictability of the misspecified y model is quit low (about %10), the proposed estimator outperforms the IPW in terms of MSE even when the $g(\cdot)$ function is completely misspecified. It also outperforms the conventional propensity score regression estimator for moderate sample sizes and performs similarly for large sample sizes. Our result indicates that our method performs well and as such does not appear to suffer from the flaws identified in Kang and Schafer (2007). See the commentary given by Gruber and van der Laan (2011) about estimating the causal effect using a doubly robust estimator under dual misspecification.

5.3 Comparison with the standard conditional approach

Here, we want to compare the proposed joint penalized likelihood approach with the conditional response penalized likelihood approach. Our simulation study in this section reveals that missing

Table 4: Comparing the joint and conditional penalized likelihood approaches.

Method	Bias	S.D	MSE	Bias	S.D	MSE
$\sigma = 4$	$n = 5000$			$n = 30000$		
LASSO ¹	0.003	0.122	0.015	0.002	0.050	0.002
LASSO ²	0.121	0.117	0.028	0.118	0.047	0.016
Saturated	0.014	0.288	0.083	0.002	0.119	0.014

LASSO¹ represents the results based on the proposed penalized joint likelihood approach. LASSO² represents the results based on the conditional response penalized likelihood approach using the adaptive LASSO.

a confounder which is weakly related to the response but strongly related to the treatment may lead to an inconsistent estimator. This type of confounder can be ignored by applying the conventional penalized covariate selection methods. Assume that

$$D \sim \text{Bernoulli} \left(\frac{\exp\{0.2x_1 - 2x_2 + 3x_8 - x_9 + x_{10}\}}{1 + \exp\{0.2x_1 - 2x_2 + 3x_8 - x_9 + x_{10}\}} \right)$$

and $Y \sim \text{Normal}(d + 2x_1 + 0.05x_2 + 5x_3 + 5x_4, \sigma)$, where the X_k are independent $\text{Normal}(1, 4^2)$ random variables.

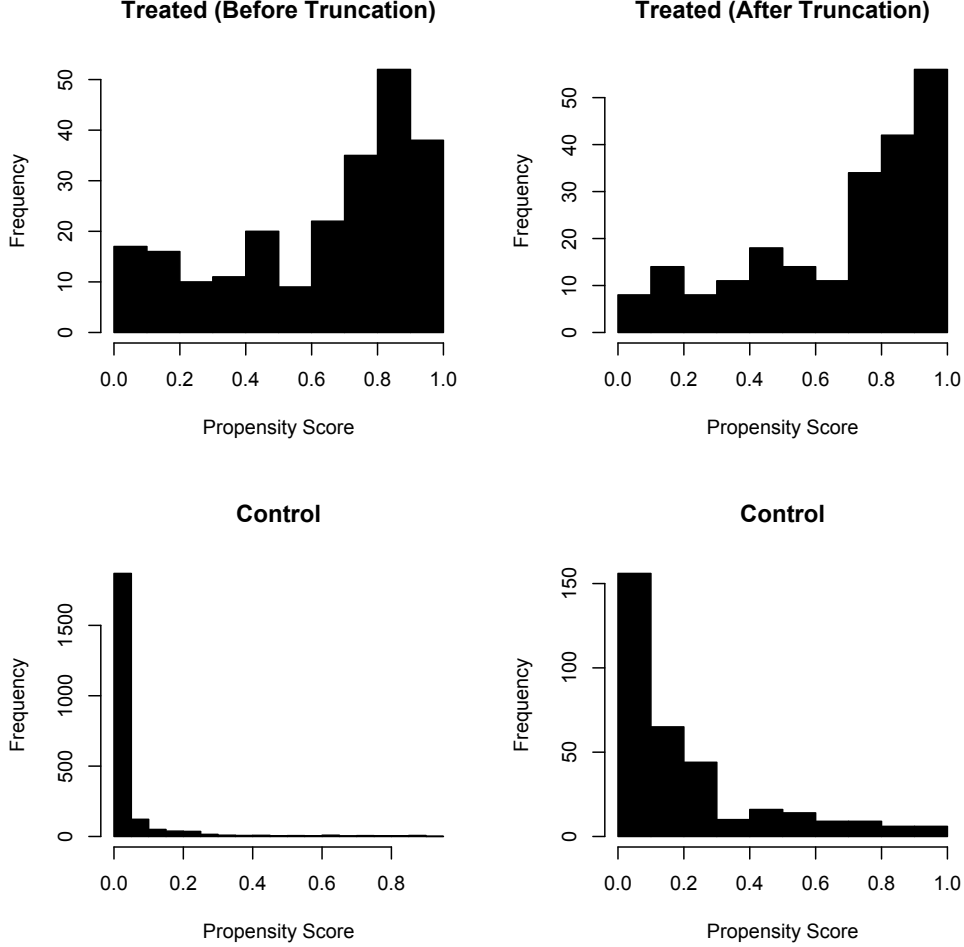
Table 4 indicates that missing x_2 which is weakly related to the response produces a bias and as sample size increases, the bias dominates the variance, which results in an inconsistent estimator.

6 Application to Real Data

The National Supported Work Demonstration (NSWD) aimed to provide work experience to individuals who lacked basic skills. In 1970, a total of 6,616 workers were sampled at random from different cities across the United States by the Manpower Demonstration Research Corporation (MDRC). Qualified individuals were assigned to a treatment group to receive training provided by the NSW program. Some baseline variables such as age (X_1), education (X_2), African American (X_3), Hispanic (X_4), married status (X_5), no degree (X_6) and earnings in 1975 (X_7) were recorded for each participant. Like LaLonde (1986), we focus on the male workers among the participants and consider their earnings in 1978 as the response variable. Our treated observations group consists of a subset of the NSWD data with 297 workers and the control group includes individuals from the Panel Study of Income Dynamics (PSID) with 2490 workers. Additional details are provided in Dehejia and Wahba (1999).

The initial propensity score is estimated using logistic regression including all the covariates.

Figure 2: The NSW data: Estimated propensity score overlap in treated and control groups before and after truncation.



The first step in the propensity score causal adjustment method is to check for sufficient overlap between treatment and control groups (Rubin (2008)). The two plots on the left side of Figure 2 reveal that there are some subclasses (bins) that include only treated or control individuals which may result in a poor causal effect estimate. To overcome this problem, we truncated those individuals fall in empty bins. The right side of Figure 2 shows that sufficient overlap can be achieved by truncating the empty bins; there are treated and untreated workers in all the propensity score subclasses.

After ensuring confounder balance inside each propensity subclass, we need to specify the conditional mean model. Here, we fit the model $\mathbb{E}[Y] = \beta_0 + \beta_1 s + g(x; \alpha)$, where β_1 is the treatment effect parameter and Y is the log response. Interaction or the higher order of the

Table 5: Penalized ATE estimators based on the SCAD and LASSO penalty functions.

Method	ATE	S.D.	C.I.	X_1	X_2	X_3	X_4	X_5	X_6	X_7
SCAD	-0.552	0.091	(-0.734,-0.370)	–	✓	–	✓	✓	✓	–
LASSO	-0.565	0.094	(-0.753,-0.377)	–	✓	–	✓	✓	✓	–
Saturated	-0.502	0.107	(-0.716,-0.288)	✓	✓	✓	✓	✓	✓	✓
Unadjusted	-1.376	0.072	(-1.520,-1.232)	–	–	–	–	–	–	–

propensity score can be added to the response model if needed.

We applied our penalized technique to the NSW data. Table 5 presents the results; *Saturated* and *unadjusted* rows refer to the propensity score model including all, and none, of the variables respectively. Figure 3 shows how estimated coefficients vary by changing the tuning parameter. Note that the maximum(minimum) effective number of coefficients (df) corresponds to $\lambda = 0(1)$. The penalized covariate selection technique keeps “Education”, “Hispanic”, “Married Status” and “No degree” as the only significant confounders in the propensity score model. Although we do not have a large number of covariates in this example, our variable selection technique results in more than 27% reduction in variance compared with the saturated model.

7 Discussion

We use a double robust regression model for estimating the treatment effect by incorporating the propensity score in the conditional expectation of the response, and establish a penalized likelihood estimator which satisfies the Oracle property under standard assumptions. The proposed technique involves the treatment and outcome models simultaneously, and does not ignore covariates which are just related to the outcome. Note that running a penalized regression on treatment or outcome models separately may ignore important confounders which are just related to the outcome or just to the treatment assignment models. Our covariate selection approach reduces the variance while attaining no residual confounding. The selected covariates may be used in other causal techniques as well as the proposed regression method. Note that the oracle property of the penalized regression estimators is a pointwise asymptotic feature and does not necessarily hold for all the points in the parameter space (Leeb and Pötscher (2008)). However in this paper, in our proposed penalized estimator, we do not penalize the causal effect. As such, our parameter space is partially sparse. In this case, our causal effect estimator $\hat{\beta}_1$ is uniformly consistent and $\hat{\beta}_1 - \beta_1 = O_p(\sqrt{n})$ (Leeb and Pötscher (2005)).

Our technique can also be used to estimate the direct treatment effect in a longitudinal setting; however, if the total effect is the parameter of interest, other causal adjustment techniques such as *Inverse Probability of Treatment Weighting* need to be used. One of the drawbacks of

propensity score regression-based models is that any non-linear term in the true response model has to be added in the propensity score model to ensure the consistency of the treatment effect estimator which is not easily detectable in some cases.

Appendix I

In this appendix, we prove the results stated in the text. Lemma 1 is an adaptation of the results given by Ibragimov and Has' Minskii (1981), page 36. It holds under the following assumptions:

- C1. The parameter space Ξ is a bounded open set in $\mathcal{R}^p$.
- C2. The joint penalized density $f_p(z; \eta)$, where $z_i = (y_i, d_i, x_i)$ is a continuous function of η on Ξ^c for almost all $z \in \mathcal{Z}$, where $\mathcal{Z}$ and Ξ^c represent the sample space $(y_i, d_i, \mathbf{x}_i)$ and the closure of Ξ respectively.
- C3. For all $\eta \in \Xi$ and all $\gamma > 0$,

$$\kappa_{\eta,n}(\gamma) = \inf_{\|\eta - \eta^*\| > \gamma} r^2(\eta, \eta^*),$$

where

$$r^2(\eta, \eta^*) = \int_{\mathcal{Z}} \{f^{1/2}(z; \eta) - f^{1/2}(z; \eta^*)\}^2 d\tau$$

- C4. For $\eta \in \Xi^c$

$$w_\eta(\delta) = \left[\int_{\mathcal{Z}} \sup_{\|h\| \leq \delta} \{f^{1/2}(z; \eta) - f^{1/2}(z; \eta + h)\}^2 d\tau \right] \rightarrow 0 \text{ as } \delta \rightarrow 0.$$

- C5. $f(z; \eta)$ has finite Fisher information at each $\eta \in \Xi$.

Assumption C3 is the identifiability condition, essentially requiring the distance between the averaged densities over the response and the covariates for two different values of the parameters η and η^* be positive. Assumption C4 is referred to as the smoothness condition; it states that the distance of the joint densities over η and η^* when $\eta \rightarrow \eta^*$ should approach zero as the sample size goes to infinity.

Lemma 1 *Suppose assumption C1-C4 are satisfied. Then for any fixed $\eta \in \Xi$*

$$\mathbb{E}_\eta \left[\sup_{\Gamma} \prod_{i=1}^n \frac{f_p^{1/2}(z_i; \eta + b)}{f_p^{1/2}(z_i; \eta)} \right] \leq \exp \left\{ -\frac{n}{2} \left[\kappa_{\eta,n}(\frac{\gamma}{2}) - 2w_{\eta+b_0,n}(\delta) + p_{\lambda_n}(|\eta + b^m|) - p_{\lambda_n}(|\eta|) \right] \right\}, \quad (8)$$

where Γ is the sphere of radius δ , situated in its entirety in the region $\|b\| > \gamma/2$, b_0 is the center of Γ and $\inf_{\Gamma} p_{\lambda_n}(|\eta + b|) = p_{\lambda_n}(|\eta + b^m|)$.

Proof The proof follows from the proof of Theorem 1.4.3 in Ibragimov and Has' Minskii (1981).
Let

$$R_n(b) = \prod_{i=1}^n \frac{f_p(z_i; \eta + b)}{f_p(z_i; \eta)} = \prod_{i=1}^n \frac{f(z_i; \eta + b) e^{-p_{\lambda_n}(|\eta + b|)}}{f(z_i; \eta) e^{-p_{\lambda_n}(|\eta|)}}.$$

We want to find an upper bound for the expectation $\mathbb{E}_\eta \left[\sup_{\Gamma} R_n^{1/2}(b) \right]$, where Γ is the sphere of a radius δ situated in its entirety in the region $||b|| > \frac{1}{2}\gamma$. If b_0 is the center of Γ , then

$$\begin{aligned} \sup_{\Gamma} R_n^{1/2}(b) &= \sup_{\Gamma} \prod_{i=1}^n \left(\frac{f(z_i; \eta + b) e^{-p_{\lambda_n}(|\eta + b|)}}{f(z_i; \eta) e^{-p_{\lambda_n}(|\eta|)}} \right)^{1/2} \leq \prod_{i=1}^n \sup_{\Gamma} e^{\frac{1}{2}p_{\lambda_n}(|\eta|) - \frac{1}{2}p_{\lambda_n}(|\eta + b|)} \\ &\quad \prod_{i=1}^n f^{-1/2}(z_i; \eta) \left(f^{1/2}(z_i; \eta + b_0) + \sup_{|h| \leq \delta} |f^{1/2}(z_i; \eta + b_0 + h) - f^{1/2}(z_i; \eta + b_0)| \right). \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{E}_\beta \left[\sup_{\Gamma} R_n^{1/2}(b) \right] &\leq \prod_{i=1}^n \sup_{\Gamma} e^{\frac{1}{2}p_{\lambda_n}(|\eta|) - \frac{1}{2}p_{\lambda_n}(|\eta + b|)} \left(\int_{\mathcal{Z}} f^{1/2}(z; \eta) f^{1/2}(z; \eta + b_0) d\tau \right. \\ &\quad \left. + \int_{\mathcal{Z}} \sup_{|h| \leq \delta} f^{1/2}(z; \eta) |f^{1/2}(z; \eta + b_0 + h) - f^{1/2}(z; \eta + b_0)| d\tau \right). \end{aligned}$$

We further note that

$$\begin{aligned} \int_{\mathcal{Z}} f^{1/2}(z; \eta) f^{1/2}(z; \eta + b_0) d\tau &= \frac{1}{2} \left(\int_{\mathcal{Z}} f(z; \eta) d\tau + \int_{\mathcal{Z}} f(z; \eta + b_0) d\tau \right. \\ &\quad \left. - \int_{\mathcal{Z}} [f^{1/2}(z; \eta) - f^{1/2}(z; \eta + b_0)]^2 d\tau \right) \\ &\leq 1 - \frac{1}{2} r^2(\eta + b_0) \leq 1 - \frac{\kappa_\eta(\frac{\gamma}{2})}{2} \end{aligned} \tag{9}$$

and

$$\int \sup_{|h| \leq \delta} f^{1/2}(z; \eta) |f^{1/2}(z; \eta + b_0 + h) - f^{1/2}(z; \eta + b_0)| d\tau \leq w_{b_0}(\delta). \tag{10}$$

The last inequality follows from the Cauchy-Schwarz inequality. Finally,

$$\mathbb{E}_\beta \left[\sup_{\Gamma} R_n^{1/2}(b) \right] \leq \exp \left\{ -\frac{n}{2} \left[\kappa_\eta\left(\frac{\gamma}{2}\right) - 2w_{b_0}(\delta) + p_{\lambda_n}(|\eta + b^m|) - p_{\lambda_n}(|\eta|) \right] \right\}$$

where $\sup_{\Gamma} e^{-p_{\lambda_n}(|\eta + b|)} = e^{-p_{\lambda_n}(|\eta + b^m|)}$, using the inequality $1 + a \leq e^a$. $\blacksquare$

Lemma 2 Let $Z_1, \dots, Z_n$ be independent and identically distributed with a density $f(Z, \eta)$ that

satisfies the conditions of Lemma 1. If the penalty function satisfies P3, then

$$R_n(\eta_2) = \prod_{i=1}^n \left(\frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \eta_1, 0)} \right) < 1. \quad (11)$$

Proof $R_n(\eta_2)$ can be written as

$$\prod_{i=1}^n \left(\frac{f(z_i; \eta_1, \eta_2) e^{-\sum_{j=s}^p p_{\lambda_n}(|\eta_j|)}}{f(z_i; \eta_1, 0)} \right)$$

By theorem 1.1 in Chapter II of Ibragimov and Has' Minskii (1981), it can be written as

$$R_n(\eta_2) = \exp \left\{ \left[\sum_{i=1}^n \frac{\partial \ln f_p(z_i; \eta_1, 0)}{\partial \eta_2} \right] \|\eta_2\| - n \sum_{j=s}^p p'_{\lambda_n}(|\eta_j|) - \frac{1}{2} \eta_2 I(\eta_1, 0) \eta_2 + \psi_n(\eta_2) \right\},$$

where $p(|\psi_n(\eta_2)| > \epsilon) \rightarrow 0$. Since $\sum_{i=1}^n \partial \ln f(z_i; \eta_1, 0) / \partial \eta_2 = O_p(\sqrt{n})$, equivalent to the condition P3, the desired inequality holds if

$$\sqrt{n} \sum_{j=s}^p p'_{\lambda_n}(|\eta_j|) > \|\eta_2\| = O_p(1) \quad \blacksquare$$

Proof of Theorem 1: For fixed $\gamma > 0$, the exterior of the sphere $\|b\| \leq \gamma$ can be covered by N spheres Γ_k , $k = 1, \dots, N$ of radius δ with centers b_k . The small value δ is chosen such that all the N spheres are located in the $\|b\| > \gamma/2$ and also $2w_{b_k} \leq \kappa_\eta(\gamma/2)/2$. In view of the inequality (8), we have

$$\begin{aligned} P(|\hat{\eta}_n - \eta_0| > \gamma) &\leq \sum_{k=1}^N P(|\hat{\eta}_n - \eta_0| \in \Gamma_k) \leq \sum_{k=1}^N P(\sup_{\Gamma} R_n(b_k) \geq R(0)) \\ &\leq \sum_{k=1}^N \exp \left\{ -\frac{n}{2} \left[\kappa_\eta\left(\frac{\gamma}{2}\right) - 2w_{b_k}(\delta) + p_{\lambda_n}(|\eta + b^m|) - p_{\lambda_n}(|\eta|) \right] \right\} \\ (\text{Since } p_{\lambda_n}(0) = 0) &\leq N \exp \left\{ -\frac{n}{2} \left[\frac{1}{2} \kappa_\eta\left(\frac{\gamma}{2}\right) + \sum_{j=1}^s p_{\lambda_n}(|\eta_j + b_j^m|) - p_{\lambda_n}(|\eta_j|) \right] \right\}, \end{aligned}$$

where the second inequality follows from lemma 1 and Markov's inequality. We need to show that $\kappa_\eta(\frac{\gamma}{2})/4$ dominates $\sum_{j=s}^p p_{\lambda_n}(|\eta_j + b_j^m|) - p_{\lambda_n}(|\eta_j|)$. Using Taylor's expansion, we have

$$\begin{aligned} \sum_{j=1}^s p_{\lambda_n}(|\eta_j + b_j^m|) - p_{\lambda_n}(|\eta_j|) &= \sum_{j=1}^s p'_{\lambda_n}(\eta_j) \text{sign}(\eta_j) b_j^m + \frac{1}{2} p_{\lambda_n}(\eta_j) (b_j^m)^2 \\ &\leq s \max_{\eta_j \neq 0} \{p'_{\lambda_n}(|\eta_j|)\} \|b\| + \frac{s}{2} \max_{\eta_j \neq 0} \{p''_{\lambda_n}(|\eta_j|)\} \|b\|^2 \\ &\leq s\gamma \max_{\eta_j \neq 0} \{p'_{\lambda_n}(|\eta_j|)\} + \frac{s\gamma^2}{2} \max_{\eta_j \neq 0} \{p''_{\lambda_n}(|\eta_j|)\} \end{aligned} \quad (12)$$

By choosing λ_n such that condition P2 holds, $\kappa_\eta(\frac{\gamma}{2})/4$ dominates the RHS of (12). Thus,

$$P(|\hat{\eta}_n - \eta_0| > \gamma) \leq N \exp \left\{ -\frac{n}{4} \kappa_\eta\left(\frac{\gamma}{2}\right) \right\},$$

and hence we have strong consistency, as

$$P\left(\bigcup_{m=n}^{\infty} |\hat{\eta}_{2m}| > \gamma\right) \leq \frac{N \exp \left\{ -\frac{n}{4} \kappa_\eta\left(\frac{\gamma}{2}\right) \right\}}{1 - \exp \left\{ -\frac{1}{4} \kappa_\eta\left(\frac{\gamma}{2}\right) \right\}} \rightarrow 0 \text{ as } n \rightarrow \infty. \quad \blacksquare$$

Proof of Theorem 2: Consider $\eta_t = (\eta_{t1}, 0)$ and partition $\eta = (\eta_1, \eta_2)$. We need to show that in the neighbourhood $\|\eta - \eta_t\| < O(h_n)$ where $h_n \rightarrow 0$ as $n \rightarrow \infty$,

$$\prod_{i=1}^n \frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \hat{\eta}_1, 0)} < 1.$$

It can be written as

$$\prod_{i=1}^n \frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \hat{\eta}_1, 0)} = \prod_{i=1}^n \left(\frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \eta_1, 0)} \right) \left(\frac{f_p(z_i; \eta_1, 0)}{f_p(z_i; \hat{\eta}_1, 0)} \right) < \prod_{i=1}^n \frac{f_p(z_i; \eta_1, \eta_2)}{f_p(z_i; \eta_1, 0)} < 1.$$

By the result of Lemma 2, the last inequality holds with probability one as $n \rightarrow \infty$. $\blacksquare$

References

- Antoniadis, A. (1997), "Wavelets in statistics: a review," *Statistical Methods and Applications*, 6(2), 97–130.
- Bang, H., and Robins, J. (2005), "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61(4), 962–972.
- Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and adaptive estimation for semiparametric models*, Johns Hopkins Series in the Mathematical Sciences, Baltimore, MD: Johns Hopkins University Press.

- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Sturmer, T. (2006a), “Variable selection for propensity score models,” *American Journal of Epidemiology*, 163(12), 1149–1156.
- Brookhart, M. A., and van der Laan, M. J. (2006b), “A semiparametric model selection criterion with applications to the marginal structural model,” *Computational Statistics & Data Analysis*, 50(2), 475–498.
- Chamberlain, G. (1987), “Asymptotic Efficiency in Estimation with Conditional Moment Restrictions,” *Journal of Econometrics*, 34, 305–334.
- Crainiceanu, C., Dominici, F., and Parmigiani, G. (2008), “Adjustment uncertainty in effect estimation,” *Biometrika*, 95(3), 635.
- Davidian, M., Tsiatis, A., and Leon, S. (2005), “Semiparametric estimation of treatment effect in a pretest–posttest study with missing data,” *Statistical Science*, 20(3), 261.
- 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), 1053–1062.
- Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986), “Semiparametric estimates of the relation between weather and electricity sales,” *Journal of the American Statistical Association*, 81(394), 310–320.
- Fan, J., and Li, R. (2001), “Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties,” *Journal of the American Statistical Association*, 96(456), 1348–1261.
- Greenland, S. (2008), “Invited commentary: variable selection versus shrinkage in the control of multiple confounders,” *American Journal of Epidemiology*, 167(5), 523.
- Gruber, S., and van der Laan, M. J. (2011), “Commentary: Consistent causal effect estimation under dual misspecification, and implications for confounder selection procedures,” *Statistical Methods*, .
- Hafeman, D. M., and VanderWeele, T. J. (2010), “Alternative assumptions for the identification of direct and indirect effects,” *Epidemiology*, 21(3), 1531–5487.
- Huber, P., and Ronchetti, E. (1981), *Robust statistics*, Vol. 1 Wiley.
- Ibragimov, I. A., and Has’ Minskii, R. Z. (1981), *Statistical Estimation–Asymptotic Theory* Springer.

- Kang, J., and Schafer, J. (2007), “Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data,” *Statistical Science*, 22(4), 523–539.
- LaLonde, R. J. (1986), “Evaluating the econometric evaluations of training programs with experimental data,” *The American Economic Review*, 76(4), 604–620.
- Leeb, H., and Pötscher, B. (2005), “Model selection and inference: Facts and fiction,” *Econometric Theory*, 21(1), 21–59.
- Leeb, H., and Pötscher, B. (2008), “Sparse estimators and the oracle property, or the return of Hodges’ estimator,” *Journal of Econometrics*, 142(1), 201–211.
- Little, R. J. A., and An, H. (2004), “Robust likelihood-based analysis of multivariate data with missing values,” *Statistica Sinica*, 14, 949–968.
- Neugebauer, R., and van der Laan, M. (2005), “Why prefer double robust estimators in causal inference?,” *Journal of Statistical Planning and Inference*, 129(1-2), 405–426.
- Petersen, M. L., Sinisi, S. E., and van der Laan, M. J. (2006), “Estimation of direct causal effects,” *Epidemiology*, 17(3), 276–284.
- Porter, K., Gruber, S., van der Laan, M., and Sekhon, J. (2011), “The relative performance of targeted maximum likelihood estimators,” *UC Berkeley Division of Biostatistics Working Paper Series*, p. 279.
- Robins, J. (1999), Robust estimation in sequentially ignorable missing data and causal inference models,, in *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, Vol. 6.
- Robins, J. M., and Greenland, S. (1992), “Identifiability and exchangeability for direct and indirect effects,” *Epidemiology*, 3(2), 143–155.
- Robins, J. M., Mark, S. D., and Newey, W. K. (1992), “Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders,” *Biometrics*, 48(2), 479–495.
- Robins, J. M., Richardson, T. S., and Spirtes, P. (2010), “On Identification and Inference for Direct Effects,” *Epidemiology*, p. In Press.
- Robinson, P. (1988), “Root-N-consistent semiparametric regression,” *Econometrica: Journal of the Econometric Society*, pp. 931–954.
- Rosenbaum, P. (2010), “Causal Inference in Randomized Experiments,” *Design of Observational Studies*, pp. 21–63.

- Rosenbaum, P. R., and Rubin, D. B. (1983), “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, 70(1), 41–55.
- Rubin, D. B. (2008), “For objective causal inference, design trumps analysis,” *The Annals of Applied Statistics*, 2(3), 808–840.
- Schafer, J. L., and Kang, J. D. Y. (2005), “Discussion of “Semi-parametric estimation of treatment effect in a pretest–posttest study with missing data” by M. Davidian et al.,” *Statistical Science*, 20, 292–295.
- Schisterman, E. F., Cole, S., and Platt, R. W. (2009), “Overadjustment bias and unnecessary adjustment in epidemiologic studies,” *Epidemiology*, 20(4), 488.
- Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society, Series B*, 58(1), 267–288.
- Tsiatis, A. A. (2006), *Semiparametric theory and missing data* Springer Verlag.
- van der Laan, M., and Gruber, S. (2010), “Collaborative double robust targeted maximum likelihood estimation,” *The international journal of biostatistics*, 6(1), 17.
- van der Laan, M., and Robins, J. (2003), *Unified methods for censored longitudinal data and causality* Springer Verlag.
- Vansteelandt, S., Bekaert, M., and Claeskens, G. (2010), “On model selection and model misspecification in causal inference,” *Statistical Methods in Medical Research*, pp. 1477–1534.
- Zou, H. (2006), “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, 101(476), 1418–1429.
- Zou, H., and Hastie, T. (2005), “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society, Series B*, 67(2), 301–320.

Appendix II: Penalty Functions

- LASSO (Tibshirani (1996)):

$$p_\lambda(|\beta_j|) = \lambda|\beta_j|$$

- SCAD (Fan and Li (2001)): for $a > 2$

$$p_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta_j| < \lambda \\ -\frac{(|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |\beta_j| < a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta_j| > a\lambda \end{cases}$$

so that the first derivative takes the form

$$p'_\lambda(|\beta_j|) = \lambda \left\{ \mathbb{1}(|\beta_j| < \lambda) + \frac{(a\lambda - |\beta_j|)_+}{(a-1)\lambda} \mathbb{1}(|\beta_j| > \lambda) \right\}$$

- EN (Zou and Hastie (2005)):

$$p_\lambda(|\beta_j|) = \lambda_1|\beta_j| + \lambda_2\beta_j^2$$

- HARD (Antoniadis (1997)):

$$p_\lambda(|\beta_j|) = \lambda^2 - (|\beta_j| - \lambda)^2 \mathbb{1}(|\beta_j| < \lambda)$$

Figure 3: The NSW data: The left and right plots are the estimated coefficients using the LASSO and SCAD, respectively. The horizontal axis represents the effective number of coefficients (df).

