Doubly robust tests of exposure effects under high-dimensional confounding

Oliver Dukes,
Department of Applied Mathematics, Computer Science and Statistics
Ghent University, Belgium

Vahe Avagyan
Mathematical and Statistical Methods Group
Wageningen University and Research, The Netherlands

and Stijn Vansteelandt
Department of Applied Mathematics, Computer Science and Statistics
Ghent University, Belgium
and Department of Medical Statistics
London School of Hygiene and Tropical Medicine, U.K.

email: oliver.dukes@ugent.be

Abstract

After variable selection, standard inferential procedures for regression parameters may not be *uniformly valid*; there is no finite-sample size at which a standard test is guaranteed to approximately attain its nominal size. This problem is exacerbated in high-dimensional settings, where variable selection becomes unavoidable. This has prompted a flurry of activity in developing uniformly valid hypothesis tests for a low-dimensional regression parameter (e.g. the causal effect of an exposure A on an outcome Y) in high-dimensional models. So far there has been limited focus on model misspecification, although this is inevitable in high-dimensional settings. We propose tests of the null that are uniformly valid under sparsity conditions weaker than those typically invoked in the literature, assuming working models for the exposure and outcome are both correctly specified. When one of the models is misspecified, by amending the procedure for estimating the nuisance parameters, our tests continue to be valid; hence they are *doubly robust*. Our proposals are straightforward to implement using existing software for penalized maximum likelihood estimation and do not require sample-splitting. We illustrate them in simulations and an analysis of data obtained from the Ghent University Intensive Care Unit.

Key words: Causal inference; doubly-robust estimation; high-dimensional inference; post-selection inference.

1 Introduction

We will consider a study design which collects i.i.d. data on an outcome Y, an exposure of interest A and a vector of covariates L, some of which may confound the relationship between A and Y. A common means of assessing the effect of A on Y is to fit a regression model, adjusted for A and the covariates; the estimate of the coefficient for A is then used to obtain inference on the exposure effect. In practice, there is often little prior knowledge on which variables in a given data set are confounders, and furthermore how one should model the association between these confounders and outcome. Hence, data-adaptive procedures are typically employed in order to select the variables to adjust for and/or choose a model for their dependence on Y. In particular, data-adaptive model selection becomes increasingly necessary when the dimension of L is close to or greater than the number of observations.

However, obtaining hypothesis tests and confidence intervals that approximately enjoy their nominal size/coverage after model selection is challenging. The estimate of the effect of A obtained directly via regularization techniques - e.g. using a penalized maximum likelihood estimator (PMLE) - will inherit a so-called regularization bias. Furthermore, the moderate-sample distribution of this estimator will typically be non-normal (Leeb and Pötscher, 2005). This is because convergence to the asymptotic normal distribution is not *uniform* with respect to the parameters indexing the model for Y. Therefore, there exists no finite n such that normal-based tests and intervals are guaranteed to perform well. This issue applies more generally to post-regularization estimators (where the model selected via regularization is refitted using the chosen covariates) and routinely-used stepwise variable selection strategies. Standard inferential procedures also ignore the additional uncertainty created during the model-selection process.

In the mathematical statistics literature, this has prompted the development of methods to obtain *uniformly valid* inference for a low-dimensional regression parameter in a high-dimensional model. Initial focus was given to tests and confidence intervals for a coefficient in a regression model fit using the Lasso (Belloni et al., 2014; van de Geer et al., 2014; Zhang and Zhang, 2014); after which attention has turned to more general data-adaptive methods (Ning and Liu, 2017; Chernozhukov et al., 2018). The key insight has been that one should perform selection based on an additional working model for the association between A and L (in addition to Y with A and L). The majority of the recent proposals rely on strong assumptions on sparsity e.g. the number of relevant covariates in L, which needs to be much smaller than the square root of the sample size n. In biostatistics, there were earlier developments in Targeted Maximum Likelihood Estimation (TMLE), where data-adaptive methods are incorporated into the estimation of causal effects (van der Laan and Rose, 2011). Much of the theory on TMLE is developed under Donsker conditions from the empirical process literature (but not all of it e.g. Zheng and van der Laan (2011)). These conditions are usually too restrictive in settings where the dimension of covariates is allowed to grow with sample size.

In this work, we describe how to obtain uniformly valid tests of the causal null hypothesis for a regression parameter in a high-dimensional Generalized Linear Model (GLM). Our tests require postulation of working models for the conditional mean of the exposure and the outcome given covariates. We will work under parametric models, as this is what is typically done in practice. First, we describe a procedure for estimating the nuisance parameters which yields a valid test so long as all working models are correct. However, given that regularization/model selection is required because we do not know the true models to start with, some degree of misspecification is likely. This is felt most acutely when the number of covariates in the data set is very large relative to the number of observations. We then show how to amend the earlier procedure for nuisance parameter estimation, so that the test statistic will converge uniformly to a limiting normal distribution if either working model is correct. Hence the test can be made *uniformly doubly robust*. This is in contrast to several existing proposals, which give doubly robust estimators but not inference (van der Laan and Rose, 2011; Farrell, 2015; Chernozhukov et al., 2018; Shah and Peters, 2019). Furthermore, we will show that when both working models are correct, then in certain cases the test will continue to attain its nominal size under sparsity conditions weaker than those invoked in the literature. Our test statistic is straightforward to construct, and all procedures for estimating the nuisance parameters can be performed using existing penalized regression software. Sample-splitting is not required, which makes our proposal much simpler to implement and less affected by regularization bias in moderate sample sizes.

The paper is organized as follows: in Section 2, we state the null hypothesis we are interested in testing and describe issues with obtaining valid inference in the high-dimensional setting. Section 3 presents the score test statistic. In Sections 4 and 5 we describe specific procedures for estimating the nuisance parameters, first when all working models are correct and then under misspecification. We also discuss the asymptotic properties of the various methods. We illustrate the methods via simulation studies in Section 6 and an analysis of data from the Ghent University Intensive Care Unit in Section 7, where we consider the effect of a change in glycemia level on mortality in critically ill patients.

2 Motivation

We consider a test of the null hypothesis H_0 that Y is independent of A within strata defined by L, or

$$H_0: \quad Y \perp\!\!\!\perp A | L. \tag{1}$$

We will let the exposure A be binary e.g. it is coded as 1 if an individual undergoes a particular medical treatment and 0 otherwise; extensions to more general exposures will be discussed later in the paper. Then given the standard structural conditions in the causal inference literature, in particular that L is sufficient to adjust for confounding, the null hypothesis also expresses the absence of a causal effect of A on Y, conditional on L.

Simple methods for testing the null hypothesis of conditional independence are available via the regression framework. Standard score tests of H_0 require estimation of $E(Y|L)$, since under the null, $E(Y|A=a,L) = E(Y|L)$ (full conditional independence implies mean conditional independence). In realistic settings where L has multiple continuous components, non-parametric estimators of this functional may perform poorly. A common strategy is to instead postulate a parametric regression model $\mathcal{B}$ for the mean of Y conditional on the covariates:

$$E(Y|L) = m(L;\beta),$$

where $m(L;\beta)$ is a known function smooth in an unknown finite-dimensional parameter β. Then, via maximum likelihood estimation of GLMs, under the pre-specified model $\mathcal{B}$ one can obtain a consistent and uniformly asymptotically normal (UAN) score statistic for testing H_0 (where uniformity is with respect to β). This means that there exists a finite-sample size such that for any value of β within the parameter space, the test statistic will be approximately normally distributed. We note that although this is a valid test of H_0, it is only consistent in the direction of alternatives that obey some form of mean conditional independence between Y and A; the same holds for subsequent tests described in this paper.

Unfortunately, this standard methodology does not straightforwardly extend to high-dimensional settings. In low-dimensional settings, one can perform a score test of the causal null H_0 based on the asymptotic distribution of an unbiased (unscaled) score test statistic $U(\beta)$; for likelihood estimation of canonical GLMs, $U(\beta) = A\{Y - m(L;\beta)\}$. Then let $\tilde{\beta}$ denote an estimate of β obtained either directly via some regularization method or after model selection. Following a Taylor expansion,

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(\tilde{\beta}) =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(\beta) + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_i(\beta)}{\partial\beta}\sqrt{n}(\tilde{\beta}-\beta) \\ &+ O_P(\sqrt{n}||\tilde{\beta}-\beta||_2^2)\end{aligned} \tag{2}$$

where $||.||_2$ denotes the Euclidean norm. For fixed β, by appealing to the oracle properties of $\tilde{\beta}$ it may be argued that the right hand side of (2) is asymptotically normal. Indeed, assuming that $\tilde{\beta}$ converges sufficiently quickly, the remainder term $O_P(\sqrt{n}||\tilde{\beta}-\beta||_2^2)$ converges to zero. But in general this does not prevent the existence of converging sequences β_n for which $\sqrt{n}(\tilde{\beta}-\beta_n)$ and thus the test statistic has a complex, non-normal distribution. One root cause of this is the discrete nature of many data-adaptive methods e.g. stepwise selection; in some samples $\tilde{\beta}$ will be forced to zero whereas in others it will be allowed to take on its estimated value. This discrete behavior persists with increasing sample size under certain sequences β_n. The convergence of the resulting score test statistic to the limiting standard normal is hence not uniform over the parameter space (Leeb and Pötscher, 2005; Dukes and Vansteelandt, 2019). This is troubling, as one wishes there to be a finite n where the normal approximation is guaranteed to hold well, regardless of the (unknown) true values of the nuisance parameters, in order to guarantee that the procedure will work well in finite samples.

3 A uniformly valid test of the causal null hypothesis

We introduce in this section the statistic we will use for testing H_0 in a high-dimensional setting. Let us now formally define γ to be the nuisance parameter indexing the parametric model $\mathcal{A}$ for the conditional mean of A given L:

$$E(A|L) = \pi(L;\gamma),$$

where $\pi(L;\gamma)$ is a known function smooth in an unknown finite-dimensional parameter γ; the conditional mean $E(A|L)$ is known as the *propensity score* for binary A. Our analysis is based on the score function

$$U(\eta) \equiv \{A - \pi(L;\gamma)\}\{Y - m(L;\beta)\}.$$

where $\eta = (\gamma^T, \beta^T)^T$. This will require initial estimates of γ and β under working models $\mathcal{A}$ and $\mathcal{B}$ respectively. It is natural to model the dependence of A on L using a logistic regression e.g. $\pi(L;\gamma) = \text{expit}(\gamma^T L)$; if A were continuous, one might postulate a linear or log-linear model instead and the proposal can then be easily adapted. The form that model $\mathcal{B}$ takes will depend on the nature of the outcome. If Y is continuous and unconstrained, one might postulate a linear model e.g. $m(L;\beta) = \beta^T L$.

One can then construct a test statistic

$$T_n = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n U_i(\hat{\eta})}{\sqrt{\frac{1}{n}\sum_{i=1}^n \left\{U_i(\hat{\eta}) - \bar{U}(\hat{\eta})\right\}^2}}$$

that we will compare to the standard normal distribution. Here, $\bar{U}(\hat{\eta}) = n^{-1}\sum_{i=1}^n U_i(\hat{\eta})$ and $\hat{\eta}$ is an estimate of η; in what follows, we will focus on regularized estimation of this parameter. Note that in evaluating the functions $\pi(L;\gamma)$ and $m(L;\beta)$ at their limiting values, it follows that the mean of $U(\eta)$ under the null is equal to

$$E[\{E(A|L) - \pi(L;\gamma)\}\{E(Y|L) - m(L;\beta)\}]$$

which equals zero if either model $\mathcal{A}$ or $\mathcal{B}$ is correct a.k.a under the union model $\mathcal{A} \cup \mathcal{B}$. Hence we refer to the score $U(\eta)$ as doubly robust.

4 Estimation of η when all models are correct

4.1 Proposal

We will first consider data generating processes where both models $\mathcal{A}$ or model $\mathcal{B}$ are correctly specified e.g. we will work under the intersection submodel $\mathcal{A} \cap \mathcal{B}$. In high-dimensional settings, under model $\mathcal{A} \cap \mathcal{B}$ one can estimate γ and β as $\hat{\gamma}$ and $\hat{\beta}$ respectively using any sufficiently fast-converging sparse estimator. For example, with binary A and continuous Y and using standard PMLE with a Lasso penalty, $\hat{\gamma}$ and $\hat{\beta}$ can be obtained as:

$$\hat{\gamma} = \arg\min_{\gamma} \frac{1}{n} \sum_{i=1}^{n} [\log\{1 + \exp(\gamma^T L_i)\} - A_i(\gamma^T L_i)] + \lambda_\gamma ||\gamma||_1$$

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \beta^T L_i)^2 + \lambda_\beta ||\beta||_1$$

(Tibshirani, 1996), where $\lambda_\gamma > 0$ and $\lambda_\beta > 0$ are penalty parameters. To keep the notation simple, we have omitted the dependence of $\hat{\gamma}$ and $\hat{\beta}$ on λ_γ and λ_β respectively (as well as on n). Plugging the resulting estimates into T_n will yield a test statistic that under the null follows a standard normal distribution. We note that our Theorem 1 is quite general and covers other penalized m-estimation methods (including variants on the Lasso).

To give some intuition about why one can plug $\hat{\gamma}$ and $\hat{\beta}$ into T_n and yet obtain a UAN test statistic, by repeating the expansion in (2) for the score $U(\hat{\eta})$, we observe the first-order term

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_i(\eta)}{\partial \eta}\sqrt{n}(\hat{\eta}-\eta). \tag{3}$$

We can control this term under the model $\mathcal{A}\cap\mathcal{B}$, since $\partial U(\eta)/\partial\eta$ will then have expectation zero at the limiting value of η. For example, when $\pi(L;\gamma) = \text{expit}(\gamma^T L)$ and $m(L;\beta) = \beta^T L$, then using the law of iterated expectation

$$E\{\partial U(\eta)/\partial\beta\} = E[E\{A-\pi(L;\gamma)|L\}L] = 0 \quad \text{and}$$

$$E\{\partial U(\eta)/\partial\gamma\} = E\left[\pi(L;\gamma)\{1-\pi(L;\gamma)\}E\{Y-m(L;\beta)|L\}L\right] = 0$$

under the null. This property helps to ensure that term (3) is asymptotically negligible, regardless of the complex behavior of $\hat{\eta}$. Such phenomena (in the context of doubly robust estimators) is well-understood when L is low-dimensional (Vermeulen and Vansteelandt, 2015). What is surprising is that it continues to hold in high-dimensional settings, even when non-regular estimators are used for η (Belloni et al., 2016).

We recommend selecting the penalty parameters in practice via cross validation, although there are limited theoretical results available on its validity in this context (Chetverikov et al., 2016), and our inferences assume that these parameters are fixed. The standard conditions are that $\lambda_\gamma = o(\sqrt{\log(p \vee n)/n})$ and $\lambda_\beta = o(\sqrt{\log(p \vee n)/n})$ (where $a \vee b$ denotes the maximum of a and b), which are required for our theoretical results (see Appendix A). In practice, we also recommend refitting both working models; model refitting is typically done in the literature in order to improve finite-sample performance (Belloni et al., 2016; Ning and Liu, 2017). Our theory can allow for this by appealing to results on Post-Lasso estimators (Belloni et al., 2014, 2016). For any vector $a \in \mathbb{R}^p$, let us define its support as $\text{support}(a) = \{j \in \{1, .., p\} : a_j \neq 0\}$; then we refit each model $\mathcal{A}$ and $\mathcal{B}$ using $\text{support}(\gamma) \subseteq \text{support}(\hat{\gamma})$ and $\text{support}(\beta) \subseteq \text{support}(\hat{\beta})$ respectively.

4.2 Asymptotic properties

Let $\mathcal{P}$ be the class of laws that obey the intersection submodel $\mathcal{A}\cap\mathcal{B}$; then we are interested in convergence under a sequence of laws $P_n \in \mathcal{P}$. We will allow for p to increase with n, and for the values of the parameters γ and β to depend on n, and hence also models $\mathcal{A}$ and $\mathcal{B}$ (although the notation with respect to the models will be suppressed). This is done in order to better gain insight into the finite-sample behavior of the test statistic when L is high-dimensional. Let γ_n and β_n be the population values of the parameters indexing models $\mathcal{A}$ and $\mathcal{B}$ respectively. Finally, let $\mathbb{P}_{P_n}()$ denote a probability taken with respect to the local data generating process P_n.

Theorem 1. *Let us define the active set of variables as $S_\gamma = support(\gamma_n)$ and $S_\beta = support(\beta_n)$. Furthermore, let s_γ denote the cardinality $|S_\gamma|$ and likewise $s_\beta = |S_\beta|$, and p denote the length of the vectors γ_n and β_n . Suppose, in addition to Assumptions 1 and 2 in Appendix A, the following sparsity conditions hold:*

(i) $(s_\gamma + s_\beta)\log(p \vee n) = o(n)$

(ii) $s_\gamma s_\beta \log^2(p \vee n) = o(n)$.

Then under H_0 and the intersection submodel, for all estimators satisfying Assumption 2 in Appendix A, we have for any $t \in \mathbb{R}$:

$$\lim_{n\to\infty} \sup_{P_n\in\mathcal{P}} |\mathbb{P}_{P_n}(T_n \leq t) - \Phi(t)| = 0 \tag{4}$$

Remark. The key assumptions required for this result to hold are given in Appendix A. Therein, Assumption 1 contains mild moment conditions, whereas Assumption 2 requires sufficiently fast estimation of $\pi(L;\gamma)$ and $m(L;\beta)$ in the empirical ℓ_2-norm. Condition (i) requires that both $s_\gamma << n$ and $s_\beta << n$; such conditions are quite standard in order to guarantee consistency of the sparse estimators. Condition (ii) implies that we can allow for s_γ to be large if s_β is small, and vice versa. We view this as useful given that in many medical settings, doctors may rely on a limited number of factors when deciding on a patient's treatment. Hence it may even be plausible that the exposure model is 'ultra-sparse' e.g. $s_\gamma << \sqrt{n}$. In contrast, it appears less likely that a model for a clinical outcome (e.g. disease occurrence) can be well approximated by a small number of covariates. These conditions are essentially equivalent to those given in Chernozhukov et al. (2018) and in the Supplementary Appendix of Belloni et al. (2014), where sample-splitting/cross-fitting is used; indeed, we obtain slightly sharper results than Belloni et al. (2014) (who require that $n^{2/r} s \log(p \vee n) = o(n)$ for some $r > 4$ for uniformly valid inference) by focusing on binary exposures.

Remark. In Farrell (2015), uniformly valid inference for the marginal treatment effect is obtained under the stronger assumption that $s_\beta^2 = o(n)$ (ignoring log factors). To obtain uniformly valid estimators and tests based on trading-off assumptions on s_γ and s_β, it turns out to be crucial that first order terms like (3) have expectation zero, conditional only on $(L_i)_{i=1}^n$; in many estimation problems this is not possible, because fitting a model for Y requires adjusting for/conditioning on the exposure, so the estimated coefficients depend on $(A_i)_{i=1}^n$. One way to get round this could be to use sample-splitting (Chernozhukov et al., 2018). However, under the null our score test statistic factorizes into a component involving A and L, and a component involving Y and L. That factorization is much like what one normally achieves via sample-splitting, and hence we can avoid doing this. Indeed, the test statistic also converges to the standard normal under the weaker null hypothesis that $E(Y|A, L) = E(Y|L)$, although this result requires stronger sparsity conditions (namely those of Theorem 2). Stronger conditions would also be required for the above procedure if the working models are refitted using the union of the selected covariates $\text{support}(\hat{\beta}) \cup \text{support}(\hat{\gamma})$ (Belloni et al., 2014), since the variables selected in the model for $E(Y|L)$ depend additionally on the data $(A_i)_{i=1}^n$ (Farrell, 2015).

Remark. Although we have focused on sparse estimators, Theorem 1 should hold for more general machine learning methods for estimating $\pi(L;\gamma)$ and $m(L;\beta)$. The key conditions required are that there is consistency in the prediction error, and shrinkage in the product of the ℓ_2 norms of the errors as $o_p(n^{-1/2})$; see Appendix A and Chernozhukov et al. (2018) for further details.

5 Estimation of η under model misspecification

5.1 Proposal

Under the union model $\mathcal{A} \cup \mathcal{B}$, plugging in an arbitrary high-quality sparse estimator $\hat{\eta}$ of η into T_n will not generally lead to the test statistic converging uniformly to the standard normal. Hence although the score is doubly robust, plugging in $\hat{\eta}$ will not yield *uniformly valid, doubly robust inference.* This can be seen by replicating the Taylor expansion in (2) for the score $U(\hat{\eta})$; the gradient $\partial U(\eta)/\partial \eta$ is no longer guaranteed to be mean zero and one would generally need to approximate $\sqrt{n}(\hat{\eta} - \eta)$ to assess the variability in the score function under the union model. However, as previously discussed, approximating this term well is generally not possible in the high-dimensional setting.

We will handle the problematic term (3) by using the gradient $\partial U(\eta)/\partial \eta$ in order to estimate η, so as to ensure that $n^{-1}\sum_{i=1}^n \partial U_i(\eta)/\partial \eta$ is approximately equal to zero at the estimator of the nuisance parameter. This leaves (aside from the remainder) only the score function $U(\eta)$, which we will show is UAN. Specifically, one can estimate η by solving the following subgradient estimating equations:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \beta} U_i(\hat{\eta}_{BR}) + \lambda_\gamma g(\hat{\gamma}_{BR}) \tag{5}$$

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \gamma} U_i(\hat{\eta}_{BR}) + \lambda_\beta g(\hat{\beta}_{BR}) \tag{6}$$

Here, $g(a)$ denotes a vector of elements $g(a_j)$, where $g(a_j) = \text{sign}(a_j)$ for $j = 1, ..., p$ if $a_j \neq 0$ and $g(a_j) \in [-1, 1]$ otherwise. The penalty term in (5) corresponds to the subgradient of the ℓ_1 norm $||\gamma||_1$ with respect to γ and likewise for the penalty in (6); hence our procedure amounts to ℓ_1-penalized m-estimation. Whilst our procedure requires that the initial working models $\mathcal{A}$ and $\mathcal{B}$ are of the same dimension, using a Lasso penalty will tend to return nuisance parameter estimates with different numbers of non-zero components. We again recommend refitting each working model using the covariates selected via penalization. The above procedure extends the bias-reduced doubly robust estimation methodology of Vermeulen and Vansteelandt (2015) to incorporate penalization; we thus use $\hat{\gamma}_{BR}$ and $\hat{\beta}_{BR}$ to refer to the resulting estimators of γ and β respectively.

5.1.1 Example 1: continuous outcome

Returning to the example of Section 3, we might postulate a linear model for the outcome and a logistic model for the exposure. In this case, the solutions to the equations (5) and (6) specify the optima of the following convex optimization problems:

$$\hat{\gamma}_{BR} = \arg\min_{\gamma} \frac{1}{n}\sum_{i=1}^{n} \log\{1 + \exp(\gamma^T L_i)\} - A_i(\gamma^T L_i) + \lambda_\gamma ||\gamma||_1 \tag{7}$$

$$\hat{\beta}_{BR} = \arg\min_{\beta} \frac{1}{2n}\sum_{i=1}^{n} [\text{expit}(\hat{\gamma}_{BR}^T L_i)\{1 - \text{expit}(\hat{\gamma}_{BR}^T L_i)\}(Y_i - \beta^T L_i)^2] + \lambda_\beta ||\beta||_1. \tag{8}$$

Hence one can estimate γ by fitting a logistic regression model with a Lasso penalty, and then estimate β by fitting a linear regression model again with a Lasso penalty and weights constructed using the estimates $\hat{\gamma}_{BR}$.

5.1.2 Example 2: binary outcome

A more appropriate working model for the conditional mean of binary Y might be $E(Y|A=0,L) = \text{expit}(\beta^T L)$. Hence we now have two minimization problems

$$\hat{\gamma}_{BR} = \arg\min_{\gamma} \frac{1}{n}\sum_{i=1}^{n}[\text{expit}(\hat{\beta}_{BR}^T L_i)\{1-\text{expit}(\hat{\beta}_{BR}^T L_i)\}][\log\{1+\exp(\gamma^T L_i)\} - A_i(\gamma^T L_i)]$$
$$+ \lambda_\gamma ||\gamma||_1$$
$$\hat{\beta}_{BR} = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}[\text{expit}(\hat{\gamma}_{BR}^T L_i)\{1-\text{expit}(\hat{\gamma}_{BR}^T L_i)\}][\log\{1+\exp(\beta^T L_i)\} - A_i(\beta^T L_i)]$$
$$+ \lambda_\beta ||\beta||_1.$$

An additional complication is then that solving each set of equations requires initial estimates of the other nuisance parameter. There are then two possible approaches one might take; one is to estimate γ and β together by maximizing a joint penalized likelihood. Alternatively, one could use the iterative procedure described in Algorithm 1 in Appendix B, which could be easily adapted for other types of outcome.

Avagyan and Vansteelandt (2017) and Tan (2019) make a closely related proposal in the context of estimation the average treatment effect; however, focusing on hypothesis testing and conditional treatment effects enables some simplifications. Our approach for nuisance parameter estimation is based on weighted ℓ_1-penalized maximum likelihood estimation, so is both easier to implement using existing software and likely to be less computationally demanding in high-dimensional settings. We will also be able to obtain sharper results (in terms of conditions on sparsity) on the theoretical properties of the test statistic; see Section 5.2. When the exposure model is linear, our estimator of the unscaled test statistic $n^{-1/2}\sum_{i=1}^{n} U_i(\hat{\eta}_{BR})$ reduces to the 'decorrelated score' approach of Ning and Liu (2017). This work thus extends the robustness of their score function to the construction of a test that is UAN under the model $\mathcal{A} \cup \mathcal{B}$. Specifically, by allowing for arbitrary exposure and outcome models and scaling the statistic by a 'sandwich estimator' of the variance of $U_i(\hat{\eta}_{BR})$, our test has greater robustness to misspecification than the proposals of Ning and Liu (2017), under equivalent assumptions on s_γ and s_β (see Section 5.2).

5.2 Asymptotic properties

We will now study convergence of T_n under a sequence of laws $P_n \in \mathcal{P}^*$, where $\mathcal{P}^*$ represents a class of laws that obey the union model $\mathcal{A} \cup \mathcal{B}$; hence this class is much larger than that considered in Section 4.2.

Theorem 2. *Suppose, in addition to Assumptions 1, 3 and 4 in Appendix A, the following sparsity condition holds:*

(iii) $(s_\gamma^2 + s_\beta^2)\log^2(p \vee n) = o(n)$.

Then under H_0 and the union model $\mathcal{A} \cup \mathcal{B}$, using estimators $\hat{\gamma}_{BR}$ and $\hat{\beta}_{BR}$, we have for any $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \sup_{P_n \in \mathcal{P}^*} |\mathbb{P}_{P_n}(T_n \leq t) - \Phi(t)| = 0. \tag{9}$$

Remark. This theorem states that under the key ultra-sparsity condition (iii), our proposed test is uniformly doubly robust over the parameter space. This condition entails that the number of non-zero coefficients in models $\mathcal{A}$ and $\mathcal{B}$ are small relative to the square root of the overall sample size; this is much stronger than conditions (i) and (ii). Such an assumption is common however in the growing literature on high-dimensional inference (Belloni et al., 2016; Ning and Liu, 2017), where model misspecification is not generally permitted. Assumptions 3 and 4 require sufficiently rapid estimation of the coefficients γ and β (notably in ℓ_1-norm) as well as the functions $\pi(L;\gamma)$ and $m(L;\beta)$.

An advantage of our proposal is that any *a priori* knowledge on the distribution of the exposure can be easily incorporated into the test statistic. Indeed, note that when $E(A|L)$ is known, then following our proposal, one can estimate β using standard ℓ_1-penalized regression without weights; hence the proposal reduces to the one described in Section 4. This is because there is no gradient with respect to the parameters in model $\mathcal{A}$ and the exposure model is guaranteed to be correct.

Corollary 1. *Suppose that A is randomized, with randomization probability $E(A|L) = \pi(L; \gamma^*)$ for known γ^*. If γ^* is plugged into T_n, one can obtain a uniformly valid test using the weaker sparsity condition $s_\beta \log(p \vee n) = o(n)$ regardless of whether model $\mathcal{B}$ is correctly specified.*

Remark. This corollary of Theorems 1 and 2 states that when we know the randomization probabilities, one can rely merely on this weaker sparsity condition in order to get a valid test, even if model $\mathcal{B}$ is misspecified.

When both working models are correctly specified and the outcome regression model is linear, then we have some additional robustness to violations of sparsity, as the following theorem illustrates:

Theorem 3. *When $m(L; \beta)$ is linear in β, and we restrict ourselves to the class of laws in $\mathcal{P}^*$ that obey the intersection submodel $\mathcal{A} \cap \mathcal{B}$, then (using estimators $\hat{\gamma}_{BR}$ and $\hat{\beta}_{BR}$) the score test statistic converges uniformly as in (9) under H_0, Assumptions 1-7 and the conditions (i) $(s_\gamma + s_\beta) \log(p \vee n) = o(n)$ and*

(iv) $s_\gamma s^* \log^2(p \vee n) = o(n)$

where $s^* = s_\gamma \vee s_\beta$.

Remark. In the context of linear models for Y, our proposal is thus 'sparsity-adaptive', in the sense that when both models are correct, our proposal is valid under conditions similar to those required for Theorem 1. As the example in Section 5.1.1 shows, estimated weights dependent on $(A_i)_{i=1}^n$ are only required in this setting when fitting the outcome model; the proof of Theorem 3 hinges on showing that estimating the weights is of lesser impact than estimating β at fixed weights. For non-linear outcome models, an equivalent result will be difficult to obtain, in light of the fact that fitting the exposure model also requires weights that are dependent on $(Y_i)_{i=1}^n$; however, a general result could be shown using sample-splitting (by estimating the weights in a sample separate to the one used in constructing the test statistic). Nonetheless, this illustrates the trade off between modelling and sparsity conditions; if we wish to obtain inference under the union model then we generally need stronger conditions on s_γ or s_β.

Remark. The sparsity conditions required for Theorem 3 reduce exactly to those of Theorem 1 (and thus Chernozhukov et al. (2018)) when $s_\gamma \leq s_\beta$. However, the converse does not hold; even if $s_\beta = 0$, strong assumptions on s_γ are still required. This reflects that estimation of β could be harmed by poor quality estimates of the weights. We do not see this asymmetry as a serious disadvantage as in many settings, we would expect that model $\mathcal{A}$ is more likely to be sparse.

6 Simulation Study

In this section, we conduct a simulation analysis to compare the performance of the proposed hypothesis test with that of different tests of the causal null hypothesis. In our study, we consider the following tests for the null hypothesis (1):

1. A naïve post-selection approach where a t-test is considered for the null using a

linear regression after the standard post-selection of β based on ℓ_1-penalized linear regression. We study the performance of this approach both when the exposure is forced (i.e., the treatment effect ψ is not penalised) and not forced to be included in the model.

2. Tests based on the 'post-double selection' (PDS) and 'partialling out' (PO) approaches proposed in the econometrics literature by Belloni et al. (2014), where Y is regressed on L and A on L also (both using linear models with a Lasso penalty). PDS then uses a second-stage regression where a linear model for Y is fitted using ordinary least squares, adjusted for A and the union of the covariates selected at the first stage, whereas PO emulates the estimator from Robinson (1988) for the partially linear model, also using refitting. Both approaches were implemented in `R` using the package `hdm`.

3. The procedure described in Section 4, valid under model $\mathcal{A} \cap \mathcal{B}$, where a score test is considered using standard logistic regression and linear regression after the post-selection of parameters γ and β based on ℓ_1-penalized logistic regression and ℓ_1-penalized linear regression, respectively (hereafter, PMLE-DR).

4. The procedure described in Section 5, valid under model misspecification, where a score test is considered for the null using standard logistic regression and weighted linear regression after the post-selection of parameters γ and β based on ℓ_1-penalized logistic regression (7) and ℓ_1-penalized weighted linear regression (8), respectively (hereafter, BR-DR).

Note that all the considered approaches require the selection of penalty parameters. In our simulation study, we use 10-fold cross validation technique to choose the tuning parameters for the naïve approach as well as for PMLE-DR and BR-DR. We obtain λ_γ

and λ_β using R package glmnet through the argument lambda.min; this selects the value which minimises the mean cross-validated error. In the hdm package, PDS and PO and implemented using pre-specified values for the penalty parameters (Chernozhukov et al., 2016); in order to study the impact of using different penalties, we also performed PDS as described in Belloni et al. (2014) using cross-validation instead of the pre-specified values.

In the simulation analysis, we generate n mutually independent vectors $\{(Y_i, A_i, L_i^T)^T\}$, $i = 1, ..., n$. Here, $L_i = (L_{i,1}, ..., L_{i,p})$ is a mean zero multivariate normal covariate with covariance matrix Σ; either $\Sigma = \mathrm{I}_{p\times p}$ (uncorrelated covariates) or $\Sigma = [\sigma_{i,j}]_{1\leq i,j\leq p}$, where $\sigma_{i,j} = 2^{-1-|i-j|}$ (correlated covariates). For simplicity we consider a binary exposure model and linear outcome model. We let for each $i = 1, ..., n$, the dichotomous exposure A_i take on values 0 or 1 with $P(A_i = 1|L_i) \equiv \pi(L_i)$, the outcome Y_i be normally distributed with mean $m(L_i)$ and unit variance, conditional on L_i and A_i. Further, the simulated data are analysed using the following parametric working models: $\pi(L;\gamma) = \text{expit}(\gamma_0 + \sum_{i=1}^{p} \gamma_i L_i)$ and $m(L, \beta) = \beta_0 + \sum_{i=1}^{p} \beta_i L_i$, where $\beta_0 = 1$, $\gamma_0 = 2$. The nuisance parameters $\beta = (\beta_1, ..., \beta_p) \in \mathbb{R}^p$ and $\gamma = (\gamma_1, ..., \gamma_p) \in \mathbb{R}^p$ are defined as $b = \left(2\frac{\log(20)}{n^{1/2}}, 2\frac{\log(19)}{n^{1/2}}, ..., 2\frac{\log(2)}{n^{1/2}}, 0_{20}, ..., 0_{81}, 10\frac{\log(2)}{n^{1/2}}, ..., 10\frac{\log(20)}{n^{1/2}}, 0_{101}, ..., 0_p\right)$, $\beta = 2{\cdot}b{\cdot}(\sum_{i=1}^p b^2)^{-1/2}$, $g = \left(40\frac{\log(20)}{n^{1/2}}, 40\frac{\log(19)}{n^{1/2}}, ..., 40\frac{\log(2)}{n^{1/2}}, 0_{20}, ..., 0_p\right)$, $\gamma = 3\cdot g\cdot(\sum_{i=1}^p g^2)^{-1/2}$ where the subscripts indicate the index (i.e., position) of 0 in the vector. The considered settings for nuisance parameters are challenging in the sense that there are confounders that are strongly predictive of the exposure and weakly predictive of the outcome. Moreover, there are covariates which are moderately predictive of the outcome but are not associated with the exposure. In order to evaluate the impact of model misspecification, we next generate data with the following outcome model: $m(L, \beta) = 1 + \beta^T \left(|L_{.,[1:3]}|; L_{.,[4:p]}\right)$. Finally, for the data generating mechanism described above, we perform 1,000 Monte Carlo runs for $n = 200$ and $p = 200$, $n = 500$ and $p = 500$, $n = 200$ and $p = 100$, and $n = 200$ and $p = 250$.

Tables 1 and 2 show the Type I errors based on 1,000 replications. The simulation results show that the PMLE-DR and BR-DR approaches have rejection rates close to the nominal level of 5%, so long as the outcome model is correctly specified. On the other hand, we observe that even when both models are correctly specified, the naïve approaches provide high rejection rates. Moreover, these rates do not diminish with larger sample size. This poor performance is well aligned with the theory of Leeb and Pötscher (2005). We also observe that the rejection rates of PDS and PO are relatively high. Part of this poor performance appeared to be due to the particular data-driven procedure for selecting the penalty parameters, which led to an insufficient number of covariates being selected. However, even when using cross validation for PDS, there was still a discrepancy between the methods. When the covariates were correlated and the outcome model was incorrect, the PMLE-DR test was mildly anti-conservative relative to the BR-DR test; the fact that PDS (which is not generally doubly robust) performed relatively well in this setting indicates that the type of misspecification considered may not be particularly damaging. In Appendix C, we also consider additional settings under the modified sparsity in the propensity score model, as well as heteroscedasticity; similar results to those in 1 and 2 are seen across settings.

7 Data analysis

Glycemic control in critically ill patients is still the subject of controversy, in terms of the optimal limits in which glucose levels are best kept. In the Leuven II randomised trial (Van den Berghe et al., 2001), strict glycemic control (with the maintenance of glycemia between 80 and 110 milligram per deciliter (mg/dl)) resulted in reduced mortality. Later multi-center studies could not replicate these findings, including the NICE-SUGAR trial (Finfer et al., 2009). Current guidelines usually recommend glycemic control between 140

Table 1: Type I errors at the 5% significance level based on 1,000 replications: $\Sigma = \mathrm{I}_{p\times p}$.

Correct models

Methods	$n = 200$ $p = 200$	$n = 500$ $p = 500$	$n = 200$ $p = 100$	$n = 200$ $p = 250$
Standard naïve (forced)	0.548	0.809	0.291	0.600
Standard naïve (not forced)	0.275	0.555	0.168	0.316
PDS (pre-specified)	0.517	0.761	0.536	0.498
PO (pre-specified)	0.508	0.748	0.519	0.482
PDS (CV)	0.074	0.074	0.068	0.072
PMLE-DR	0.055	0.054	0.070	0.075
BR-DR	0.053	0.069	0.078	0.073

Incorrect outcome model

Methods	$n = 200$ $p = 200$	$n = 500$ $p = 500$	$n = 200$ $p = 100$	$n = 200$ $p = 250$
Standard naïve (forced)	0.368	0.586	0.210	0.425
Standard naïve (not forced)	0.175	0.313	0.115	0.197
PDS (pre-specified)	0.319	0.634	0.345	0.327
PO (pre-specified)	0.317	0.615	0.345	0.309
PDS (CV)	0.073	0.072	0.060	0.070
PMLE-DR	0.056	0.053	0.070	0.059
BR-DR	0.046	0.059	0.081	0.050

and 180mg/dl. In the Ghent University Intensive Care Unit (UZ Ghent ICU) a glycemic protocol is used, targeting values between 80 and 150 mg/dl. In practice, glycemia in patients often falls outside of this range, partly due to a lack of compliance in following the protocol. We sought to investigate the relationship between glycemic control and 30-day mortality, using routinely collected data from the UZ Ghent ICU on a large representative cohort of intensive care patients. Specifically, we aimed to test the null hypothesis of no effect of a change in glycemia level (from <110 to ≥110 mg/dl, and then from ≤150 to >150 mg/dl) at any day of follow-up on death within 30 days from ICU entry. We restricted the analysis to patients that were alive in the ICU for at least 48 hours, thus removing patients who died immediately upon arrival in the ICU.

Table 2: Type I errors at the 5% significance level based on 1,000 replications: $\Sigma = [\sigma_{i,j}]_{1 \leq i,j \leq p}$.

Correct models

	$n = 200$	$n = 500$	$n = 200$	$n = 200$
Methods	$p = 200$	$p = 500$	$p = 100$	$p = 250$
Standard naïve (forced)	0.448	0.387	0.190	0.522
Standard naïve (not forced)	0.178	0.164	0.100	0.193
PDS (pre-specified)	0.143	0.064	0.103	0.131
PO (pre-specified)	0.115	0.064	0.085	0.108
PDS (CV)	0.066	0.055	0.059	0.075
PMLE-DR	0.057	0.063	0.063	0.047
BR-DR	0.043	0.049	0.051	0.046

Incorrect outcome model

	$n = 200$	$n = 500$	$n = 200$	$n = 200$
Methods	$p = 200$	$p = 500$	$p = 100$	$p = 250$
Standard naïve (forced)	0.312	0.277	0.144	0.363
Standard naïve (not forced)	0.139	0.113	0.070	0.140
PDS (pre-specified)	0.094	0.057	0.079	0.091
PO (pre-specified)	0.077	0.050	0.060	0.077
PDS (CV)	0.071	0.061	0.057	0.073
PMLE-DR	0.063	0.057	0.060	0.056
BR-DR	0.030	0.051	0.044	0.041

Data were obtained from the electronic patient data management system of the UZ Ghent ICU. The potential confounders were split up into variables assessed at admission into the intensive care unit and variables where data were collected over time. For covariates that were measured repeatedly, we took the mean of the measurements taken within the previous 48 hours to the considered day of follow up in the ICU for continuous covariates, and the maximum value for categorical covariates. Measurements on glycemia were usually recorded multiple times per day, so in order to create the exposure, we took the mean of the measurements from within the first 6 hours of the day. For this illustration, any patients with missing data on the exposure, outcome or confounders were removed from the dataset. In order to perform our test, at each day we assumed (in individuals still alive) a logistic regression model for the probability of glycemia level $\geq$110 mg/dl (or >150 mg/dl) as well as a logistic model for death within 30 days of entering hospital. To avoid the issues associated with time-varying confounding described e.g. in Robins (1997), in each regression model we adjusted only for covariate data, as well as previous exposures, collected prior to the glycemia measurements on a given day. We then used an amended version of the test for binary outcomes described in Section 5.1.2 (implemented using Algorithm 1 described in Appendix C), allowing for potential misspecification in either the exposure or outcome model. Given that the data consisted of multiple observations per individual, then letting t denote a particular day, we replaced $U_i(\hat{\eta})$ with $\sum_t U_{it}(\hat{\eta})$ in the statistic T_n (for $t = 3, ..., 30$). In our modeling, we included all confounders selected by clinical experts, as well as quadratic terms of continuous variables and all two-way interactions between main effects.

We obtained data on 12,105 patients entering the intensive care unit; after restricting to patients still alive at day 3, 10,885 individuals remained. Further removing patients entering prior to 2013 left 4,682 individuals, with a final dataset of 4,120 after removing those with missing data. Given that patients were assessed on multiple days, there were 24,863 observations in the dataset; the median number of contributed observations was 3 (the minimum was 1 and the maximum 28). In this final cohort, 768 (18.6%) of individuals died in hospital within 30 days of entering the ICU. Considering the mean glycemia values for patients within the first 6 hours of day 3, the average of these values among all patients was 131.6 (minimum: 45, maximum: 492). 927 (23.3%) patients had mean glycemia at day 3 <110 mg/dl, 2,208 (55.5%) had a level $\geq$110mg/dl and $\leq$150mg/dl and 841 (21.2%) had a level >150 mg/dl. After generating interactions, there were 148 covariates to adjust for in the analysis. Looking at a change at each day from <110 to $\geq$110 mg/dl, the test statistic T_n was -1.42 with a p-value of 0.156 whereas changing from $\leq$150 to >150 mg/dl gave a test statistic of 6.98 (p <0.001). Hence, at the 5% level, we saw evidence of a difference in 30 day mortality based on a change from moderate ($\leq$150mg/dl) to high (>150 mg/dl) glycemia levels on a given day. On the other hand, in comparing those with low (<110mg/dl) vs. higher ($\geq$110 mg/dl) glycemia levels, we did not observe a statistically significant difference at the 5% level.

8 Discussion

We have proposed a general framework for constructing uniformly valid tests of GLM parameters in high-dimensional settings. We hope to have clarified why locally doubly robust methods (in this case, doubly robust under the null) have a privileged position in the literature (Farrell, 2015); if all working models are correct, one can obtain a uniformly valid test by plugging in any sufficiently fast-converging sparse estimator of the nuisance parameters. If one of the working models is misspecified, then one can still obtain uniformly valid inference, so long as a specific estimation procedure for the nuisance parameters is used. We have also indicated why score tests might be preferable in high-dimensional settings, since then the outcome model can be fit under the null hypothesis, enabling one to weaken conditions on sparsity.

In future work, we will extend our procedures to the estimation of regression parameters and the construction of confidence intervals. Consider the model $\mathcal{M}$ defined by the restriction

$$g\{E(Y|A=a,L=l)\} - g\{E(Y|A=0,L=l)\} = \psi a$$

where $g(\cdot)$ is a known link function. The score $U(\eta)$ implies an estimator of ψ, the conditional causal effect of A on Y. Let $H(\psi) = Y - \psi A$ when $g()$ is the identity link and $H(\psi) = Y\exp(-\psi A)$ when $g()$ is the log link; then estimation of ψ can be based on the function

$$U(\psi, \eta) = \{A - \pi(L; \gamma)\}\{H(\psi) - m(L; \beta)\} \tag{10}$$

(Robins et al., 1992). An estimator of ψ based on (10) is consistent under model $\mathcal{M} \cap (\mathcal{A} \cup \mathcal{B})$. The goal of constructing uniformly valid confidence intervals could require a revision of the conditions given in Sections 4.2 and 5.2, since we are no longer working under the null. It also remains an open question for which settings doubly robust estimators can be constructed. For example, there currently exists no doubly robust estimator for the Cox proportional hazards model or probit models. In practice it may be more feasible to construct estimators and confidence intervals that are locally doubly robust e.g. under the null, and in this context enjoy the properties of the tests described in this paper.

When ψ is multivariate, equations (5) and (6) deliver more estimating equations than there are unknown nuisance parameters. To ensure that standard errors are valid, one would also need to ensure that the estimating functions of each component of ψ are orthogonal to those of the remaining components. Such a development would not only be advantageous in terms of testing for and estimating interaction terms, but also for obtaining uniformly valid inference in high-dimensional data with mediators and/or time dependent confounders. Indeed, the estimators described above are special cases of *g-estimators* (Robins et al., 1992), developed for fitting *structural nested models* in complex longitudinal studies. Because it turns out to be essentially impossible to correctly specify sequential regression models for an outcome, it is unlikely that existing proposals for high-dimensional inference can be adapted to test the hypothesis of no causal effect of any treatment *regime* on Y a.k.a the *g-null hypothesis* (Robins, 1997). In contrast, although we perform selection on both the outcome and exposure models (in order for the relevant gradients to be set to zero), in the proposal of Section 5 only the latter needs to be correctly specified in order to obtain a valid test of the g-null.

Acknowledgements

Oliver Dukes is supported by the Strategic Basic Research PhD grant 1S05916N from the Research Foundation - Flanders (FWO). Vahe Avagyan and Stijn Vansteelandt are supported by the FWO research project G016116N and Special Research Fund (BOF) research project BOF.244.2017.0004.01. The authors are grateful to Prof. Johan Decruyenaere and Prof. Kirsten Colpaert for permission to use the Ghent University Intensive Care Unit dataset, and to Bram Gadeyne and Martijn Busselen for assistance in managing the data.

A Appendix A

Define $\mathbb{N}_p = \{1, 2, ..., p\}$. We use $\mathbb{E}_{P_n}[]$ for taking expectation w.r.t. the local data generating process (DGP), whereas $\mathbb{E}_n[]$ refers to sample expectations. Similarly, $\mathbb{P}_{P_n}[]$ and $\text{var}_{P_n}[]$ denote probabilities and variances taken w.r.t. the local DGP respectively. We define the ℓ_∞ norm of any matrix A as $||A||_\infty = \max_{i,j} |A_{ij}|$. For a vector $\delta \in \mathbb{R}^p$ and indices $T \subset \{1, ..., p\}$, let δ_T denote the vector where $\delta_{Tj} = \delta_j$ if $j \in T$ and $\delta_{Tj} = 0$ otherwise. Also, $T^C = \{1, ..., p\} \backslash T$. For sequences a_n and b_n, we use $a_n \lesssim b_n$ to denote that $a_n \leq C b_n$ for some constant C ($a_n \gtrsim b_n$ is similarly defined). We view the gradients $\partial U_i(\hat{\eta})/\partial \eta$ and $\partial U_i(\eta_n)/\partial \eta$ as row vectors.

A.1 Proof of Theorem 1

In order to prove Theorem 1, we will rely on the following assumptions.

Assumption 1. *(Data generating process) There exist constants* $C_1, C_2, C_3 < \infty$, $c_4, c_5 > 0$ *and* $4 < r < \infty$ *such that:*

(i) $\mathbb{E}_{P_n}\{|Y - m(L; \beta_n)|^4 | L\} \leq C_1$ *w.p. 1.*

(ii) $\mathbb{E}_{P_n}\{|Y - m(L; \beta_n)|^r\} \leq C_2$.

(iii) $\max_{i \leq n} \|L_i\|_\infty \leq C_3 < \infty$ *w.p. 1.*

(iv) $c_4 \leq \mathbb{E}_{P_n}[\{A - \pi(L; \gamma_n)\}^2 | L]$ *and* $c_5 \leq \mathbb{E}_{P_n}[\{Y - m(L; \beta_n)\}^2 | L]$ *w.p. 1.*

Remark. Assumption 1(i) allows one to bound the conditional variance of $Y - m(L; \beta_n)$ given L and also implies a bound on the variance given A and L. Assumption 1(ii) places a bound on the higher order moments of $Y - m(L; \beta_n)$, and is required to show uniform consistency of the variance estimator of $U(\hat{\eta})$ and uniform asymptotic normality of the test statistic. We note that Assumptions 1(i)-(ii) allow for non-Gaussianity and heteroscedasticity with respect to the error term $Y - m(L; \beta_n)$. Assumption 1(iii) requires L to be restricted to a bounded set, which is an assumption commonly made in the literature (sometimes as a primitive condition for proving consistency of estimators) (van de Geer et al., 2014; Farrell, 2015; Ning and Liu, 2017). Assumption 1(iv) places additional bounds on the conditional variance, and implies a type of 'positivity' condition such that there must be some variation in A at different levels of L.

Assumption 2. *(Rates for prediction error with unweighted estimators)*

(i) $\mathbb{E}_n[\{\pi(L_i; \gamma_n) - \pi(L_i; \hat{\gamma})\}^2] = O_{P_n}(s_\gamma \log(p \vee n)/n)$.

(ii) $\mathbb{E}_n[\{m(L_i; \beta_n) - m(L_i; \hat{\beta})\}^2] = O_{P_n}(s_\beta \log(p \vee n)/n)$.

Remark. Results 2(i)-2(ii) follow from the results of Belloni and Chernozhukov (2013), Belloni et al. (2014), Farrell (2015) and Belloni et al. (2016) on Lasso and Post-Lasso-based estimators. Rates on quantities like $\mathbb{E}_n[\{\gamma_n^T L_i - \hat{\gamma}^T L_i\}^2]$ also follow from those papers.

Proof. The proof will proceed in four steps. In the first step, we show that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i(\hat{\eta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i(\eta_n) + o_{P_n}(1) \tag{A.1}$$

in the second, that

$$\frac{\mathbb{E}_n\{U_i(\eta_n)\}}{\sqrt{\frac{1}{n}\mathbb{E}_{P_n}\{U_i(\eta_n)^2\}}} \xrightarrow{d} \mathcal{N}(0,1) \tag{A.2}$$

in the third, that

$$\mathbb{E}_n[U_i(\hat{\eta})^2 - \mathbb{E}_n\{U_i(\hat{\eta})\}^2]^{-1} = \mathbb{E}_{P_n}\{U_i(\eta_n)^2\}^{-1} + o_{P_n}(1) \tag{A.3}$$

Finally, we will use these results to show result (4) in the main paper.

Step 1.

Consider the sample mean of $U_i(\hat{\eta})$:

$$\mathbb{E}_n\{U_i(\hat{\eta})\} = \mathbb{E}_n\{U_i(\eta_n) + U_i(\hat{\eta}) - U_i(\eta_n)\}$$

After some algebra, we have

$$\sqrt{n}\mathbb{E}_n\{U_i(\hat{\eta})\} = \sqrt{n}\mathbb{E}_n\{U_i(\eta_n)\} + R_1 + R_2 + R_3$$

where

$$\begin{aligned}
R_1 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i - \pi(L_i;\gamma_n)\}\{m(L_i;\beta_n) - m(L_i;\hat{\beta})\},\\
R_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Y_i - m(L_i;\beta_n)\}\{\pi(L_i;\gamma_n) - \pi(L_i;\hat{\gamma})\}\\
R_3 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(L_i;\hat{\beta}) - m(L_i;\beta_n)\}\{\pi(L_i;\hat{\gamma}) - \pi(L_i;\gamma_n)\}
\end{aligned}$$

We aim to show that R_1, R_2 and R_3 are all $o_{P_n}(1)$ under model $\mathcal{A} \cap \mathcal{B}$.

For R_1, under the null as defined in (1),

$$\begin{aligned}
&\mathbb{E}_{P_n}\{R_1|(Y_i, L_i)_{i=1}^n\}\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathbb{E}_{P_n}\{A_i|(Y_i, L_i)_{i=1}^n\} - \pi(L_i;\gamma_n)]\{m(L_i;\beta_n) - m(L_i;\hat{\beta})\}\\
&= 0
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{P_n}\{R_1^2|(Y_i, L_i)_{i=1}^n\}\\
&= \mathbb{E}_n\left(\mathbb{E}_{P_n}[\{A_i - \pi(L_i;\gamma_n)\}^2|(Y_i, L_i)_{i=1}^n]\{m(L_i;\beta_n) - m(L_i;\hat{\beta})\}^2\right)\\
&\leq C\mathbb{E}_n[\{m(L_i;\beta_n) - m(L_i;\hat{\beta})\}^2]
\end{aligned}$$

where C is a constant. Furthermore, invoking Assumption 2(ii) and sparsity condition (i), we have

$$C\mathbb{E}_n[\{m(L_i;\beta_n) - m(L_i;\hat{\beta})\}^2] = o_{P_n}(1)$$

and $\mathbb{E}_{P_n}[R_1^2] = o(1)$. Hence one can then apply Chebyshev's inequality to show that $|R_1| = o_{P_n}(1)$.

Similarly, for R_2,

$$\begin{aligned}
&\mathbb{E}_{P_n}[R_2^2|(A_i, L_i)_{i=1}^n] \\
&= \mathbb{E}_n\Big[\mathbb{E}_{P_n}[\{Y_i - m(L_i;\beta_n)\}^2|(A_i, L_i)_{i=1}^n]\{\pi(L_i;\gamma_n) - \pi(L_i;\hat{\gamma})\}^2\Big] \\
&\leq C\mathbb{E}_n[\{\pi(L_i;\gamma_n) - \pi(L_i;\hat{\gamma})\}^2],
\end{aligned}$$

where C is a constant. This inequality follows from Assumption 1(i). Invoking Assumption 2(i) and sparsity condition (i), we have

$$C\mathbb{E}_n[\{\pi(L_i;\gamma_n) - \pi(L_i;\hat{\gamma})\}^2] = o_{P_n}(1)$$

so $\mathbb{E}_{P_n}[R_2^2] = o(1)$ and using Chebyshev's inequality, $|R_2| = o_{P_n}(1)$.

Finally, considering R_3, by Hölder's inequality

$$\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\{m(L_i;\hat{\beta}) - m(L_i;\beta_n)\}\{\pi(L_i;\hat{\gamma}) - \pi(L_i;\gamma_n)\}\right| \\
&\leq \sqrt{n}\mathbb{E}_n[\{m(L_i;\hat{\beta}) - m(L_i;\beta_n)\}^2]^{1/2}\mathbb{E}_n[\{\pi(L_i;\hat{\gamma}) - \pi(L_i;\gamma_n)\}^2]^{1/2}
\end{aligned}$$

Then given the joint sparsity condition (ii) on s_γ and s_β, and Assumptions 2(i) and 2(ii), it follows that

$$\sqrt{n}\mathbb{E}_n[\{m(L_i;\hat{\beta}) - m(L_i;\beta_n)\}^2]^{1/2}\mathbb{E}_n[\{\pi(L_i;\hat{\gamma}) - \pi(L_i;\gamma_n)\}^2]^{1/2} = o_{P_n}(1)$$

Therefore $|R_3| = o_{P_n}(1)$ and we have result (A.1).

Step 2.

Under the null, we have that

$$\begin{aligned}\mathrm{var}_{P_n}\{U_i(\eta_n)\} &= \mathbb{E}_{P_n}\{U_i(\eta_n)^2\}\\ &= \mathbb{E}_{P_n}[\{A_i - \pi(L_i;\gamma_n)\}^2\{Y_i - m(L_i;\beta_n)\}^2]\end{aligned}$$

and by Assumptions 1(ii) and 1(iv), $\mathbb{E}_{P_n}\{U_i(\eta_n)^2\}$ is bounded away from zero (necessary for the inversion) and above uniformly in n.

For some $\epsilon > 0$, such that $4 + 2\epsilon \leq r$

$$\begin{aligned}&\mathbb{E}_{P_n}\{|U_i(\eta_n)|^{2+\epsilon}\}\\ &\leq \mathbb{E}_{P_n}\{|A_i - \pi(L_i;\gamma_n)|^{4+2\epsilon}\}^{1/2}\mathbb{E}_{P_n}\{|Y_i - m(L_i;\beta_n)|^{4+2\epsilon}\}^{1/2}\\ &\leq C\end{aligned}$$

where C is a constant, by Assumption 1(ii). This verifies the Lyapunov condition, such that using this result (and the fact that $\mathbb{E}_{P_n}\{U_i(\eta_n)^2\}$ is finite) one can then invoke the Lyapunov central limit theorem for triangular arrays to get result (A.2). We rely on array asymptotics here in order to allow for the data-generating process to change with n.

Step 3.

Since $\mathbb{E}_{P_n}\{U_i(\eta_n)^2\}$ is bounded away from zero uniformly in n and $\mathbb{E}_{P_n}\{U_i(\eta_n)\} = 0$, given the previous steps it suffices to show that $\mathbb{E}_n\{U_i(\hat{\eta})^2\} = \mathbb{E}_{P_n}\{U_i(\eta_n)^2\} + o_{P_n}(1)$. We will first obtain the result

$$\mathbb{E}_n\{U_i(\eta_n)^2\} = \mathbb{E}_{P_n}\{U_i(\eta_n)^2\} + o_{P_n}(1) \tag{A.4}$$

We have

$$\begin{aligned}
&\mathbb{P}_{P_n}\left[|\mathbb{E}_n\{U_i(\eta_n)^2\} - \mathbb{E}_{P_n}\{U_i(\eta_n)^2\}|^2 > \epsilon\right] \\
&\leq \frac{1}{\epsilon^2}\mathbb{E}_{P_n}\left[\left|\frac{1}{n}\sum_{i=1}^{n} U_i(\eta_n)^2 - \mathbb{E}_{P_n}\{U_i(\eta_n)^2\}\right|^2\right] \\
&\leq \frac{1}{\epsilon^2 n^2}\left(2 - \frac{1}{n}\right)\sum_{i=1}^{n}\mathbb{E}_{P_n}\left[\left|U_i(\eta_n)^2 - \mathbb{E}_{P_n}\{U_i(\eta_n)^2\}\right|^2\right]
\end{aligned}$$

where we first apply Chebyshev's inequality. The second uses the Von Bahr-Esseen inequality: let $q \in [1, 2]$, then for independent mean-zero variables $X_1, ..., X_n$, we have

$$E\left(\left|\sum_{i=1}^{n} X_i\right|^q\right) \leq \left(2 - \frac{1}{n}\right)\sum_{i=1}^{n} E(|X_i|^q)$$

(von Bahr and Esseen, 1965).

Since

$$\begin{aligned}
&\mathbb{E}_{P_n}\left([U_i(\eta_n)^2 - \mathbb{E}_{P_n}\{U_i(\eta_n)^2\}]^2\right) \\
&= \text{var}_{P_n}\{U_i(\eta_n)^2\} \\
&= \text{var}_{P_n}\left(\{A_i - \pi(L_i;\gamma_n)\}^2\mathbb{E}_{P_n}[\{Y_i - m(L_i;\beta_n)\}^2|A_i, L_i]\right) \\
&\quad + \mathbb{E}_{P_n}\left(\{A_i - \pi(L_i;\gamma_n)\}^4\text{var}_{P_n}[\{Y_i - m(L_i;\beta_n)\}^2|A_i, L_i]\right)
\end{aligned}$$

then firstly

$$\begin{aligned}
&\text{var}_{P_n}\left(\{A_i - \pi(L_i;\gamma_n)\}^2\mathbb{E}_{P_n}[\{Y_i - m(L_i;\beta_n)\}^2|A_i, L_i]\right) \\
&\leq C\mathbb{E}_{P_n}[\{A_i - \pi(L_i;\gamma_n)\}^2]^2 = O(1)
\end{aligned}$$

where C is a constant, using Assumption 1(i). Secondly,

$$\begin{aligned}
&\mathbb{E}_{P_n}\left(\{A_i - \pi(L_i;\gamma_n)\}^4\text{var}_{P_n}[\{Y_i - m(L_i;\beta_n)\}^2|A_i, L_i]\right) \\
&\leq C\mathbb{E}_{P_n}[\{A_i - \pi(L_i;\gamma_n)\}^4] = O(1)
\end{aligned}$$

where C is again a constant, invoking Assumptions 1(i) and 1(ii). Result (A.4) then follows.

It remains to show that

$$\mathbb{E}_n\{U_i(\hat{\eta})^2\} = \mathbb{E}_n\{U_i(\eta_n)^2\} + o_{P_n}(1) \tag{A.5}$$

By adding and subtracting $\mathbb{E}_n[\{A_i - \pi(L_i;\hat{\gamma})\}^2\{Y_i - m(L_i;\beta_n)\}^2]$ and applying the triangle inequality, then

$$\begin{aligned}
&|\mathbb{E}_n[\{A_i - \pi(L_i;\hat{\gamma})\}^2\{Y_i - m(L_i;\hat{\beta})\}^2 - \{A_i - \pi(L_i;\gamma_n)\}^2\{Y_i - m(L_i;\beta_n)\}^2]| \\
&\leq \big|\mathbb{E}_n\big([\{A_i - \pi(L_i;\hat{\gamma})\}^2 - \{A_i - \pi(L_i;\gamma_n)\}^2]\{Y_i - m(L_i;\beta_n)\}^2\big)\big| \\
&\quad + \big|\mathbb{E}_n\big([\{Y_i - m(L_i;\hat{\beta})\}^2 - \{Y_i - m(L_i;\beta_n)\}^2]\{A_i - \pi(L_i;\hat{\gamma})\}^2\big)\big| \\
&= |R_4| + |R_5|
\end{aligned}$$

Looking first at $|R_5|$, after some algebra we have

$$\begin{aligned}
&\big|\mathbb{E}_n\big([\{Y_i - m(L_i;\hat{\beta})\}^2 - \{Y_i - m(L_i;\beta_n)\}^2]\{A_i - \pi(L_i;\hat{\gamma})\}^2\big)\big| \\
&\leq |\mathbb{E}_n[\{m(L_i;\hat{\beta}) - m(L_i;\beta_n)\}^2\{A_i - \pi(L_i;\hat{\gamma})\}^2]| \\
&\quad + |2\mathbb{E}_n[\{Y_i - m(L_i;\beta_n)\}\{m(L_i;\hat{\beta}) - m(L_i;\beta_n)\}\{A_i - \pi(L_i;\hat{\gamma})\}^2]| \\
&= |R_{5a}| + |R_{5b}|.
\end{aligned}$$

Then,

$$|R_{5a}| \leq \max_{i\leq n}\{A_i - \pi(L_i;\hat{\gamma})\}^2\mathbb{E}_n[\{m(L_i;\hat{\beta}) - m(L_i;\beta_n)\}^2] = o_{P_n}(1)$$

following Assumption 2(ii), the fact that A is binary and sparsity condition (ii). Furthermore,

$$\begin{aligned}
|R_{5b}| \leq 2\max_{i\leq n}&\{A_i - \pi(L_i;\hat{\gamma})\}^2\mathbb{E}_n[\{Y_i - m(L_i;\beta_n)\}^2]^{1/2} \\
&\times \mathbb{E}_n[\{m(L_i;\hat{\beta}) - m(L_i;\beta_n)\}^2]^{1/2}
\end{aligned}$$

As before, one can bound $\max_{i\leq n}\{A_i - \pi(L_i;\hat{\gamma})\}^2$, and $\mathbb{E}_n[\{m(L_i;\hat{\beta}) - m(L_i;\beta_n)\}^2]^{1/2} = o_{P_n}(1)$ by Assumption 2(ii). For $\mathbb{E}_n[\{Y_i - m(L_i;\beta_n)\}^2]^{1/2}$, note that by Assumption 1(ii), $\mathbb{E}_{P_n}\{|Y_i - m(L_i;\beta_n)|^4\} = O(1)$ and hence $\mathbb{E}_{P_n}[\{Y_i - m(L_i;\beta_n)\}^2]^{1/2} = O(1)$ since the bound on the higher-order moment implies the existence of the lower-order moment. To bound the sample average $\mathbb{E}_n[\{Y_i - m(L_i;\beta_n)\}^2]^{1/2}$, by the Von Bahr-Esseen inequality:

$$\begin{aligned}
&\mathbb{P}_{P_n}\left(\left|\mathbb{E}_n\{|Y_i - m(L_i;\beta_n)|^4\} - \mathbb{E}_{P_n}\{|Y_i - m(L_i;\beta_n)|^4\}\right|^q > \epsilon\right)\\
&\leq \frac{1}{\epsilon^q n^q}\left(2 - \frac{1}{n}\right)\sum_{i=1}^{n}\mathbb{E}_{P_n}\left[\left||Y_i - m(L_i;\beta_n)|^4 - \mathbb{E}_{P_n}\{|Y_i - m(L_i;\beta_n)|^4\}\right|^q\right]
\end{aligned}$$

for $q \in [1,2]$. Applying Minkowski's inequality and using Assumption 1(ii):

$$\begin{aligned}
&\mathbb{E}_{P_n}\left[\left||Y_i - m(L_i;\beta_n)|^4 - \mathbb{E}_{P_n}\{|Y_i - m(L_i;\beta_n)|^4\}\right|^q\right]\\
&\leq \left(\mathbb{E}_{P_n}\left\{|Y_i - m(L_i;\beta_n)|^{4q}\right\}^{1/q} + \mathbb{E}_{P_n}\left[\mathbb{E}_{P_n}\{|Y_i - m(L_i;\beta_n)|^4\}^q\right]^{1/q}\right)^q\\
&= O(1),
\end{aligned}$$

hence $\mathbb{E}_n\{|Y_i - m(L_i;\beta_n)|^4\} = O_{P_n}(1)$ and also $\mathbb{E}_n[\{Y_i - m(L_i;\beta_n)\}^2]^{1/2} = O_{P_n}(1)$. Therefore $|R_5| = o_{P_n}(1)$.

Similarly, for R_4 we have

$$\begin{aligned}
|R_4| \leq &\mathbb{E}_n[\{\pi(L_i;\hat{\gamma}) - \pi(L_i;\gamma_n)\}^2\{Y_i - m(L_i;\beta_n)\}^2]\\
&+ 2\max_{i\leq n}|A_i - \pi(L_i;\gamma_n)|\mathbb{E}_n\{|Y_i - m(L_i;\beta_n)|^4\}^{1/2}\\
&\times \mathbb{E}_n[\{\pi(L_i;\hat{\gamma}) - \pi(L_i;\gamma_n)\}^2]^{1/2}
\end{aligned}$$

By invoking Assumptions 1(ii), 2(i) and the sparsity condition (i), one can show that the second term on the right hand side of the inequality is $o_{P_n}(1)$. Regarding the first term,

$$\begin{aligned}
&\mathbb{E}_n[\{\pi(L_i;\hat{\gamma}) - \pi(L_i;\gamma_n)\}^2\{Y_i - m(L_i;\beta_n)\}^2] \\
&\leq \max_{i\leq n} |\pi(L_i;\hat{\gamma}) - \pi(L_i;\gamma_n)|\mathbb{E}_n[\{\pi(L_i;\hat{\gamma}) - \pi(L_i;\gamma_n)\}^2]^{1/2} \\
&\quad \times \mathbb{E}_n\{|Y_i - m(L_i;\beta_n)|^4\}^{1/2} \\
&= o_{P_n}(1)
\end{aligned}$$

using Hölder's inequality, Assumptions 1(ii), 2(i), the previous result that $\mathbb{E}_n\{|Y_i - m(L_i;\beta_n)|^4\} = O_{P_n}(1)$ and the sparsity condition (i). We have shown (A.5) and result (A.3) follows.

Step 4.

Consider a sequence $P_n \in P$ such that for any $t \in \mathbb{R}$

$$\lim_{n\to\infty} |\mathbb{P}_{P_n}(T_n \leq t) - \Phi(t)| > 0.$$

This directly contradicts the results given above that the test statistic T_n converges to a normal distribution with mean 0 and variance 1 under any subsequence P_n in P. □

A.2 Proof of Theorem 2

In the proofs of Theorems 2 and Theorem 3, we will rely on some additional assumptions. Specifically, we will make use of the following rates:

Assumption 3. *(Rates on error of estimated coefficients) Recall that $s^* = s_\gamma \vee s_\beta$; then*

(i) $||\gamma_n - \hat{\gamma}_{BR}||_1 = O_{P_n}(s^*\sqrt{\log(p \vee n)/n})$.

(ii) $||\beta_n - \hat{\beta}_{BR}||_1 = O_{P_n}(s^*\sqrt{\log(p \vee n)/n})$.

(iii) $||\gamma_n - \hat{\gamma}_{BR}||_2 = O_{P_n}(\sqrt{s^* \log(p \vee n)/n})$.

(iv) $||\beta_n - \hat{\beta}_{BR}||_2 = O_{P_n}(\sqrt{s^* \log(p \vee n)/n})$.

Assumption 4. *(Rates on prediction error for weighted estimators)*

(i) $\mathbb{E}_n[\{\pi(L_i;\gamma_n) - \pi(L_i;\hat{\gamma}_{BR})\}^2] = O_{P_n}(s^* \log(p \vee n)/n)$.

(ii) $\mathbb{E}_n[\{m(L_i;\beta_n) - m(L_i;\hat{\beta}_{BR})\}^2] = O_{P_n}(s^* \log(p \vee n)/n)$.

Remark. Our proposed nuisance parameter estimators are obtained via (weighted) ℓ_1 penalized regression. The rates in 3 and 4 again follow from the results of Belloni et al. (2016) on weighted ℓ_1-penalized regression (e.g. their Theorem 4); see also Ning and Liu (2017).

To obtain these rates in Assumptions 3 and 4, we need assumptions on the order of the penalty (which we exploit in the proof); specifically, we need that

$$\lambda_\gamma = O\left(\sqrt{\frac{\log(p \vee n)}{n}}\right) \tag{A.6}$$

$$\lambda_\beta = O\left(\sqrt{\frac{\log(p \vee n)}{n}}\right). \tag{A.7}$$

These are standard assumption on the order of the penalty level in the literature, when working either under the intersection submodel (Farrell, 2015; Belloni et al., 2016), or the union model (Avagyan and Vansteelandt, 2017).

Proof. Repeating the previous decomposition of

$\sqrt{n}\mathbb{E}_n\{U_i(\hat{\eta}_{BR})\} - \sqrt{n}\mathbb{E}_n\{U_i(\eta_n)\}$, for R_1, we now have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i - \pi(L_i;\gamma_n)\}\{m(L_i;\beta_n) - m(L_i;\hat{\beta}_{BR})\} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i - \pi(L_i;\hat{\gamma}_{BR})\}\{m(L_i;\beta_n) - m(L_i;\hat{\beta}_{BR})\} \qquad \text{(A.8)} \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\pi(L_i;\hat{\gamma}_{BR}) - \pi(L_i;\gamma_n)\}\{m(L_i;\beta_n) - m(L_i;\hat{\beta}_{BR})\} \qquad \text{(A.9)}
\end{aligned}$$

Then for (A.8), note that following a Taylor expansion,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i - \pi(L_i;\hat{\gamma}_{BR})\}\{m(L_i;\beta_n) - m(L_i;\hat{\beta}_{BR})\} \\
&= -\sqrt{n}\mathbb{E}_n\left\{\frac{\partial U_i(\hat{\eta}_{BR})}{\partial \beta}\right\}(\beta_n - \hat{\beta}_{BR}) + O_{P_n}(\sqrt{n}||\beta_n - \hat{\beta}_{BR}||_2^2)
\end{aligned}$$

and by Hölder's inequality,

$$\begin{aligned}
&\left|\sqrt{n}\mathbb{E}_n\left\{\frac{\partial U_i(\hat{\eta}_{BR})}{\partial \beta}\right\}(\beta_n - \hat{\beta}_{BR})\right| \\
&\leq \sqrt{n}\lambda_\gamma||\beta_n - \hat{\beta}_{BR}||_1
\end{aligned}$$

using the stationarity conditions for the ℓ_1-penalised estimator of γ_0. Therefore, given (A.6), Assumptions 3(ii), 3(iv) and sparsity condition (iii),

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i - \pi(L_i;\hat{\gamma}_{BR})\}\{m(L_i;\beta_n) - m(L_i;\hat{\beta}_{BR})\}\right| = o_{P_n}(1)$$

Considering the other term (A.9), along the same lines as in the proof of Theorem 1, one can show that

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\pi(L_i;\hat{\gamma}_{BR}) - \pi(L_i;\gamma_n)\}\{m(L_i;\beta_n) - m(L_i;\hat{\beta}_{BR})\}\right| = o_{P_n}(1)$$

using Hölder's inequality, sparsity condition (iii) and Assumptions 4(i) and 4(ii). Therefore $|R_1| = o_{P_n}(1)$. One can re-apply the argument given immediately above to show that $|R_3| = o_{P_n}(1)$.

By noting that

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Y_i - m(L_i;\hat{\beta}_{BR})\}\{\pi(L_i;\gamma_n) - \pi(L_i;\hat{\gamma}_{BR})\}\\&= -\sqrt{n}\mathbb{E}_n\left\{\frac{\partial U_i(\hat{\eta}_{BR})}{\partial\gamma}\right\}(\gamma_n - \hat{\gamma}_{BR}) + O_{P_n}(\sqrt{n}||\gamma_n - \hat{\gamma}_{BR}||_2^2)\end{aligned}$$

and

$$\begin{aligned}&\left|\sqrt{n}\mathbb{E}_n\left\{\frac{\partial U_i(\hat{\eta}_{BR})}{\partial\gamma}\right\}(\gamma_n - \hat{\gamma}_{BR})\right|\\&\leq \sqrt{n}\lambda_\beta||\gamma_n - \hat{\gamma}_{BR}||_1\end{aligned}$$

one can also repeat the above arguments to show that $|R_2| = o_{P_n}(1)$, given (A.7), Assumptions 3(ii), 3(iv), 4(i), 4(ii) and sparsity condition (iii). Result (A.1) follows immediately and the main result follows by essentially repeating Steps 2-4 from the proof of Theorem 1. □

A.3 Proof of Corollary 1

Proof. As discussed in the main paper, when $E(A|L) = \pi(L;\gamma^*)$ and γ^* is known, the proposal in Section 5 for estimating β reduces to standard (unweighted) PMLE. Then repeating Step 1 of the previous proof,

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i - \pi(L_i;\gamma^*)\}\{Y_i - m(L_i;\hat{\beta})\}\\&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i - \pi(L_i;\gamma^*)\}\{Y_i - m(L_i;\beta_n)\}\\&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i - \pi(L_i;\gamma^*)\}\{m(L_i;\beta_n) - m(L_i;\hat{\beta})\}\end{aligned}$$

Let us define $R_1^* = \sqrt{n}\mathbb{E}_n[\{A_i - \pi(L_i;\gamma^*)\}\{m(L_i;\beta_n) - m(L_i;\hat{\beta})\}]$.

Then,

$$\begin{aligned}
&\mathbb{E}_{P_n}\{R_1^{*2}|(Y_i, L_i)_{i=1}^n\} \\
&= \mathbb{E}_n\left(\mathbb{E}_{P_n}[\{A_i - \pi(L_i; \gamma^*)\}^2|(L_i)_{i=1}^n]\{m(L_i; \beta_n) - m(L_i; \hat{\beta})\}^2\right) \\
&\leq C\mathbb{E}_n[\{m(L_i; \beta_n) - m(L_i; \hat{\beta})\}^2]
\end{aligned}$$

where C is a constant. Invoking 4(ii) and sparsity condition (i), we have

$$C\mathbb{E}_n[\{m(L_i; \beta_n) - m(L_i; \hat{\beta})\}^2] = o_{P_n}(1),$$

hence $\mathbb{E}_{P_n}[R_1^{*2}] = o(1)$ and $|R_1^*| = o_{P_n}(1)$ using Chebyshev's Inequality. Note that sparsity condition (iii) has not been invoked. □

A.4 Auxiliary results on weighted estimators

Here, we restrict to settings where $m(L; \beta_n) = \beta_n^T L$ and $\pi(L; \gamma_n) = \text{expit}(\gamma_n^T L)$. In this case, note that $\hat{\gamma}_{BR} = \hat{\gamma}$, since no weights are used in estimating this parameter. In order to make transparent the dependence of the estimator $\hat{\beta}_{BR}$ on the weights, we introduce the notation $\hat{\beta}(\hat{\gamma})$ for when γ (required for the weights) is estimated from the data and $\hat{\beta}(\gamma_n)$ otherwise.

In what follows, we will give a lemma (key to proving Theorem 3) regarding the quantity $\hat{\beta}(\gamma_n) - \hat{\beta}(\hat{\gamma})$. This will be helpful for understanding the impact of using estimated weights on the distribution of the test statistic. Several additional assumptions are required:

Assumption 5. *(Fast convergence of estimating equations)*

$$\begin{aligned}
&\left\|\mathbb{E}_n[w(L_i; \gamma_n)[Y_i - m\{L_i; \hat{\beta}(\gamma_n)\}]L_i - w(L_i; \hat{\gamma})[Y_i - m\{L_i; \hat{\beta}(\hat{\gamma})\}]L_i]\right\|_\infty \\
&= O_{P_n}\left(\frac{\sqrt{\log p}}{n}\right) \qquad \text{(A.10)}
\end{aligned}$$

Remark. This assumption requires the difference between the estimating equations for $\hat{\beta}(\hat{\gamma})$ and $\hat{\beta}(\gamma_n)$ to shrink very quickly. In low-dimensional settings, this difference is exactly zero by virtue of the estimation procedure. In the high-dimensional setting, even stronger results than (A.10) are available for our proposed estimators, if we represent our estimator of β_0 as the solution to estimating equations with a bridge penalty:

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\gamma}U_i(\hat{\eta}_{BR}) + \lambda_\beta\delta|\hat{\beta}_{BR}|^{\delta-1}\circ\text{sign}(\hat{\beta}_{BR}).$$

Here, $\circ$ is the Hadamard product operator. As $\delta \to 1+$, the penalty term corresponds to the subgradient of the ℓ_1 or Lasso norm penalty $||\beta||_1$ with respect to β. Using this representation of our estimator, we can see that

$$\begin{aligned}
&\mathbb{E}_n[w(L_i;\gamma_n)[Y_i - m\{L_i;\hat{\beta}(\gamma_n)\}]L_i - w(L_i;\hat{\gamma})[Y_i - m\{L_i;\hat{\beta}(\hat{\gamma})\}]L_i]\\
&= \sqrt{n}\lambda_\beta\delta[|\hat{\beta}(\gamma_n)|^{\delta-1}\circ\text{sign}(\hat{\beta}(\gamma_n)) - |\hat{\beta}(\hat{\gamma})|^{\delta-1}\circ\text{sign}(\hat{\beta}(\hat{\gamma}))]\\
&= \sqrt{n}\lambda_\beta\delta(\delta-1)|\hat{\beta}(\hat{\gamma})|^{\delta-2}\circ\text{sign}(\hat{\beta}(\hat{\gamma}))\{\hat{\beta}(\gamma_n) - \hat{\beta}(\hat{\gamma})\}\\
&\quad + O_{P_n}(\sqrt{n}||\hat{\beta}(\gamma_n) - \hat{\beta}(\hat{\gamma})||_2^2)
\end{aligned}$$

where the final equality follows from a Taylor expansion around $\hat{\beta}(\hat{\gamma})$. For any finite n, we can choose δ to be close enough to 1 such that

$$\sqrt{n}\lambda_\beta\delta(\delta-1)|\hat{\beta}(\hat{\gamma})|^{\delta-2}\circ\text{sign}(\hat{\beta}(\hat{\gamma}))\{\hat{\beta}(\gamma_n) - \hat{\beta}(\hat{\gamma})\}$$

is negligible, since $\hat{\beta}(\gamma_n) - \hat{\beta}(\hat{\gamma})$ is assumed not to diverge as $\delta \to 1+$.

Assumption 6. *(High-dimensional model selection)*
Let $\hat{s}_\beta$ denote the cardinality $|support\{\hat{\beta}(\hat{\gamma})\} \cup support\{\hat{\beta}(\gamma_n)\}|$; then $\hat{s}_\beta = O_{P_n}(s^*)$.

Remark. This assumption states that the number of non-zero entries common to both $\hat{\beta}(\hat{\gamma})$ and $\hat{\beta}(\gamma_n)$ is of similar order to s^*, and is satisfied when $|\text{support}\{\hat{\beta}(\hat{\gamma})\}| = O_{P_n}(s^*)$; note that it does not require perfect model selection (Belloni et al., 2014).

Before giving Lemma 1, we will also review several regularity conditions necessary for the consistency of ℓ_1-penalized estimators. This is done for clarity, since they will be invoked in the following proofs. For example, we will use the concentration bound

$$\|\mathbb{E}_n[\{A_i - \pi(L_i;\gamma_n)\}L_i]\|_\infty = O_{P_n}\left(\sqrt{\frac{\log(p \vee n)}{n}}\right); \tag{A.11}$$

this can be shown to hold either using the theory of moderate deviations for self-normalised sums (De la Peña et al., 2009; Belloni et al., 2012), or using sub-Gaussian primitive conditions. Furthermore, restricted eigenvalues conditions are also required. For the estimator $\hat{\beta}(\hat{\gamma})$ we need that for any constant $\zeta \geq 1$, there exists a finite constant $\kappa > 0$ such that

$$\min\left[\frac{\hat{\delta}^T M \hat{\delta}}{\left\|\hat{\delta}\right\|_2^2} : \hat{\delta} \in \mathbb{R}^p, \left\|\hat{\delta}_{T^C}\right\|_1 \leq \zeta\left\|\hat{\delta}_T\right\|_1, \hat{\delta} \neq 0\right] \geq \kappa > 0 \tag{A.12}$$

where $\hat{\delta} = \hat{\beta}(\hat{\gamma}) - \beta_n$ and $M = \mathbb{E}_n\{L_i w(L_i;\gamma_n) L_i^T\}$.

Remark. If the cone constraint on $\hat{\delta}$ in (A.12) is satisfied (and the equivalent condition is satisfied for $\tilde{\delta} = \hat{\beta}(\gamma_n) - \beta_n$), then by the triangle inequality,

$$\begin{aligned}\left\|\hat{\beta}(\hat{\gamma})_{T^C} - \hat{\beta}(\gamma_n)_{T^C}\right\|_1 &= \left\|\hat{\beta}(\hat{\gamma})_{T^C} - \beta_{nT^C} + \beta_{nT^C} - \hat{\beta}(\gamma_n)_{T^C}\right\|_1 \\ &\leq \left\|\hat{\delta}_{T^C}\right\|_1 + \left\|\tilde{\delta}_{T^C}\right\|_1 \\ &\leq \zeta\left(\left\|\hat{\delta}_T\right\|_1 + \left\|\tilde{\delta}_T\right\|_1\right).\end{aligned}$$

Defining the set

$$\Delta_{\zeta,T} = \left\{\hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n) \in \mathbb{R}^p\backslash\{0\} : \left\|\hat{\beta}(\hat{\gamma})_{T^C} - \hat{\beta}(\gamma_n)_{T^C}\right\|_1 \leq \zeta\left(\left\|\hat{\delta}_T\right\|_1 + \left\|\tilde{\delta}_T\right\|_1\right)\right\}$$

and the minimum restricted eigenvalue of M as

$$\varphi_\zeta^2(M) = \min_{\hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n) \in \Delta_{\zeta,T}, |T| \leq s^*}\left[\frac{\{\hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n)\}^T M \{\hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n)\}}{\left\|\hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n)\right\|_2^2}\right],$$

condition (A.12) thus implies that

$$\varphi_\zeta^2(M) \geq \kappa > 0. \tag{A.13}$$

For a comprehensive discussion of suitable regularity conditions for proving the consistency of ℓ_1-penalized estimators, we refer the interested reader to Bühlmann and van de Geer (2011).

Lemma 1. *In addition to the sparsity conditions (i) and (iv), suppose that the Assumptions 1(iii), 2(i), 3(i), 4(ii), 5 and 6 hold. Then it follows that for a given sequence P_n we have that*

$$\begin{aligned}&\left\|\hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n)\right\|_1 \\ &= O_{P_n}\left(\max_{i\leq n}|\epsilon_i|\sqrt{\frac{s_\gamma s^* \log(p \vee n)}{n}} + \frac{s^*\sqrt{\log(p\vee n)}(s_\gamma\sqrt{\log(p\vee n)}+1)}{n}\right)\end{aligned} \tag{A.14}$$

where $\epsilon = Y - m(L; \beta_n)$.

Proof. Many of the arguments are similar to those in Appendix E of Ning and Liu (2017). Let $\bar{\delta} = \hat{\beta}(\hat{\gamma}) - \hat{\beta}(\gamma_n)$; then

$$\begin{aligned}\bar{\delta}^T M \bar{\delta} &= \mathbb{E}_n\left(w(L_i;\gamma_n)(\bar{\delta}^T L_i)[Y_i - m\{L_i;\hat{\beta}(\gamma_n)\}]\right) \\ &\quad - \mathbb{E}_n\left(w(L_i;\gamma_n)(\bar{\delta}^T L_i)[Y_i - m\{L_i;\hat{\beta}(\hat{\gamma})\}]\right) \\ &= \mathbb{E}_n\left(w(L_i;\gamma_n)(\bar{\delta}^T L_i)[Y_i - m\{L_i;\hat{\beta}(\gamma_n)\}]\right) \\ &\quad - \mathbb{E}_n\left(w(L_i;\hat{\gamma})(\bar{\delta}^T L_i)[Y_i - m\{L_i;\hat{\beta}(\hat{\gamma})\}]\right) \\ &\quad + \mathbb{E}_n\left[\{w(L_i;\hat{\gamma}) - w(L_i;\gamma_n)\}(\bar{\delta}^T L_i)\{Y_i - m(L_i;\beta_n)\}\right] \\ &\quad + \mathbb{E}_n\left(\{w(L_i;\hat{\gamma}) - w(L_i;\gamma_n)\}(\bar{\delta}^T L_i)[m(L_i;\beta_n) - m\{L_i;\hat{\beta}(\hat{\gamma})\}]\right) \\ &= R_6 + R_7 + R_8 + R_9.\end{aligned}$$

Firstly,

$$\begin{aligned}
&|R_6 + R_7| \\
&\leq \|\mathbb{E}_n[w(L_i;\gamma_n)\{Y_i - \tilde{m}(L_i)\}L_i - w(L_i;\hat{\gamma})\{Y_i - \hat{m}(L_i)\}L_i]\|_\infty \left\|\bar{\delta}\right\|_1 \\
&\leq C' \frac{\sqrt{\log(p \vee n)}}{n} \left\|\bar{\delta}\right\|_1
\end{aligned}$$

where C' is a constant, due to Assumption 5. For R_8,

$$\begin{aligned}
|R_8| &= \left| \mathbb{E}_n \left[\left\{ \frac{w(L_i;\hat{\gamma}) - w(L_i;\gamma_n)}{w(L_i;\gamma_n)(\hat{\gamma}^T L_i - \gamma_n^T L_i)} \right\} \sqrt{w(L_i;\gamma_n)}(\bar{\delta}^T L_i) \right.\right. \\
&\qquad \left.\left. \times \sqrt{w(L_i;\gamma_n)}(\hat{\gamma}^T L_i - \gamma_n^T L_i)\{Y_i - m(L_i;\beta_n)\} \right] \right| \\
&\leq \max_{i \leq n} |\epsilon_i| (\bar{\delta}^T M \bar{\delta})^{1/2} \mathbb{E}_n \left[(\hat{\gamma}^T L_i - \gamma_n^T L_i)^2 \right]^{1/2} \\
&\lesssim \max_{i \leq n} |\epsilon_i| (\bar{\delta}^T M \bar{\delta})^{1/2} \sqrt{\frac{s_\gamma \log(p \vee n)}{n}}
\end{aligned}$$

under Assumption 2 and using the properties of the logistic link function. Next,

$$\begin{aligned}
|R_9| &\leq \mathbb{E}_n[w(L_i;\gamma_n)|\hat{\gamma}^T L_i - \gamma_n^T L_i||\bar{\delta}^T L_i||m(L_i;\beta_n) - m\{L_i;\hat{\beta}(\hat{\gamma})\}|] \\
&\leq \max_{i \leq n} \|L_i\|_\infty \|\hat{\gamma} - \gamma_n\|_1 \mathbb{E}_n[w(L_i;\gamma_n)|\bar{\delta}^T L_i||m(L_i;\beta_n) - m\{L_i;\hat{\beta}(\hat{\gamma})\}|] \\
&\leq \max_{i \leq n} \|L_i\|_\infty \|\hat{\gamma} - \gamma_n\|_1 \\
&\quad \times (\bar{\delta}^T M \bar{\delta})^{1/2} \mathbb{E}_n \left([m(L_i;\beta_n) - m\{L_i;\hat{\beta}(\hat{\gamma})\}]^2 \right)^{1/2} \\
&\lesssim (\bar{\delta}^T M \bar{\delta})^{1/2} \left(\frac{s_\gamma \sqrt{s^*} \log(p \vee n)}{n} \right)
\end{aligned}$$

invoking Assumptions 1(iii), 3(i), 4 and (ii).

Putting this together,

$$\begin{aligned}
\bar{\delta}^T M \bar{\delta} \leq & C' \frac{\sqrt{\log p}}{n} \left\|\bar{\delta}\right\|_1 \\
& + (\bar{\delta}^T M \bar{\delta})^{1/2} C^* \left(\max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma \log(p \vee n)}{n}} + \frac{s_\gamma \sqrt{s^*} \log(p \vee n)}{n} \right) \qquad \text{(A.15)}
\end{aligned}$$

where C^* is a constant. Let us consider two cases; firstly, assume

$$(\bar{\delta}^T M \bar{\delta})^{1/2} \leq C^* \left(\max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma \log(p \vee n)}{n}} + \frac{s_\gamma \sqrt{s^*} \log(p \vee n)}{n} \right)$$

holds. Note (A.13) implies that

$$(\bar{\delta}^T M \bar{\delta})^{1/2} \geq \varphi_\zeta(M) \left\| \bar{\delta} \right\|_2 \gtrsim \frac{1}{\sqrt{\hat{s}_\beta}} \left\| \bar{\delta} \right\|_1 .$$

Invoking Assumption 6 and combining the lower and upper bound,

$$\left\| \bar{\delta} \right\|_1 \lesssim \max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma s^* \log(p \vee n)}{n}} + \frac{s_\gamma s^* \log(p \vee n)}{n} \tag{A.16}$$

On the other hand, if

$$(\bar{\delta}^T M \bar{\delta})^{1/2} \geq C^* \left(\max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma \log(p \vee n)}{n}} + \frac{s_\gamma \sqrt{s^*} \log(p \vee n)}{n} \right)$$

then rearranging (A.15), it follows that

$$\begin{aligned} &(\bar{\delta}^T M \bar{\delta})^{1/2} \left\{ (\bar{\delta}^T M \bar{\delta})^{1/2} - C^* \left(\max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma \log(p \vee n)}{n}} + \frac{s_\gamma \sqrt{s^*} \log(p \vee n)}{n} \right) \right\} \\ &\leq C' \frac{\sqrt{\log(p \vee n)}}{n} \left\| \bar{\delta} \right\|_1 \end{aligned}$$

Using (A.13), we have

$$\begin{aligned} &(\bar{\delta}^T M \bar{\delta})^{1/2} - C^* \left(\max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma \log(p \vee n)}{n}} + \frac{s_\gamma \sqrt{s^*} \log(p \vee n)}{n} \right) \\ &\lesssim \frac{\sqrt{s^* \log(p \vee n)}}{n} \end{aligned}$$

so

$$(\bar{\delta}^T M \bar{\delta})^{1/2} \lesssim \max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma \log(p \vee n)}{n}} + \frac{s_\gamma \sqrt{s^*} \log(p \vee n) + \sqrt{s^* \log(p \vee n)}}{n}$$

and again by (A.13),

$$\left\| \bar{\delta} \right\|_1 \lesssim \max_{i \leq n} |\epsilon_i| \sqrt{\frac{s_\gamma s^* \log(p \vee n)}{n}} + \frac{s^* \sqrt{\log(p \vee n)} (s_\gamma \sqrt{\log(p \vee n)} + 1)}{n}. \tag{A.17}$$

Taking the union of the bounds (A.16) and (A.17) completes the proof. □

A.5 Proof of Theorem 3

Assumption 7. *(Regularity conditions on the errors)* $\max_{i\le n}|\epsilon_i|\sqrt{s_\gamma s^*}\log(p\vee n)=o(\sqrt{n})$ *w.p. 1.*

Remark. This can be most simply shown to hold if $\max_{i\le n}|\epsilon_i| = O_{P_n}(1)$. Belloni et al. (2012) and Farrell (2015) suppose that $\max_{i\le n}|\epsilon_i| = O_{P_n}(n^{1/r})$ for some $r>2$, such that larger values of r allow one to relax assumptions on sparsity in exchange for stronger conditions on the distributions of the errors. If the ϵ_i are normal, then r can be arbitrarily large. Alternatively, one can place the stronger sub-Gaussian conditions on ϵ_i, whereby $\max_{i\le n}|\epsilon_i| = O_{P_n}(\sqrt{\log n})$.

We now give the proof.

Proof. Decomposing $\sqrt{n}\mathbb{E}_n\{U_i(\hat\eta_{BR})\}-\sqrt{n}\mathbb{E}_n\{U_i(\eta_n)\}$ as in the proof of Theorem 1, one can show $|R_2|=o_{P_n}(1)$ along the lines of the proof of Theorem 1, appealing to Assumptions 1(i), 2(i) and sparsity condition (i). Similarly, one can show that R_3 is $o_{P_n}(1)$ using the joint sparsity condition (ii) and Assumptions 2(i) and 4(ii).

Then for R_1,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i-\pi(L_i;\gamma_n)\}[m(L_i;\beta_n)-m\{L_i;\hat\beta(\hat\gamma)\}]\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i-\pi(L_i;\gamma_n)\}[m(L_i;\beta_n)-m\{L_i;\hat\beta(\gamma_n)\}]\\
&\quad+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i-\pi(L_i;\gamma_n)\}[m\{L_i;\hat\beta(\gamma_n)\}-m\{L_i;\hat\beta(\hat\gamma)\}]\\
&=R_{1a}+R_{1b}
\end{aligned}$$

One can show $|R_{1a}| = o_{P_n}(1)$ using Assumption 4(ii) and sparsity condition (i). For R_{1b},

$$
\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i - \pi(L_i;\gamma_n)\}[m\{L_i;\hat{\beta}(\gamma_n)\} - m\{L_i;\hat{\beta}(\hat{\gamma})\}]\right| \\
&\leq \sqrt{n}\,\|\mathbb{E}_n[\{A_i - \pi(L_i;\gamma_n)\}L_i]\|_\infty \left\|\hat{\beta}(\gamma_n) - \hat{\beta}(\hat{\gamma})\right\|_1 \\
&= O_{P_n}(\sqrt{\log(p \vee n)}) \\
&\quad \times O_{P_n}\left(\max_{i\leq n}|\epsilon_i|\sqrt{\frac{s_\gamma s^* \log(p\vee n)}{n}} + \frac{s^*\sqrt{\log(p\vee n)}(s_\gamma\sqrt{\log(p\vee n)}+1)}{n}\right)
\end{aligned}
$$

by (A.11) and Lemma 1. Hence under sparsity assumptions (i), (iv) and Assumption 7, $|R_{1b}| = o_{P_n}(1)$ and thus $R_1 = o_{P_n}(1)$. By then repeating Steps 2-4 from the proof of Theorem 1, the main result follows. □

Remark. We note that it follows from the above proof that sharper results are also available in linear models under misspecification than are given in Theorem 2. Namely, when either model $\mathcal{A}$ or the linear model for $\mathcal{B}$ is misspecified, ultra-sparsity is only required in the correct model. For example, if $E(A|L) = \pi(L;\gamma)$, then we require $s_\gamma^2 = o(n)$ but only $s_\beta = o(n)$ (ignoring log factors).

B Appendix B

In the figure on the following page, we describe an iterative method for γ and β (based on the reasoning in Section 5), when both are parameters indexing logistic models. In practice, one can take the penalty terms obtained via cross validation during the first iteration of the algorithm ($j = 1$) and use the same terms in subsequent iterations.

C Appendix C

Here we include some additional simulation results. Compared with the setting considered in Section 6 of the main paper, we allowed for a more dense model for the exposure A.

Algorithm 1 An algorithm for estimating η when Y is binary

```
1. Estimate $\gamma$ and $\beta$ as $\hat{\gamma}^{(0)}$ and $\hat{\beta}^{(0)}$ using (unweighted) $\ell_1$-penalized logistic regression.
   Let $\check{\gamma}^{(0)}$ and $\check{\beta}^{(0)}$ denote the refitted estimates.

2. Calculate the weights $w(L_i;\hat{\gamma}^{(0)}) = \text{expit}(\hat{\gamma}^{(0)'}L_i)\{1-\text{expit}(\hat{\gamma}^{(0)'}L_i)\}$, $w(L_i;\hat{\beta}^{(0)}) = \text{expit}(\hat{\beta}^{(0)'}L_i)\{1-\text{expit}(\hat{\beta}^{(0)'}L_i)\}$, $w(L_i;\check{\gamma}^{(0)})$ and $w(L_i;\check{\beta}^{(0)})$. Calculate the objective function
   $\check{\nu}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \log\{1+\exp(\check{\gamma}^{(0)'}L_i)\} - A_i(\check{\gamma}^{(0)'}L_i) + \log\{1+\exp(\check{\beta}^{(0)'}L_i)\} - Y_i(\check{\beta}^{(0)'}L_i)$

3. Set $j = 0$ and carry out the following recursive algorithm:
   (a) Set $j = j + 1$.
   (b) Using the initial estimates, re-estimate $\gamma$ and $\beta$ as the solutions $\hat{\gamma}^{(j)}$ and $\hat{\beta}^{(j)}$ to
       $0 = \sum_{i=1}^{n} w(L_i;\hat{\beta}^{(j-1)'})\{A_i - \text{expit}(\gamma^T L_i)\}L_i + \lambda_\gamma \delta |\gamma|^{\delta-1} \circ \text{sign}(\gamma)$
       $0 = \sum_{i=1}^{n} w(L_i;\hat{\gamma}^{(j-1)'})\{Y_i - \text{expit}(\beta^T L_i)\}L_i + \lambda_\beta \delta |\beta|^{\delta-1} \circ \text{sign}(\beta)$
       Similarly, using $w(L_i;\check{\gamma}^{(j-1)})$ and $w(L_i;\check{\beta}^{(j-1)})$, obtain the refitted $\check{\gamma}^{(j)}$ and $\check{\beta}^{(j)}$.
   (c) Re-evaluate the objective function as:
       $\check{\nu}^{(j)} = \frac{1}{n}\sum_{i=1}^{n} \left[\log\{1+\exp(\check{\gamma}^{(j)'}L_i)\} - A_i(\check{\gamma}^{(j)'}L_i)\right] w(L_i;\check{\beta}^{(j-1)'}) + \left[\log\{1+\exp(\check{\beta}^{(j)'}L_i)\} - Y_i(\check{\beta}^{(j)'}L_i)\right] w(L_i;\check{\gamma}^{(j-1)'})$
   (d) If $|\check{\nu}^{(j)} - \check{\nu}^{(j-1)}| < 0.0001$, stop the algorithm, and set $\check{\gamma}_{BR} = \check{\gamma}^{(j)}$ and $\check{\beta}_{BR} = \check{\beta}^{(j)}$.
```

Specifically $g = \left(40\frac{\log(20)}{n^{1/2}}, 40\frac{\log(19)}{n^{1/2}}, ..., 40\frac{\log(2)}{n^{1/2}}, 0_{20}, 0_{21}, 2\frac{\log(2)}{n^{1/2}}, ..., 2\frac{\log(60)}{n^{1/2}}, 0_{81}, ..., 0_p\right)$ now. Data generation process under the alternative outcome model (to evaluate the impact of fitting a misspecified linear model) was the same as in the main paper.

We further relaxed the assumption in Section 6 of homoscedastic errors. Specifically, the outcome Y_i was normally distributed conditional on L_i with mean $1+\beta_i L_i$ and standard deviation $\sqrt{\left(\frac{(1+\beta_i L_i)^2}{E_n((1+\beta_i L_i)^2)}\right)}$. In settings where we fitted a misspecified outcome model Y_i was normally distributed with mean $1 + \beta^T \left(|L_{.,[1:3]}|; L_{.,[4:p]}\right)$ and standard deviation $\sqrt{\left(\frac{\{1+\beta^T\left(|L_{.,[1:3]}|; L_{.,[4:p]}\right)\}^2}{E_n[\{1+\beta^T\left(|L_{.,[1:3]}|; L_{.,[4:p]}\right)\}^2]}\right)}$.

Table 3: Type I errors based on 1,000 replications in settings with a denser propensity score: $\Sigma = \mathrm{I}_{p\times p}$.

Correct models

Methods	$n = 200$ $p = 200$	$n = 500$ $p = 500$	$n = 200$ $p = 100$	$n = 200$ $p = 250$
Standard naïve (forced)	0.520	0.813	0.289	0.602
Standard naïve (not forced)	0.272	0.522	0.171	0.336
PDS (pre-specified)	0.526	0.779	0.523	0.517
PO (pre-specified)	0.517	0.758	0.521	0.502
PDS (CV)	0.074	0.070	0.066	0.079
PMLE-DR	0.061	0.053	0.067	0.077
BR-DR	0.054	0.061	0.064	0.054

Incorrect outcome model

Methods	$n = 200$ $p = 200$	$n = 500$ $p = 500$	$n = 200$ $p = 100$	$n = 200$ $p = 250$
Standard naïve (forced)	0.363	0.576	0.205	0.456
Standard naïve (not forced)	0.169	0.304	0.114	0.196
PDS (pre-specified)	0.344	0.611	0.342	0.310
PO (pre-specified)	0.338	0.596	0.325	0.309
PDS (CV)	0.064	0.068	0.068	0.073
PMLE-DR	0.051	0.050	0.063	0.071
BR-DR	0.055	0.052	0.069	0.041

Table 4: Type I errors based on 1,000 replications in settings with a denser propensity score: $\Sigma = [\sigma_{i,j}]_{1\leq i,j\leq p}$.

Correct models

Methods	$n = 200$ $p = 200$	$n = 500$ $p = 500$	$n = 200$ $p = 100$	$n = 200$ $p = 250$
Standard naïve (forced)	0.431	0.368	0.197	0.505
Standard naïve (not forced)	0.173	0.174	0.088	0.203
PDS (pre-specified)	0.132	0.082	0.096	0.143
PO (pre-specified)	0.102	0.071	0.078	0.094
PDS (CV)	0.056	0.071	0.064	0.068
PMLE-DR	0.050	0.062	0.046	0.062
BR-DR	0.047	0.041	0.043	0.041

Incorrect outcome model

Methods	$n = 200$ $p = 200$	$n = 500$ $p = 500$	$n = 200$ $p = 100$	$n = 200$ $p = 250$
Standard naïve (forced)	0.315	0.265	0.156	0.353
Standard naïve (not forced)	0.126	0.116	0.090	0.137
PDS (pre-specified)	0.097	0.066	0.070	0.091
PO (pre-specified)	0.073	0.066	0.056	0.074
PDS (CV)	0.063	0.060	0.060	0.068
PMLE-DR	0.055	0.063	0.048	0.064
BR-DR	0.038	0.043	0.048	0.037

References

Avagyan, V. and Vansteelandt, S. (2017). Honest data-adaptive inference for the average treatment effect under model misspecification using penalised bias-reduced double-robust estimation. *arXiv:1708.03787 [stat]*. arXiv: 1708.03787.

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.

Table 5: Type I errors based on 1,000 replications in settings with heteroscedastic errors: $\Sigma = \mathrm{I}_{p\times p}$.

Correct models

Methods	$n = 200$ $p = 200$	$n = 500$ $p = 500$	$n = 200$ $p = 100$	$n = 200$ $p = 250$
Standard naïve (forced)	0.532	0.815	0.279	0.613
Standard naïve (not forced)	0.260	0.550	0.171	0.324
PDS (pre-specified)	0.543	0.779	0.549	0.531
PO (pre-specified)	0.502	0.738	0.517	0.497
PDS (CV)	0.073	0.060	0.062	0.079
PMLE-DR	0.043	0.060	0.057	0.077
BR-DR	0.060	0.062	0.069	0.054

Incorrect outcome model

Methods	$n = 200$ $p = 200$	$n = 500$ $p = 500$	$n = 200$ $p = 100$	$n = 200$ $p = 250$
Standard naïve (forced)	0.375	0.590	0.203	0.436
Standard naïve (not forced)	0.152	0.309	0.105	0.199
PDS (pre-specified)	0.348	0.627	0.369	0.345
PO (pre-specified)	0.310	0.594	0.348	0.322
PDS (CV)	0.066	0.064	0.071	0.068
PMLE-DR	0.045	0.058	0.055	0.067
BR-DR	0.042	0.069	0.072	0.049

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

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., and Wei, Y. (2016). Post-Selection Inference for Generalized Linear Models With Many Controls. *Journal of Business & Economic Statistics*, 34(4):606–619.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer Series in Statistics. Springer Berlin Heidelberg, Berlin, Heidelberg.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Hansen, C., and Spindler, M. (2016). hdm: High-Dimensional Metrics. *The R Journal*, 8(2):185–199.

Chetverikov, D., Liao, Z., and Chernozhukov, V. (2016). On cross-validated Lasso. *arXiv:1605.02214 [math, stat]*. arXiv: 1605.02214.

De la Peña, V., Lai, T. L., and Shao, Q.-M. (2009). *Self-normalized processes: limit theory and statistical applications*. Probability and its applications. Springer, Berlin. OCLC: ocn244765605.

Dukes, O. and Vansteelandt, S. (2019). How to obtain valid tests and confidence intervals after propensity score variable selection? *Statistical Methods in Medical Research*, page 0962280219862005.

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

Finfer, S., Chittock, D. R., Su, S. Y.-S., Blair, D., and Foster, D. (2009). Intensive versus conventional glucose control in critically ill patients. *The New England Journal of Medicine*, 360(13):1283–1297.

Leeb, H. and Pötscher, B. M. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory*, 21(1):21–59.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195.

Robins, J. M. (1997). Causal Inference from Complex Longitudinal Data. In Berkane, M., editor, *Latent Variable Modeling and Applications to Causality*, Lecture Notes in Statistics, pages 69–117. Springer New York.

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.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, 56(4):931–954.

Shah, R. D. and Peters, J. (2019). The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *The Annals of Statistics*. arXiv: 1804.07203.

Tan, Z. (2019). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics*. arXiv: 1801.09817.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Van den Berghe, G., Wouters, P., Weekers, F., Verwaest, C., Bruyninckx, F., Schetz, M., Vlasselaers, D., Ferdinande, P., Lauwers, P., and Bouillon, R. (2001). Intensive Insulin

Therapy in Critically Ill Patients. *New England Journal of Medicine*, 345(19):1359–1367.

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning*. Springer Series in Statistics. Springer New York, New York, NY.

Vermeulen, K. and Vansteelandt, S. (2015). Bias-Reduced Doubly Robust Estimation. *Journal of the American Statistical Association*, 110(511):1024–1036.

von Bahr, B. and Esseen, C.-G. (1965). Inequalities for the rth absolute moment of a sum of random variables, $1 \leqq r \leqq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zheng, W. and van der Laan, M. J. (2011). Cross-Validated Targeted Minimum-Loss-Based Estimation. In *Targeted Learning*, pages 459–474. Springer New York, New York, NY.