Statistical Science
2023, Vol. 38, No. 1, 68–91
https://doi.org/10.1214/22-STS850

Double-Estimation-Friendly Inference for High-Dimensional Misspecified Models

Rajen D. Shah and Peter Bühlmann

Abstract. All models may be wrong—but that is not necessarily a problem for inference. Consider the standard t-test for the significance of a variable X for predicting response Y while controlling for p other covariates Z in a random design linear model. This yields correct asymptotic type I error control for the null hypothesis that X is conditionally independent of Y given Z under an *arbitrary* regression model of Y on (X, Z), provided that a linear regression model for X on Z holds. An analogous robustness to misspecification, which we term the "double-estimation-friendly" (DEF) property, also holds for Wald tests in generalised linear models, with some small modifications.

In this expository paper, we explore this phenomenon, and propose methodology for high-dimensional regression settings that respects the DEF property. We advocate specifying (sparse) generalised linear regression models for both Y and the covariate of interest X; our framework gives valid inference for the conditional independence null if either of these hold. In the special case where both specifications are linear, our proposal amounts to a small modification of the popular debiased Lasso test. We also investigate constructing confidence intervals for the regression coefficient of X via inverting our tests; these have coverage guarantees even in partially linear models where the contribution of Z to Y can be arbitrary. Numerical experiments demonstrate the effectiveness of the methodology.

Key words and phrases: Conditional independence, high-dimensional inference, Debiased Lasso, generalised linear models, double robustness.

1. INTRODUCTION

In this expository article, we describe a concept of insensitivity or robustness against model misspecification in linear and generalised linear models. Our starting point is the observation that inference in a misspecified linear model for the regression parameter still leads to correct statements about certain conditional independencies if the relationships between the covariates takes an appropriate form. Our aim is to popularise this main idea which, up to a few exceptions, seems to have been largely overlooked in the statistical literature and textbooks; and also to further develop the methodology and some theory for the case of high-dimensional linear and generalised linear models.

Rajen D. Shah is Professor of Statistics, Statistical Laboratory, University of Cambridge, Cambridge, United Kingdom (e-mail: r.shah@statslab.cam.ac.uk). Peter Bühlmann is Professor of Statistics, Seminar for Statistics, ETH Zurich, Zurich, Switzerland (e-mail: buehlmann@stat.math.ethz.ch).

Misspecified linear models and the t-test. We now describe a simple result (see Theorem 1) which should serve as a motivation. Consider data $(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n\times p}$ (note $\mathbf{X}$ is a vector while $\mathbf{Z}$ is a matrix) for which we have postulated a random design linear model,

$$\mathbf{Y} = \mathbf{X}\theta + \mathbf{Z}\beta^Y + \boldsymbol{\varepsilon}, \tag{1}$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$ and design matrix $(\mathbf{X}, \mathbf{Z})$ having i.i.d. Gaussian rows. The reason for distinguishing the covariate $\mathbf{X}$ from the other columns of $\mathbf{Z}$ is to focus attention on a single component of the vector of regression coefficients, namely θ. If this model is correctly specified, the t-statistic provides valid and optimal inference for θ.

Now suppose that the model (1) is misspecified and $\mathbf{Y}$ is a nonlinear function of the Gaussian covariates and a (not necessarily Gaussian) error term. Then, the standard t-test in the misspecified linear model for $\theta = 0$ still provides asymptotically valid inference for testing the null hypothesis that $\mathbf{Y}$ is conditionally independent of $\mathbf{X}$ given all other covariates $\mathbf{Z}$, in the sense that the type I error is asymptotically correctly controlled. In fact if $\mathbf{Y} = \theta\mathbf{X} + f(\mathbf{Z}, \boldsymbol{\varepsilon})$ for an essentially arbitrary measurable function f, standard confidence intervals for θ will be valid in this more general partially linear model setting. This perhaps comes as a surprise! As we will explain, it is connected to the fact that in the misspecified model, the projected parameter in the specified linear model corresponding to $\mathbf{X}$ is exactly zero when we have the conditional independence $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{Z}$; this in turn is a consequence of the regression relation between $\mathbf{X}$ and $\mathbf{Z}$ being linear due to the Gaussian assumption, that is, we have $\mathbb{E}(\mathbf{X}|\mathbf{Z}) = \mathbf{Z}\beta^X$ for some $\beta^X \in \mathbb{R}^p$.

This is just a simple motivating example, and we will relax some of the assumptions to provide a more general methodology and theory. In particular, we show that this phenomenon also extends to generalised linear models (GLMs) in the sense that if $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$, then the estimated coefficient corresponding to $\mathbf{X}$ following a generalised linear regression of $\mathbf{Y}$ on $(\mathbf{X}, \mathbf{Z})$ will have mean zero asymptotically if either the GLM is valid, or if a linear regression model for $\mathbf{X}$ on $\mathbf{Z}$ holds (and in the latter case, the GLM can be arbitrarily misspecified).

Thus, in general, basic statistical inference procedures concerning linear models and GLMs have validity beyond the restrictive parametric settings for which they are designed. Our focus in this work is studying this robustness property for which we use the term

> DEF, for "**d**ouble-**e**stimation-**f**riendly". The word "double" refers to the issue of specifying and estimating two models, and the double estimation leads then to more "friendly" results where valid inference is provided if either model is well specified.

With this term DEF we want to clearly distinguish it from double robustness, a concept whose relation to DEF is described below in Section 1.1.

A substantial part of this work considers DEF methodology in high-dimensional regression where $p \gg n$. Driven by demands from a range of application areas, but perhaps most notably genomics, high-dimensional regression has received a great deal of attention over the last two decades; see, for example, the books Bühlmann and van de Geer (2011), Hastie, Tibshirani and Wainwright (2015), Wainwright (2019) and references therein. While earlier work dealt primarily with point estimation of regression coefficients, more recently there has been a drive towards (Frequentist) uncertainty quantification, including testing for whether prespecified regression coefficients are nonzero. Much of this work has centred on the so-called *debiased Lasso* (Zhang and Zhang, 2014, van de Geer et al., 2014) which gives a construction of a coefficient estimate that unlike the more standard Lasso (Tibshirani, 1996) on which it is based, is asymptotically unbiased and normally distributed; it can therefore serve as a basis for forming confidence intervals and hypothesis tests about the unknown true coefficient vector.

The debiased Lasso has been a major advance for inference in high-dimensional settings. However, the validity of the statistical inferences it provides rests on the somewhat strong assumption that the true coefficient vector is highly sparse. For example, when testing whether $\mathbf{Y} = \mathbf{Z}\beta^Y + \boldsymbol{\varepsilon}$, that is, if the coefficient for $\mathbf{X}$ is 0, guarantees for the debiased Lasso require that $s_Y := |\{j : \beta^Y_j \neq 0\}|$ satisfies $s_Y = o(\sqrt{n}/\log(p))$. Given the preceding discussion, it is natural to ask whether the debiased Lasso is in some sense DEF. We show in this work that, with some small modifications, a version of the debiased Lasso has the DEF property. Specifically, a modified debiased Lasso gives a valid test for $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$ if either the X-model, that is the model for $\mathbf{X}$ regressed on $\mathbf{Z}$, or the Y-model is a sparse linear model. Confidence intervals derived from the debiased Lasso, however, are not DEF and do rely heavily on a sparse linear Y-model. We demonstrate that confidence intervals constructed via inverting a DEF hypothesis test can lead to much better coverage properties. While not part of the main focus of this work, we also show how a related approach may be used to construct confidence intervals for $w^T\beta^Y$, where $w \in \mathbb{R}^p$ is a possibly dense contrast vector.

In many settings, for example, when $\mathbf{X}$ is binary, a linear model for $\mathbf{X}$ on $\mathbf{Z}$ seems unlikely to hold. It would therefore be desirable to have a DEF procedure for testing the conditional independence relationship $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$ that is valid when either the Y-model or the X-model are sparse generalised linear models. For example, when both $\mathbf{Y}$ and $\mathbf{X}$ are binary we might wish to specify both models as logistic regression models. By first adapting our proposed DEF procedure to settings with linear X- and Y-models with heteroscedastic errors, we show how generalised linear models can be handled within our DEF methodology.

Below we mention some related work. We first discuss how our DEF concept and methodology relates to the literature on double robustness, and then look at other work in high-dimensional inference that bears some relation to ours here.

1.1 Relation to Double Robustness

The concept of double robustness has been developed in the context of missing values and causal effects; the latter can be seen as a missing value problem with unobserved potential outcomes. One specifies a model for the response and a model for the missingness (e.g., unobserved potential outcome), both as a function of covariates. The double robustness property is then (typically) as follows: if only one of the models is correctly specified, one can still obtain consistent estimates of average effects. This conclusion comes as a result of the bias of a doubly robust estimator taking the form of a product of estimation errors relating to each of the aforementioned models. In order for the product to tend to zero, only one of the terms in the product need tend to zero; we refer to Robins and Rotnitzky (1995), Scharfstein, Rotnitzky and Robins (1999), Kang and Schafer (2007), Cao, Tsiatis and Davidian (2009), Rotnitzky et al. (2012), among many other contributions in the literature.

While the philosophy of DEF is similar to that of double robustness in that it aims to "give the analyst two chances, instead of only one, to make a valid inference" (Bang and Robins, 2005), there are several differences. First, we are asking for valid inferential procedures, that is, hypothesis tests and confidence intervals, when either the X-model or the Y-model is misspecified. Whereas for consistency, it suffices for one of the terms composing the bias to go to zero, for our purposes this would need to vanish at a rate dominated by the variance which is typically $n^{-1/2}$. The requirement that the product of estimation error rates bounding the bias goes zero faster than $n^{-1/2}$ has been referred to as *rate double robustness* (Smucler, Rotnitzky and Robins, 2019). However, directly applying known estimation error rates for high-dimensional regression to achieve rate double robustness gives rise to procedures for hypothesis testing that require both the X and Y-models to be sparse regression models with sparsity levels $s_X, s_Y = o(\sqrt{n}/\log(p))$ (Chernozhukov et al., 2018, Shah and Peters, 2020, Dukes, Avagyan and Vansteelandt, 2020); a stronger requirement than needed for the debiased Lasso, which only assumes a sparse Y-model, and stronger still than our DEF methodology, which requires either a well-specified sparse Y-model or X-model.

In parallel work to ours, Bradic, Wager and Zhu (2019) introduce the concept of *sparsity double robustness* in the context of estimation of average treatment effects that refers to a weakening of the strong sparsity conditions imposed by rate double robustness above; however, in contrast to our DEF principle, this still requires sparse X and Y-models.

A second difference is that whereas doubly robust methods are typically semiparametrically efficient as they are often derived by considering efficient influence functions for the parameters at hand, this sort of efficiency does not necessarily arise in the more general settings covered by our idea of DEF inference. Because of these differences, we use the new terminology to distinguish the concept from double robustness.

1.2 Other Related Work

In the low-dimensional setting, early work on single-index models (Brillinger, 1983, Li and Duan, 1989, Duan and Li, 1991) has shown that OLS regression on Gaussian covariates can correctly estimate the direction of the vector of regression coefficients up to an unknown sign. This property is somewhat related to the DEF property of OLS, though deals with a rather specific form of misspecification of a linear model.

The concept of leveraging an X-model in assessing the contribution of a covariate $\mathbf{X}$ to a response $\mathbf{Y}$ while controlling for additional covariates $\mathbf{Z}$ has a long history, and the modelling of propensity scores when estimating average treatment effects is one example of this (Rosenbaum and Rubin, 1983). The work of Robins, Mark and Newey (1992) proposes to exclusively estimate an X-model in more general settings, and this idea has also appeared more recently in the model-X knockoff framework (Candès et al., 2018). The conceptual difference though is that with DEF (and also double robustness as discussed above), both the X-model and Y-model are estimated but one does not need to know which of the two models is correct.

Some recent work has looked at DEF procedures for different high-dimensional settings. Shah and Bühlmann (2018) studied a certain regularised partial correlation proposed in Ren et al. (2015); the latter work shows this test statistic is valid for testing $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$ when both the X-model and Y-models are sparse linear models, while the former shows in fact only the Y-model needs to be true for correct type I error control. As the test statistic is symmetric in $\mathbf{X}$ and $\mathbf{Y}$, we can further conclude it has the DEF property. Our proposed DEF methodology for the high-dimensional setting builds on this work, generalising it to allow for generalised linear X and Y-models. This approach is not the only possibility for DEF methodology in the high-dimensional setting, and Zhu and Bradic (2018a) look at another similar test statistic they call CorrT that delivers hypothesis tests with asymptotic type I error control in the setting where the Y-model is permitted to be a dense linear model, while the X-model must be a sparse linear model. Again, this test statistic has a DEF-like property as a consequence of its symmetry, though the dense linear model still entails some restrictions on the model class, see the discussion following Theorem 5 in Section 3.1.

Bühlmann and van de Geer (2015) consider inference with the debiased Lasso in misspecified linear models, but where the best linear predictor of the response given covariates, is sparse, and the X-model is linear. This is related to our results and methodology here, though in contrast we aim for valid inference with no sparsity requirements on one of either the X or Y-models. We note that our work also connects more generally to a thriving literature on high-dimensional inference. We refer to Dezeure et al. (2015) for a review of some of the most important developments that are related to our work here.

1.3 Organisation of the Paper

The rest of the paper is organised as follows. In Section 2, we study the low-dimensional setting and formally set out the DEF properties of standard inference procedures for linear and generalised linear models. We then turn to the high-dimensional setting and study in Section 3.1 the case where we allow either the regression model for $\mathbf{Y}$ on $\mathbf{Z}$ or that for $\mathbf{X}$ on $\mathbf{Z}$ to be linear. In Section 3.2, we detail the construction of confidence intervals in partially linear high-dimensional models using the classical duality between confidence regions and hypothesis tests. We then study the setting where the models of $\mathbf{Y}$ and $\mathbf{X}$ are generalised linear models. Some numerical experiments are presented in Section 4 and we conclude with a discussion in Section 5. The Appendix contains proofs omitted in the main text, a construction for confidence regions for $w^T\beta^Y$ based on the methodology set out in Section 3.1, a description of how square-root Lasso solutions may be computed given regular Lasso solutions, and some additional numerical experiments.

2. LOW DIMENSIONS

Recall that $(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n\times p}$ and we are interested in the relationship between $\mathbf{Y}$ and $\mathbf{X}$, and specifically testing the conditional independence $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$. We first study the DEF property of the standard t-statistic in the linear model, before turning to generalised linear models in Section 2.2.

2.1 Linear Models

Let $\tilde{\mathbf{Z}} := (\mathbf{X}, \mathbf{Z}) \in \mathbb{R}^{n\times(p+1)}$ and let $(\hat{\theta}, \hat{\beta}^Y) \in \mathbb{R} \times \mathbb{R}^p$ be the regression coefficient vector from an OLS regression of $\mathbf{Y}$ on $\tilde{\mathbf{Z}}$. Further let $\tilde{\mathbf{P}}$ and $\mathbf{P}$ be the orthogonal projections on to $\tilde{\mathbf{Z}}$ and $\mathbf{Z}$, respectively. Also define $\tilde{\sigma}^2 = \|\mathbf{Y} - \tilde{\mathbf{P}}\mathbf{Y}\|_2^2/(n-p-1)$. The usual t-statistic for testing the significance of variable $\mathbf{X}$ is given by $T_{\text{OLS}} := \hat{\theta}/\sqrt{\{(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\}_{11}\tilde{\sigma}^2}$. Denote by $\mathbf{R} := (\mathbf{I}-\mathbf{P})\mathbf{X}$ the residuals from regressing $\mathbf{X}$ on $\mathbf{Z}$.

Consider the following set of assumptions.

(Y1) We have $\mathbf{Y} = \mathbf{Z}\beta^Y + \boldsymbol{\varepsilon}$ with $\mathbb{E}(\varepsilon_i|\mathbf{Z}) = 0$, $\mathbb{E}(\varepsilon_i^2|\mathbf{Z}) = \sigma^2 > 0$, $\mathbb{E}(|\varepsilon_i|^{2+\delta}|\mathbf{Z}) < M$ for some constants $M, \delta, \sigma^2 > 0$, and the ε_i are independent conditional on $\mathbf{Z}$.

(Y2) We have $\mathbb{P}(\mathbf{R} = \mathbf{0}) \to 0$ and for some $\delta > 0$,

$$A_n := \begin{cases} \dfrac{1}{\|\mathbf{R}\|_2^{2+\delta}} \displaystyle\sum_{i=1}^n |R_i|^{2+\delta} & \text{if } \mathbf{R} \neq \mathbf{0}, \\ 0 & \text{if } \mathbf{R} = \mathbf{0}, \end{cases} \tag{2}$$

satisfies $A_n \xrightarrow{p} 0$.

Condition (Y1) formalises the particular form of the linear model we assume here (under the null hypothesis), which includes the normal linear model, for example, but is rather more general. Condition (Y2) enforces that no individual residual is too extreme. Indeed, it is sufficient that $\max_i R_i/\|\mathbf{R}\|_2 \xrightarrow{p} 0$. This would typically be satisfied if the rows of $(\mathbf{X}, \mathbf{Z})$ were i.i.d., for example, but is much weaker. We also introduce the following.

(Xj) The equivalent of (Yj) above but with $\mathbf{X}$ replaced with $\mathbf{Y}$ and vice versa, for $j \in \{1, 2\}$.

The theorem below shows that T_{OLS} has a DEF property.

THEOREM 1. *Suppose $p/n \to 0$. If either* (X1) *and* (X2) *or* (Y1) *and* (Y2) *hold, then under the null hypothesis that $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$, we have $T_{OLS} \xrightarrow{d} \mathcal{N}(0, 1)$.*

The result may be viewed as a consequence of the close relationship between the t-statistic above and the partial correlation

$$\hat{\rho} := \frac{\mathbf{X}^T(\mathbf{I}-\mathbf{P})\mathbf{Y}}{\|(\mathbf{I}-\mathbf{P})\mathbf{X}\|_2\|(\mathbf{I}-\mathbf{P})\mathbf{Y}\|_2}.$$

This can also be interpreted as a test statistic based on a score test for $\theta = 0$ when it is assumed the errors are Gaussian. One can verify that

$$T_{\text{OLS}} = \sqrt{n-p-1}\frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}, \tag{3}$$

so the distributional result for T_{OLS} follows from $\sqrt{n}\hat{\rho} \xrightarrow{d} \mathcal{N}(0, 1)$. As $\hat{\rho}$ is symmetric in $\mathbf{X}$ and $\mathbf{Y}$ it is unsurprising that this has a DEF property. Indeed, the DEF approach suited to the high-dimensional setting we present in Section 3, is based on a certain regularised partial correlation.

We also remark that under the assumption that $\mathbf{Y} = \mathbf{Z}\beta^Y + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, we have the exact distributional relationship

$$\hat{\rho}\sqrt{\frac{n-p-1}{1-\hat{\rho}^2}} \sim t_{n-p-1}.$$

The symmetry of this statistic in $\mathbf{X}$ and $\mathbf{Y}$ means that the distributional result also holds when an analogous normal linear model for $\mathbf{X}$ on $\mathbf{Z}$ holds. This may be used to yield a DEF test for conditional independence with exact type I error control in finite samples, under these additional Gaussianity assumptions.

EXAMPLE 1. The famous diabetes dataset of Efron et al. (2004) contains $p = 10$ predictors (age, sex, BMI, etc.) measured for $n = 442$ patients. We take these covariates as our matrix $\mathbf{Z} \in \mathbb{R}^{n\times p}$ and generate an additional predictor $\mathbf{X} \in \mathbb{R}^n$ with entries $X_i = \sum_j Z_{ij} + \varepsilon_i^X$ where $\varepsilon_i^X + 1 \overset{\text{i.i.d.}}{\sim} \text{Exp}(1)$. We ignore the original response of the design matrix and generate a new response $\mathbf{Y} \in \mathbb{R}^n$ that depends nonlinearly on $\mathbf{Z}$ through $Y_i = \eta_i\zeta_i$ where $\zeta_i \overset{\text{i.i.d.}}{\sim} \chi_1^2$ and

$$\eta_i = \sum_{j,k} \frac{\exp(Z_{ij}Z_{ik})}{1+\exp(Z_{ij}Z_{ik})}. \tag{4}$$

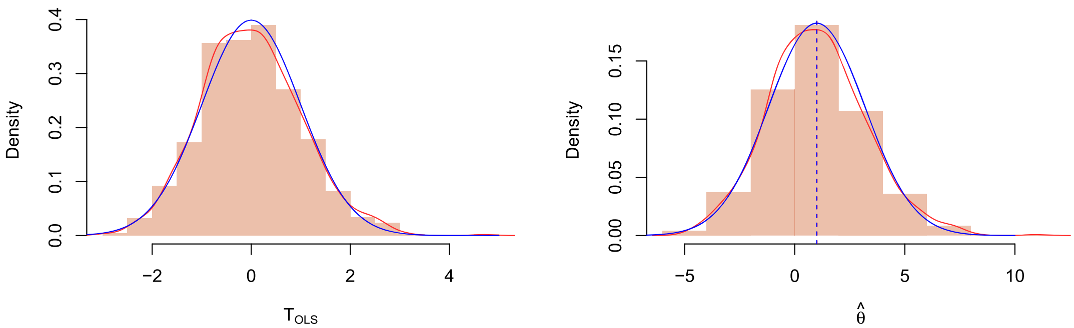

FIG. 1. *Histograms of T_{OLS} (left plot) and $\hat{\theta}$ (right plot) for the setup described in Example 1. The red curves are kernel density estimates. We see close agreement with the theoretical normal density (blue curves). The vertical dashed red lines and blue lines in the right plot are the empirical and theoretical means respectively; their proximity in this example makes them hard distinguish visually.*

In this setup, we then have $Y_i \perp\!\!\!\perp X_i|Z_i$ and the X-model is a linear regression model. Theorem 1 suggests that the t-statistic T_{OLS} corresponding to $\mathbf{X}$ should have a distribution well approximated by a standard normal. The left panel of Figure 1 plots the histogram of T_{OLS} computed on 500 simulated datasets generated through the construction above. We do indeed see a close agreement with a standard normal density, verifying the theoretical result. The right panel plots the coefficient estimate $\hat{\theta}$ corresponding to $\mathbf{X}$ when the equation for $\mathbf{Y}$ has $\mathbf{X}$ added (i.e., the null hypothesis does not hold). It is easy to see that compared to the previous setup, this coefficient will be shifted by 1, and hence asymptotically should have a Gaussian distribution centred on 1, as we observe in the plot.

2.2 Generalised Linear Models

It is well known that maximum likelihood estimators under misspecification are, given regularity conditions, asymptotically normal about a parameter vector corresponding to the model closest to the ground truth in terms of Kullback–Leibler divergence (Huber, 1967, White, 1982). This fact is typically used as reassurance that while all statistical models are wrong, provided one is working with a model that is a good enough approximation to the truth, maximum likelihood estimation is nevertheless useful. However, as we shall explain, in terms of conditional independence testing, maximum likelihood estimation of generalised linear models can form the basis of a valid test even under severe misspecification.

In this section, we will assume that the rows (X_i, Y_i, Z_i) of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in \mathbb{R}^{n\times(2+p)}$ are independent copies of the random triple $(X, Y, Z) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^p$. Consider a generalised linear model relating response vector $\mathbf{Y}$ to covariates $(\mathbf{X}, \mathbf{Z})$, or more generally, a model where the density $f_{Y|X,Z}$ of Y conditional on (X, Z) (with respect to a measure μ) takes the form

$$f_{Y|X,Z}(y|x, z) = L(x\theta + z^T\beta^Y; y) \tag{5}$$

for $(\theta, \beta^Y) \in \Theta \subseteq \mathbb{R}^{p+1}$. We will assume that L is twice differentiable in its first argument. Define $\ell := \log L$ and $U := \ell'$ where the prime denotes a derivative with respect to the first argument; we will typically suppress the dependence of U on its second argument y for simplicity. Under regularity conditions, the maximum likelihood estimator

$$(\hat{\theta}, \hat{\beta}^Y) := \underset{(t,\beta)\in\Theta}{\arg\min} -\sum_{i=1}^{n} \ell(X_i t + Z_i^T\beta; Y_i)$$

is asymptotically normal centred on (θ^*, β^*), which solve for $(t, \beta) \in \Theta$ the score equations

$$\mathbb{E}\{XU(Xt + Z^T\beta)\} = 0, \tag{6}$$

$$\mathbb{E}\{ZU(Xt + Z^T\beta)\} = 0. \tag{7}$$

When (5) holds (which includes as a special case when a generalised linear model is correct), under regularity conditions, we will have $(\theta^*, \beta^*) = (\theta, \beta^Y)$. In order for inference based on $\hat{\theta}$ to provide useful information concerning the conditional independence $X \perp\!\!\!\perp Y|Z$ when (5) does not hold, we would like $\theta^* = 0$ in the case of conditional independence. Analogously to the case with linear models discussed in the previous section, we have that regardless of the form of the Y-model, provided the X-model is linear, it holds that $\theta^* = 0$; here though we additionally require that the solution to (6) and (7) is unique to derive this conclusion.

THEOREM 2. *Suppose $X \perp\!\!\!\perp Y|Z$. Let $\beta^\dagger \in \mathbb{R}^p$ maximise the expected log-likelihood $\mathbb{E}\ell(Z^T\beta; Y)$ over β. Assume regularity conditions set out in Section A.2.1 of the Appendix. Suppose that either the Y-model is well specified so (5) holds, or the X-model is linear so $\mathbb{E}(X|Z) = Z^T\beta^X$. Then $(t, \beta) = (0, \beta^\dagger)$ satisfies the score equations (6), (7).*

Theorem 2 shows that under the X-model, the parameter corresponding to the projection of the truth on to the purported Y-model is 0 under conditional independence. A standard Wald test for whether $\theta = 0$ will, however, not be valid under general misspecification as the asymptotic variance of $\hat{\theta}$ will not necessarily be given by the $(1, 1)$

entry of the inverse Fisher information matrix for (θ, β^Y). Indeed, it is well known that, under regularity conditions, the variance of $\hat{\theta}$ is given by the sandwich formula

$$\sqrt{n}\left(\begin{pmatrix}\hat{\theta}\\ \hat{\beta}^Y\end{pmatrix} - \begin{pmatrix}\theta^*\\ \beta^*\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}(0, H^{-1}VH^{-1}), \tag{8}$$

where V is the covariance matrix of the derivative of $\ell(X\theta + Z^T\beta; Y)$ with respect to (θ, β) evaluated at (θ^*, β^*) (satisfying the score equations (6), (7)) and H is the negative expectation of the corresponding Hessian matrix:

$$V := \mathbb{E}\left(\begin{pmatrix}X\\ Z\end{pmatrix}\begin{pmatrix}X\\ Z\end{pmatrix}^T U^2(X\theta^* + Z^T\beta^*)\right),$$

$$H := -\mathbb{E}\left(\begin{pmatrix}X\\ Z\end{pmatrix}\begin{pmatrix}X\\ Z\end{pmatrix}^T U'(X\theta^* + Z^T\beta^*)\right).$$

The matrices V and H may be estimated individually using the data via several methods (MacKinnon and White, 1985). However, if either the X-model is a homoscedastic linear model, or the Y-model holds, some simplifications are possible, as the result below describes.

THEOREM 3. *Suppose $X \perp\!\!\!\perp Y|Z$ and assume regularity conditions set out in Section A.3.1 of the Appendix. Suppose either (5) holds with $U = \ell'$, or $\mathbb{E}(X|Z) = Z^T\beta^X$. We additionally assume* $\mathrm{Var}(X|Z) = \mathrm{Var}(X)$ *in the latter case. Then we have*

$$(H^{-1}VH^{-1})_{11} = -(H^{-1})_{11}\frac{\mathbb{E}\{U^2(Z^T\beta^*)\}}{\mathbb{E}\{U'(Z^T\beta^*)\}}.$$

The correction factor for the usual inverse of the Fisher information may be readily estimated by

$$\hat{C}_1 := -\frac{\sum_{i=1}^n U^2(Z_i^T\hat{\beta}^Y)}{\sum_{i=1}^n U'(Z_i^T\hat{\beta}^Y)}, \tag{9}$$

or indeed a variant of the above with $Z_i^T\hat{\beta}^Y$ replaced everywhere by $X_i\hat{\theta} + Z_i^T\hat{\beta}^Y$ which we will refer to as $\hat{C}_2$. Writing $\hat{H}$ for the empirical version of H,

$$\hat{H} := -\frac{1}{n}\sum_{i=1}^n \begin{pmatrix}X_i\\ Z_i\end{pmatrix}\begin{pmatrix}X_i\\ Z_i\end{pmatrix}^T U'(X_i\hat{\theta} + Z_i^T\hat{\beta}^Y),$$

we may define for $j = 1, 2$, the test statistics

$$T_{\mathrm{GLM},j} := \frac{\sqrt{n}\hat{\theta}}{\sqrt{\hat{C}_j(\hat{H}^{-1})_{11}}}.$$

Putting together Theorems 2 and 3 we have the following result.

THEOREM 4. *Suppose $X \perp\!\!\!\perp Y|Z$ and (8) holds where (θ^*, β^*) is the unique solution in (t, β) to (6) and (7). Assume that $\hat{H} \xrightarrow{p} H$ with H positive definite and assume the regularity conditions set out in Section A.4.1. Suppose that either the Y-model is well specified so (5) holds, or the X-model is linear so $\mathbb{E}(X|Z) = Z^T\beta^X$. Then for $j = 1, 2$ we have*

$$T_{GLM,j} \xrightarrow{d} \mathcal{N}(0, 1).$$

EXAMPLE 2. We use a similar setup to Example 1 but here generate the response $\mathbf{Y} \in \mathbb{R}^n$ according to $Y_i \overset{\text{i.i.d.}}{\sim} \mathrm{Poisson}(\mu_i)$ with

$$\log(\mu_i) = a_1\sum_j Z_{ij} + \sigma a_2\eta_i$$

with $\sigma \in \{0, 2, 4\}$ and factors a_1 and a_2 chosen so the maximum absolute value over i of the two terms above is 3 to ensure $\mathbb{E}Y_i$ does not take values that are too large. We consider testing the significance of the variable $\mathbf{X}$ using (a) standard Wald-based p-values assuming a Poisson log-linear model, (b) the equivalent using a quasi-Poisson likelihood and (c) using $T_{\mathrm{GLM},2}$. We plot in Figure 2 the empirical distribution functions of the p-values

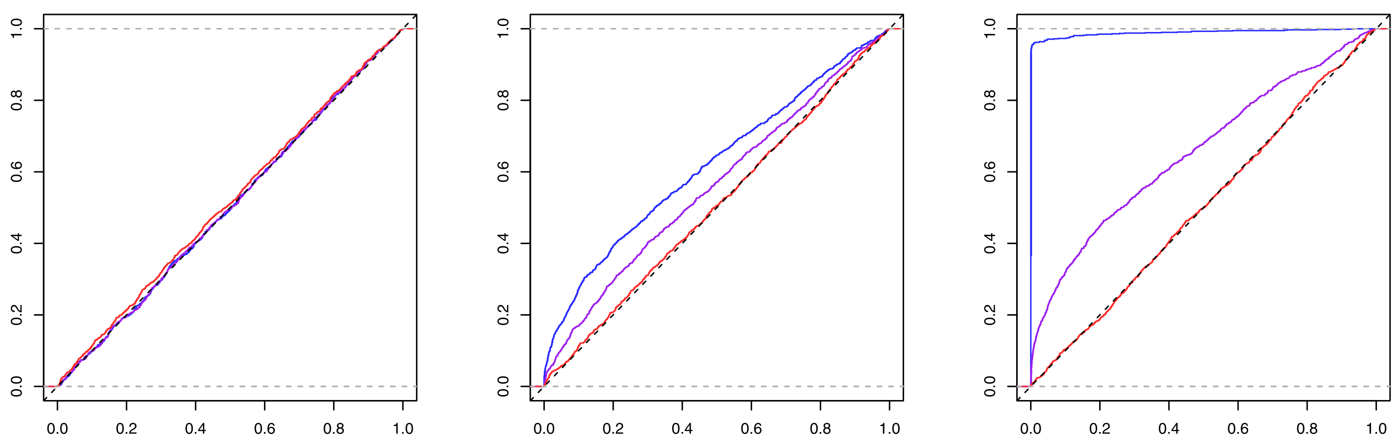

FIG. 2. *Empirical distribution functions of p-values from the simulation setups of Example 2 with $\sigma = 0, 2, 4$ from left to right. Blue, purple and red curves correspond to naive p-values (a), quasi-likelihood-based p-values (b) and p-values based on $T_{GLM,2}$ (c), respectively. Type I errors of the resulting tests are well controlled for (c), but (a) and (b) fail to maintain nominal levels under misspecification.*

observed over 500 replicates of the three settings determined by σ. As expected, for the well-specified case with $\sigma = 0$ all p-values are roughly uniformly distributed. However, for increasing levels of misspecification, the standard p-values (a) tend to be more anti-conservative, a phenomenon which occurs to a lesser extent for the quasi-likelihood-based p-values (b). The correction factor (c) ensures that p-values corresponding to $T_{\mathrm{GLM},2}$ are approximately uniform across all of the settings considered.

3. HIGH DIMENSIONS

We have seen in the previous section how classical linear and generalised linear model inferential tools have the DEF property. In the case of linear models, this could be deduced from the similarity of the standard t-statistic to partial correlation. For generalised linear models, the DEF property is perhaps more surprising. Our analysis first used the fact that maximum likelihood converges to a projection of the ground truth, and then considered the projected parameters themselves. There is, however, no analogue of the classical Huber–White results on the properties of maximum likelihood in nonlinear models under misspecification available for high-dimensional estimators. Our approach to DEF inference in high-dimensional settings will therefore be based around versions of partial correlation. We first study linear models before turning to the case of high-dimensional generalised linear models.

3.1 Linear Models

One of the most popular methods for testing the significance of predictors in high-dimensional regression problems is the so-called debiased Lasso (Zhang and Zhang, 2014). We begin by discussing this approach, in order to motivate our DEF methodology.

The debiased Lasso works as follows: first we form estimates $(\hat{\theta}, \check{\beta}^Y)$ through a Lasso regression of $\mathbf{Y}$ on $(\mathbf{X}, \mathbf{Z})$, and also conduct a Lasso regression of $\mathbf{X}$ on $\mathbf{Z}$ to give a coefficient estimate $\hat{\beta}^X$. There are a variety of choices of tuning parameters for each of these regressions; to ensure that these tuning parameters do not depend on the noise variances of the respective regressions, we may use a particular parametrisation of the Lasso known as the square-root Lasso regressions (Belloni, Chernozhukov and Wang, 2011, Sun and Zhang, 2012):

$$(\hat{\theta}, \check{\beta}^Y) := \operatorname*{arg\,min}_{(t,\beta)\in\mathbb{R}^{1+p}} \{\|\mathbf{Y} - \mathbf{X}t - \mathbf{Z}\beta\|_2/\sqrt{n} + \lambda_Y\|\beta\|_1\}, \tag{10}$$

$$\hat{\beta}^X := \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} \{\|\mathbf{X} - \mathbf{Z}\beta\|_2/\sqrt{n} + \lambda_X\|\beta\|_1\}.$$

Here we may take $\lambda_X = \lambda_Y = A\sqrt{2\log(p)/n}$ for $A > 1$. Note that we have denoted the estimate of the coefficient vector for X as $\check{\beta}^Y$ in order to distinguish it from $\hat{\beta}^Y$ introduced in (14) below. The square-root Lasso may be computed easily using standard software that computes regular Lasso solutions: see Section B in the Appendix.

We then construct a test statistic T_{DB} for assessing the conditional independence $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$ as follows:

$$T_{\mathrm{DB}} := \sqrt{n}\frac{(\mathbf{Y} - \mathbf{Z}\check{\beta}^Y)^T(\mathbf{X} - \mathbf{Z}\hat{\beta}^X)}{\|\mathbf{Y} - \hat{\theta}\mathbf{X} - \mathbf{Z}\check{\beta}^Y\|_2\|\mathbf{X} - \mathbf{Z}\hat{\beta}^X\|_2}.$$

When the Y-model is a sparse linear model so $\mathbf{Y} = \mathbf{Z}\beta^Y + \boldsymbol{\varepsilon}$ with β^Y sparse and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$, we have that $T \xrightarrow{d} \mathcal{N}(0, 1)$ as we now outline. Let us write

$$\mathbf{R} := \mathbf{X} - \mathbf{Z}\hat{\beta}^X,$$
$$\hat{\sigma} := \|\mathbf{Y} - \hat{\theta}\mathbf{X} - \mathbf{Z}\check{\beta}^Y\|_2/\sqrt{n}.$$

A consequence of the stationarity conditions (the so-called KKT conditions) for the optimisation problem defining $\hat{\beta}^X$ is that, provided $\mathbf{R} \neq \mathbf{0}$,

$$\frac{1}{\sqrt{n}}\|\mathbf{Z}^T\mathbf{R}\|_\infty/\|\mathbf{R}\|_2 \le \lambda_X. \tag{11}$$

We may thus decompose T_{DB} as follows:

$$T_{\mathrm{DB}} = \frac{1}{\hat{\sigma}}\frac{\mathbf{R}^T}{\|\mathbf{R}\|_2}\boldsymbol{\varepsilon} + \frac{1}{\hat{\sigma}}(\beta^Y - \check{\beta}^Y)^T\mathbf{Z}^T\frac{\mathbf{R}}{\|\mathbf{R}\|_2} =: \text{(i)} + \text{(ii)}.$$

Conditioning on $\mathbf{R}$, $\mathbf{R}^T\boldsymbol{\varepsilon}/\|\mathbf{R}\|_2$ is a weighted sum of the independent and identically distributed ε_i, and thus will have an asymptotic Gaussian distribution under weak conditions on $\mathbf{R}$; in fact if the ε_i are Gaussian themselves we will have $\mathbf{R}^T\boldsymbol{\varepsilon}/\|\mathbf{R}\|_2|\mathbf{R} \sim \mathcal{N}(0, \sigma^2)$ exactly, and of course the unconditional distribution will hence also be Gaussian. If $\hat{\sigma} \xrightarrow{p} \sigma$, then by Slutsky's lemma we will have that (i) converges in distribution to a standard normal. In order to guarantee this, we may appeal to known results about the square-root Lasso (Sun and Zhang, 2012). These rest on a compatibility factor ϕ^2 (van de Geer and Bühlmann, 2009) being bounded away from zero:

$$\phi^2 := \inf_{\substack{(t,\beta)\in\mathbb{R}^{1+p} \\ |t|+\|\beta_{S_Y^c}\|_1\le 3\|\beta_{S_Y}\|_1\neq 0}} \frac{\|\mathbf{X}t + \mathbf{Z}\beta\|/n}{\|\beta_{S_Y}\|_1/s_Y}; \tag{12}$$

here $S_Y := \{j : \beta_j^Y \neq 0\}$, $s_Y := |S_Y|$ and we have used the notation that for any vector $b \in \mathbb{R}^p$ and set $S \subseteq \{1, \dots, p\}$, $b_S \in \mathbb{R}^{|S|}$ is the subvector of b composed of those components of b indexed by S. Roughly speaking, designs with large compatibility factors cannot have very highly correlated columns. Provided $\phi^2 \gtrsim 1$, we have $\hat{\sigma} \xrightarrow{p} \sigma$ and also $\|\check{\beta}^Y - \beta^Y\|_1 \lesssim s_Y\sqrt{\log(p)/n}$ with high probability, when $\lambda_Y \asymp \sqrt{\log(p)/n}$ (van de Geer, 2016). This second property may be used to bound (ii) via

$$\frac{|(\beta^Y - \check{\beta}^Y)^T\mathbf{Z}^T\mathbf{R}|}{\|\mathbf{R}\|_2} \lesssim \lambda_X s_Y\sqrt{\log(p)} \lesssim s_Y\log(p)/\sqrt{n}, \tag{13}$$

where we have used Hölder's inequality and (11). Thus, in an asymptotic regime where $s_Y \log(p)/\sqrt{n} \to 0$, Slutsky's lemma gives us that $T_{\text{DB}} \xrightarrow{d} \mathcal{N}(0, 1)$.

Note that essentially no assumptions regarding a regression model for $\mathbf{X}$ on $\mathbf{Z}$ are required here; the only purpose of the square-root Lasso regression producing $\hat{\beta}^X$ is to construct the vector of residuals $\mathbf{R}$. This latter quantity may be regarded as a version of predictor $\mathbf{X}$ modified to be almost orthogonal to the remaining covariates $\mathbf{Z}$ (11) such that when normalised, the dot product with the bias term $\mathbf{Z}(\beta^Y - \check{\beta}^Y)$ is well controlled (13). Although this orthogonality comes free as a by-product of the square-root Lasso, we have, however, tacitly assumed $\mathbf{R} \neq \mathbf{0}$ to arrive at (11). If $\mathbf{R} = \mathbf{0}$ (which we have yet to observe in practice) we can simply agree to accept the null of conditional independence, so this poses no problem for type I error control. We note that the same sort of orthogonality argument may not go through for a regular Lasso estimator with tuning parameter chosen by cross-validation, for example, as control of the LHS of (11) with no assumptions on the model would be very challenging. However, empirically, we have observed that the cross-validated Lasso performs similarly to the square-root Lasso here.

Now consider the case where the X-model is a sparse linear model. While we will have control of $\|\beta^X - \hat{\beta}^X\|_1$, the equivalent of (11) with residuals $\mathbf{R}$ replaced by $\mathbf{Y} - \mathbf{Z}\check{\beta}^Y$ will not hold in general. The issue is that the latter quantity is not equal to the residuals from the Y-regression unless $\hat{\theta} = 0$. Thus the debiased Lasso is not quite DEF in that it can be sensitive to misspecification of the Y-model.

There are several options for how to restore the DEF property in this setting, but one that is particularly simple involves enforcing that $\hat{\theta} = 0$, that is, setting $\hat{\beta}^Y$ to be coefficients from a regression of $\mathbf{Y}$ on $\mathbf{Z}$ rather than the augmented design $(\mathbf{X}, \mathbf{Z})$:

$$\hat{\beta}^Y := \underset{b \in \mathbb{R}^p}{\arg\min}\{\|\mathbf{Y} - \mathbf{Z}b\|_2/\sqrt{n} + \lambda_Y \|b\|_1\}; \tag{14}$$

note this differs from the definition in (10). The resulting test statistic takes the form of a regularised partial correlation:

$$\begin{aligned} T_{\text{DEF}} &:= T_{\text{DEF}}(\mathbf{Y}, \mathbf{X}) \\ &:= \sqrt{n}\frac{(\mathbf{Y} - \mathbf{Z}\hat{\beta}^Y)^T(\mathbf{X} - \mathbf{Z}\hat{\beta}^X)}{\|\mathbf{Y} - \mathbf{Z}\hat{\beta}^Y\|_2\|\mathbf{X} - \mathbf{Z}\hat{\beta}^X\|_2}; \end{aligned} \tag{15}$$

note the inclusion of the notation $T_{\text{DEF}}(\mathbf{Y}, \mathbf{X})$ making the dependence of the test statistic on $\mathbf{Y}$ and $\mathbf{X}$ is included here for use later in Section 3.2. In the unlikely case that the denominator defining T_{DEF} above is zero, so one of the square-root Lasso solutions is degenerate, we will set $T_{\text{DEF}} = 0$; we have never observed this degeneracy to occur in any of the numerical experiments conducted. The test statistic (15) above was first studied in Ren et al. (2015) in the context of Gaussian graphical model estimation where asymptotic normality was shown when both the X-model and Y-model are sparse. The work of Shah and Bühlmann (2018) extended this result to show that the same conclusion holds when only the Y-model holds, and hence by symmetry of the test statistic, that it has the DEF property. Below we state a variant of the latter result that allows for non-Gaussian errors.

In the case that (only) the Y-model holds, we will need to assume in addition to (Y1) and (Y2) with $\mathbf{R} = \mathbf{X} - \mathbf{Z}\hat{\beta}^X$, the following conditions.

(Y3) Defining $S_Y := \{j : \beta_j^Y \neq 0\}$ and $s_Y := |S_Y|$, we have $s_Y \log(p)/\sqrt{n} \to 0$.

(Y4) $\|\hat{\beta}^Y - \beta^Y\|_1 = O_{\mathbb{P}}(s_Y\sqrt{\log(p)/n})$.

(Y5) $\|\mathbf{Y} - \mathbf{Z}\hat{\beta}^Y\|_2^2/n \xrightarrow{p} \sigma^2$.

Note that, as in the low-dimensional case, the only assumption placed on the conditional distribution of $\mathbf{X}$ given $\mathbf{Z}$ is (Y2), with $\mathbf{R} = \mathbf{X} - \mathbf{Z}\hat{\beta}^X$. This would be satisfied if we had a sparse linear X-model, but such an assumption is very far from necessary in order for (Y2) to hold. Furthermore, as shown in Shah and Bühlmann (2018), this is not necessary when the errors $\boldsymbol{\varepsilon}$ for the Y-model are Gaussian. We also introduce, in addition to (X1) and (X2) with $\mathbf{R} = \mathbf{Y} - \mathbf{Z}\hat{\beta}^Y$, the following assumptions that are relevant when the X-model holds.

(Xj) As (Yj) above, but with X and $\mathbf{X}$ interchanged with Y and $\mathbf{Y}$ everywhere, for $j \in \{3, 4, 5\}$.

We have the following result.

THEOREM 5. *Let* $\lambda_X = \lambda_Y = A\sqrt{2\log(p)/n}$ *for some* $A > 1$. *Assume that either* (Y1)–(Y5) *or* (X1)–(X5) *hold. Then under the null hypothesis that* $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$, *test statistic* T_{DEF} *defined according to* (15) *satisfies* $T_{DEF} \xrightarrow{d} \mathcal{N}(0, 1)$.

Similarly to the case with the debiased Lasso, under an alternative where $\mathbf{Y} = \mathbf{X}\theta + \mathbf{Z}\beta^Y + \boldsymbol{\varepsilon}$, if a sparse linear X-model also holds, T_{DEF} (15) has power tending to 1 when $\sqrt{n}\theta \to \infty$. We refer the reader to Ren et al. (2015) and Shah and Bühlmann (2018) for further details.

The DEF version of the debiased Lasso bears some similarities to the CorrT test developed and studied in Zhu and Bradic (2018a). However, whereas the latter relies on estimating β^Y and β^X via a family of linear programs, the DEF statistic presented here can be calculated using standard software for computing Lasso solutions such as `glmnet` (Friedman, Hastie and Tibshirani, 2010). We note further that whereas Theorem 5 only requires the weak condition that no residual from the regression relating to the misspecified is too extreme (and no condition on the residuals when the errors in the true model are Gaussian), the corresponding result (Theorem 2) in Zhu and

Bradic (2018a) requires the misspecified model to nevertheless be a linear model with the coefficient vector having bounded ℓ_2-norm. Furthermore, the sparsity condition $s = o(\sqrt{n}/(\log p)^{5/2})$ is assumed, where s is the sparsity of the coefficient vector in the well-specified model, compared to our requirement of $s = o(\sqrt{n}/\log p)$. On the other hand, the CorrT test accommodates heteroscedastic errors whereas one would need to modify our statistic to

$$\sqrt{n}\frac{\frac{1}{n}(\mathbf{R}^Y)^T\mathbf{R}^X}{\frac{1}{n}\sum_{i=1}^n(\mathbf{R}_i^Y)^2(\mathbf{R}_i^X)^2 - (\frac{1}{n}(\mathbf{R}^Y)^T\mathbf{R}^X)^2},$$

where

$$\mathbf{R}^Y := \mathbf{Y} - \mathbf{Z}\hat{\beta}^Y \quad \text{and} \quad \mathbf{R}^X := \mathbf{X} - \mathbf{Z}\hat{\beta}^X$$

in order to achieve this; see Shah and Peters (2020) which uses the denominator above more generally in nonparametric models.

3.2 Confidence Intervals via Inverting Tests

Thus far we have only discussed testing, but using the DEF statistic (15), it is straightforward to obtain confidence intervals for a parameter θ in the partially linear model

$$\mathbf{Y} = \mathbf{X}\theta + f(\mathbf{Z}, \boldsymbol{\varepsilon}), \tag{16}$$

where $\boldsymbol{\varepsilon} \perp\!\!\!\perp \mathbf{X}|\mathbf{Z}$ and $f : \mathbb{R}^{n\times p} \times \mathbb{R}^n \to \mathbb{R}^n$ under the following conditions: either $f(\mathbf{Z}, \boldsymbol{\varepsilon}) = \mathbf{Z}\beta^Y + \boldsymbol{\varepsilon}$, or a sparse linear X-model holds. Our approach for constructing a confidence region for θ utilises the well-known duality between confidence intervals and hypothesis tests; specifically we invert the DEF test, noting that under (16), we have $\mathbf{Y} - \mathbf{X}\theta \perp\!\!\!\perp \mathbf{X}|\mathbf{Z}$. We first compute test statistic

$$T_{\mathrm{DEF},t} := T_{\mathrm{DEF}}(\mathbf{Y} - \mathbf{X}t, \mathbf{X}), \tag{17}$$

that is, we subtract t times $\mathbf{X}$ from $\mathbf{Y}$ and compute the usual DEF test statistic. Then we form a $1-\alpha$ confidence region R_α via

$$R_\alpha := \{t \in \mathbb{R} : |T_{\mathrm{DEF},t}| \geq z_\alpha\},$$

where z_α is the upper $\alpha/2$ quantile of a standard normal distribution. As a consequence of Theorem 5, this confidence region has the following asymptotic validity.

COROLLARY 6. *Suppose the partially linear model* (16) *holds with* $\boldsymbol{\varepsilon} \perp\!\!\!\perp \mathbf{X}|\mathbf{Z}$ *and let* $\lambda_X = \lambda_Y = A\sqrt{2\log(p)/n}$ *for some* $A > 1$. *Suppose the assumptions of Theorem* 5 *hold with* $\mathbf{Y}$ *replaced by* $\mathbf{Y} - \mathbf{X}\theta$, *that is, in particular either* $f(\mathbf{Z}, \boldsymbol{\varepsilon}) = \mathbf{Z}\beta^Y + \boldsymbol{\varepsilon}$, *or a sparse linear* X-*model holds. Then for any* $\alpha \in (0, 1)$,

$$\mathbb{P}(\theta \in R_\alpha) = \mathbb{P}(|T_{DEF,\theta}| \geq z_\alpha) \to 1 - \alpha.$$

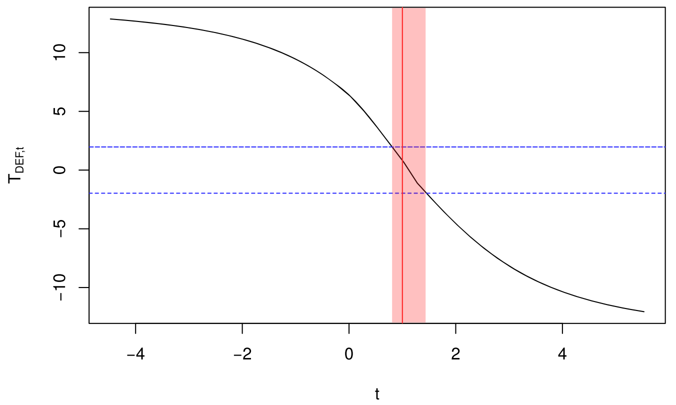

FIG. 3. *Illustration of confidence interval construction. We generated* $(\mathbf{X}, \mathbf{Z}) \in \mathbb{R}^{n\times p}$ *with independent rows distributed as* $\mathcal{N}_p(0, \Sigma)$ *with* $\Sigma_{jk} = 0.9^{|j-k|}$ *where* $(n, p) = (200, 500)$. *A response* $\mathbf{Y}$ *was generated through* $Y_i = X_i - 0.5Z_{i1} + 0.7Z_{i2} + \varepsilon_i$ *where* $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, I)$. *The plot shows* $T_{DEF,t}$ (17) *as a function of* t (*black curve*). *Horizontal dotted blue lines lie at* $\pm z_{0.05}$ *and the shaded red region enclosing the intersection points with the curve* $(t, T_{DEF,t})$ *depicts the* 95% *confidence interval*; *here this contains the true parameter* $\theta = 1$.

Interestingly, in the case where the X-model holds, f can be a fairly exotic function such that different components of $f(\mathbf{Z}, \boldsymbol{\varepsilon}) \in \mathbb{R}^n$ are dependent, provided (X6) holds. Figure 3 illustrates our construction.

Rather than directly seeking for an estimate of θ, by inverting hypothesis tests, we do not rely on being able to distinguish the contribution of $\mathbf{X}$ from among the remaining covariates $\mathbf{Z}$. Thus, for example having $\mathbf{X}$ very highly correlated with $\mathbf{Z}$ would not interfere with coverage properties of the intervals.

Of course, computing $T_{\mathrm{DEF},t}$ for all $t \in \mathbb{R}$ is not feasible. However, while R_α is not guaranteed to be an interval in general, it appears to be the case in practice and we have yet to find a counterexample. This observation allows us to find the end points of the interval via a bisection search. We use coordinate descent to solve the square-root Lasso programmes involved in computing the test statistics $T_{\mathrm{DEF},t}$, and warm start this iterative optimisation procedure at the closest point computed in the search. While this construction is computationally more intensive than the standard approach with the debiased Lasso, it is still feasible in large-scale settings. For the example shown in Figure 3, the computation of the 500 confidence intervals taking each columns of $\mathbf{Z}$ as the variable of interested (i.e., treating it as $\mathbf{X}$) took under 6 seconds on a standard laptop; this time could be further reduced by performing computations in parallel.

In Section C of the Appendix, we show how a similar technique to that described above can be used to construct confidence intervals for $w^T\beta^Y$ for some $w \in \mathbb{R}^p$ that is potentially dense, when the Y-model is a sparse linear model. This is perhaps most useful when w is an additional covariate vector for a new observation whose corresponding response has not been observed; we can thus provide a confidence interval for the mean response conditional on the observed vector of covariates.

3.3 Generalised Linear Models

We have seen in Section 3.1 how one can modify the debiased Lasso to construct a test statistic that has similar sorts of DEF properties to that enjoyed by the standard t-statistic in the low-dimensional setting. In Section 2.2, we saw how standard inference for generalised linear models has a DEF property, albeit with a slight modification needed to account for the different variances of the test statistics when the Y-model is misspecified. It is natural to ask whether inferential procedures for high-dimensional generalised linear models can be adapted to be DEF, but one could equally ask the broader question of whether we can specify sparse generalised linear X and Y-models (possibly different for each), and obtain valid inference if at most one of these is misspecified: this is the question we attempt to address here. As a first step in this direction, we consider heteroscedastic linear models, and then move on to treat generalised linear models in Section 3.3.2.

3.3.1 *Heteroscedastic linear models*. Consider the model $Y_i = Z_i^T \beta^Y + \zeta_i$ where $\mathbb{E}(\zeta_i|\mathbf{Z}) = 0$, $\mathrm{Var}(\zeta_i|\mathbf{Z}) = \sigma_Y^2/(D_{ii}^Y)^2$ and the ζ_i are independent conditional on $\mathbf{Z}$; and a similar X-model. Equivalently, we may write

$$\mathbf{D}^Y\mathbf{Y} = \mathbf{D}^Y\mathbf{Z}\Lambda^Y\beta^Y + \boldsymbol{\varepsilon}^Y, \tag{18}$$

$$\mathbf{D}^X\mathbf{X} = \mathbf{D}^X\mathbf{Z}\Lambda^X\beta^X + \boldsymbol{\varepsilon}^X \tag{19}$$

for the Y and X-models respectively, where $\mathrm{Var}(\varepsilon_i^Y) = \sigma_Y^2$, $\mathrm{Var}(\varepsilon_i^X) = \sigma_X^2$ and the diagonal matrices $\Lambda^Y, \Lambda^X \in \mathbb{R}^{p\times p}$ are such that the empirical variances of the columns of the resulting design matrices $\mathbf{D}^Y\mathbf{Z}\Lambda^Y$ and $\mathbf{D}^X\mathbf{Z}\Lambda^X$ are 1. Note we have redefined β^Y and β^X by scaling them by Λ^Y and Λ^X respectively. We will treat the diagonal matrices $\mathbf{D}^Y$ and $\mathbf{D}^X$ as known, though one of (18) and (19) may be misspecified, in which case the corresponding matrix will be meaningless. In this context, it seems natural to seek an analogue of the test statistic T_{DEF} based on the weighted square-root Lasso regressions

$$\hat\beta^Y = \operatorname*{arg\,min}_{b\in\mathbb{R}^p}\{\|\mathbf{D}^Y(\mathbf{Y} - \mathbf{Z}\Lambda^Y b)\|_2/\sqrt{n} + \lambda\|b\|_1\},$$

$$\hat\beta^X = \operatorname*{arg\,min}_{b\in\mathbb{R}^p}\{\|\mathbf{D}^X(\mathbf{X} - \mathbf{Z}\Lambda^X b)\|_2/\sqrt{n} + \lambda\|b\|_1\}.$$

The KKT conditions of the above optimisations are, however, not "compatible" in the same way as allowed for arguments similar to (13); the issue is that the design matrices in (19) and (18) are different so Theorem 5 does not directly apply. Thus we cannot conclude that the bias term is small unless, for example, both the X and Y-models specified above hold. Instead, consider orthogonalising the residuals $\tilde{\mathbf{Y}} := \mathbf{Y} - \mathbf{Z}\Lambda^Y\hat\beta^Y$ and $\tilde{\mathbf{X}} := \mathbf{X} - \mathbf{Z}\Lambda^X\hat\beta^X$ from the regressions above using the following construction:

$$\begin{aligned}(\tilde\beta^Y, \tilde\eta^Y) = \operatorname*{arg\,min}_{(b,u)\in\mathbb{R}^p\times\mathbb{R}^p} \{&\|\mathbf{D}^Y(\tilde{\mathbf{Y}} - \mathbf{Z}\Lambda^Y b)\\ &- \mathbf{D}^X\mathbf{Z}\Lambda^X u)\|_2/\sqrt{n}\\ &+ \lambda(\|b\|_1 + \|u\|_1)\}\end{aligned} \tag{20}$$

$$\begin{aligned}(\tilde\beta^X, \tilde\eta^X) = \operatorname*{arg\,min}_{(b,u)\in\mathbb{R}^p\times\mathbb{R}^p} \{&\|\mathbf{D}^X(\tilde{\mathbf{X}} - \mathbf{Z}\Lambda^X b)\\ &- \mathbf{D}^Y\mathbf{Z}\Lambda^Y u)\|_2/\sqrt{n}\\ &+ \lambda(\|b\|_1 + \|u\|_1)\}.\end{aligned} \tag{21}$$

Here we have augmented the designs with the terms $\mathbf{D}^X\mathbf{Z}$ and $\mathbf{D}^Y\mathbf{Z}$. The only purpose of these terms and the corresponding estimates $\tilde\eta^Y$ and $\tilde\eta^X$ is to ensure that the residuals from the regressions above satisfy the required near-orthogonality properties for controlling the bias term.

Consider now the case that the Y-model (18) is well specified. Let $\mathbf{R}^X := \mathbf{D}^X(\tilde{\mathbf{X}} - \mathbf{Z}\Lambda^X\tilde\beta^X) - \mathbf{D}^Y\mathbf{Z}\Lambda^Y\tilde\eta^X$. The KKT conditions for (21) yield in particular that

$$\begin{aligned}&\frac{1}{\sqrt{n}}\frac{\|\Lambda^X\mathbf{Z}^T\mathbf{D}^X\mathbf{R}^X\|_\infty}{\|\mathbf{R}^X\|_2} \le \lambda,\\ &\frac{1}{\sqrt{n}}\frac{\|\Lambda^Y\mathbf{Z}^T\mathbf{D}^Y\mathbf{R}^X\|_\infty}{\|\mathbf{R}^X\|_2} \le \lambda;\end{aligned} \tag{22}$$

note the second inequality is due to the additional $\mathbf{D}^Y\mathbf{Z}\Lambda^Y$ term included in (21). Let us also define $\mathbf{R}^Y$ to be the equivalent of $\mathbf{R}^X$, but with X and $\mathbf{X}$ interchanged everywhere with Y and $\mathbf{Y}$ respectively. With these we may define a weighted version of the test statistic T_{DEF} which is simply a scaled correlation between the weighted residuals $\mathbf{R}^X$ and $\mathbf{R}^Y$:

$$T_{\text{W-DEF}} := \sqrt{n}\frac{(\mathbf{R}^X)^T\mathbf{R}^Y}{\|\mathbf{R}^X\|_2\|\mathbf{R}^Y\|_2}.$$

Similar to the homoscedastic case, we set $T_{\text{W-DEF}} = 0$ if the denominator above is zero. We now explain why we will typically have $T_{\text{W-DEF}} \xrightarrow{d} \mathcal{N}(0,1)$ if the Y-regression holds, and hence also by symmetry, if the X-regression holds. Let us write $\hat\sigma := \|\mathbf{R}^Y\|_2/\sqrt{n}$. Now

$$\begin{aligned}\mathbf{R}^Y = \boldsymbol{\varepsilon}^Y + \mathbf{D}^Y\{&\mathbf{Z}\Lambda^Y(\beta^Y - \hat\beta^Y - \tilde\beta^Y)\}\\ &- \mathbf{D}^X\mathbf{Z}\Lambda^X\tilde\eta^Y.\end{aligned} \tag{23}$$

Thus, we have

$$\begin{aligned}\hat\sigma T_{\text{W-DEF}} &= (\boldsymbol{\varepsilon}^Y)^T\frac{\mathbf{R}^X}{\|\mathbf{R}^X\|_2}\\ &\quad + (\beta^Y - \hat\beta^Y - \tilde\beta^Y)^T\Lambda^Y\mathbf{Z}^T\mathbf{D}^Y\frac{\mathbf{R}^X}{\|\mathbf{R}^X\|_2}\\ &\quad - (\tilde\eta^Y)^T\Lambda^X\mathbf{Z}^T\mathbf{D}^X\frac{\mathbf{R}^X}{\|\mathbf{R}^X\|_2}\\ &=: \text{(i)} + \text{(ii)} + \text{(iii)}.\end{aligned}$$

Under weak conditions, the first term (i) will converge in distribution to a normal distribution. The two sets of near-orthogonality conditions (22) in conjunction with Hölder's inequality give that the two bias terms above satisfy

$$|(\mathrm{ii})| \le \sqrt{n}\lambda(\|\beta^Y - \hat{\beta}^Y\|_1 + \|\tilde{\beta}\|_1),$$
$$|(\mathrm{iii})| \le \sqrt{n}\lambda\|\hat{\eta}^Y\|_1$$

respectively. As explained in Section 3.1, we can expect that under reasonable conditions we have $\|\beta^Y - \hat{\beta}^Y\|_1 \lesssim s_Y\sqrt{\log(p)/n}$ with high probability. The additional terms $\|\tilde{\beta}^Y\|_1$ and $\|\tilde{\eta}^Y\|_1$ may be controlled similarly to $\|\beta^Y - \hat{\beta}^Y\|_1$; see Theorem 7 below.

Throughout the discussion above, we have assumed that the Y-model holds. If instead the X-model is correct, the symmetry of the test statistic allows that analogous results may be established in the same manner, justifying that $T_{\text{W-DEF}}$ has a standard normal distribution under the null hypothesis if either model is well specified. This is formalised in the result below, which assumes some additional moment conditions for the entries in $\mathbf{Z}$, and a condition on the growth rate of p compared to n.

THEOREM 7. *Suppose there exist constants $M, \delta > 0$ such that*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(|D_{ii}^Y Z_{ij}\Lambda_{ii}^Y|^{2+\delta} + |D_{ii}^X Z_{ij}\Lambda_{ii}^X|^{2+\delta}) \le M\right) \to 1,$$

and $p \le n^{c\delta}$ for some $c \in (0,1)$ and all n sufficiently large. Suppose that (Y1) *holds with the heteroscedastic Y-model* (18) *in place of the linear model and δ as above,* (Y2) *holds with $\mathbf{R} = \mathbf{R}^X$, and* (Y3)–(Y5) *hold. Suppose $\lambda = A\sqrt{2\log(p)/n}$ for some $A > 1$. Then there exists a constant $C > 0$ such that*

$$\mathbb{P}(\|\tilde{\beta}^Y\|_1 + \|\tilde{\eta}^Y\|_1 \le C\|\beta^Y - \hat{\beta}^Y\|_1) \to 1, \tag{24}$$

and moreover, under the null hypothesis that $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$, we have $T_{\text{W-DEF}} \xrightarrow{d} \mathcal{N}(0,1)$.

By symmetry, an analogous version of the result holds with every instance of Y and $\mathbf{Y}$ interchanged with X and $\mathbf{X}$, respectively.

3.3.2 *Generalised linear models.* With the methodology for heteroscedastic linear models introduced above, we can now set out a DEF test statistic for the case where we wish to specify the X and Y-models as generalised linear models. The first step is to run penalised generalised linear regressions of each of $\mathbf{Y}$ and $\mathbf{X}$ on $\mathbf{Z}$ to obtain coefficient estimates $\hat{\beta}^Y, \hat{\beta}^X \in \mathbb{R}^p$. Let μ_X and μ_Y be the respective mean functions (i.e., inverse link functions) so that if the Y-model is well specified and $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{Z}$, we have $\mathbb{E}(Y_i|Z_i) = \mu_Y(Z_i^T\beta^Y)$ where $\beta^Y \in \mathbb{R}^p$. Further define variance functions $V_{Y,i}$ for the Y-model; when the Y-model holds we will have $V_{Y,i}(\mu_Y(Z_i^T\beta^Y)) = \mathrm{Var}(Y_i|Z_i)$. We will assume for simplicity that the $V_{Y,i}$ are known and do not vary over the observations, so we may write $V_Y = V_{Y,i}$. Define the variance function V_X for the X-model analogously.

To compute a DEF test statistic for generalised linear models, we take the following steps.

1. Define the adjusted response $\tilde{\mathbf{Y}} \in \mathbb{R}^n$ by

$$\tilde{Y}_i := \frac{Y_i - \mu_Y(Z_i^T\hat{\beta}^Y)}{\mu_Y'(Z_i^T\hat{\beta}^Y)}$$

and define $\tilde{\mathbf{X}}$ analogously.

2. Define diagonal matrix $\hat{\mathbf{D}}^Y \in \mathbb{R}^{n\times n}$ by $\hat{D}_{ii}^Y = \mu_Y'(Z_i^T\hat{\beta}^Y)\{V_Y(\mu_Y(Z_i^T\hat{\beta}^Y))\}^{-1/2}$, and define $\hat{\mathbf{D}}^X$ analogously.

3. Compute test statistic $T_{\text{GLM-DEF}}$ by forming $T_{\text{W-DEF}}$ but replacing $\mathbf{X}$ and $\mathbf{Y}$ with their adjusted versions $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$, and using the diagonal matrices $\hat{\mathbf{D}}^X$ and $\hat{\mathbf{D}}^Y$ defined above.

We now explain why we can expect that $T_{\text{GLM-DEF}} \xrightarrow{d} \mathcal{N}(0,1)$ when $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$ and either the Y-model or X-model is well specified. Suppose that the Y-model holds. Then a first-order Taylor expansion yields

$$\begin{aligned} Y_i - \mu_Y(Z_i^T\hat{\beta}^Y) &= \mu_Y(Z_i^T\beta^Y) - \mu_Y(Z_i^T\hat{\beta}^Y) + \zeta_i \\ &\approx Z_i^T(\beta^Y - \hat{\beta}^Y)\mu_Y'(Z_i^T\hat{\beta}^Y) + \zeta_i, \end{aligned}$$

where $\mathbb{E}(\zeta_i|Z_i) = 0$ and $\mathrm{Var}(\zeta_i|Z_i) = V_Y(\mu_Y(Z_i^T\beta^Y))$. Thus, $\tilde{Y}_i \approx Z_i^T(\beta^Y - \hat{\beta}^Y) + \zeta_i/\mu_Y'(Z_i^T\hat{\beta}^Y)$ and hence

$$\hat{\mathbf{D}}^Y\tilde{\mathbf{Y}} \approx \hat{\mathbf{D}}^Y\mathbf{Z}(\beta^Y - \hat{\beta}^Y) + \boldsymbol{\varepsilon},$$

where $\mathbb{E}(\boldsymbol{\varepsilon}|\mathbf{Z}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}|\mathbf{Z}) = \mathbf{I}$.

Now the square-root Lasso regression involving $\tilde{\mathbf{Y}}$ used in step 3 above should have little effect as $\tilde{\mathbf{Y}}$ is essentially noise (see Theorem 7). The corresponding regression for $\tilde{\mathbf{X}}$, however, will ensure the resulting residuals are almost orthogonal to the bias term $\hat{\mathbf{D}}^Y\mathbf{Z}(\beta^Y - \hat{\beta}^Y)$. Arguing similarly to (23), we see that the overall bias should be well controlled. The variance term $\boldsymbol{\varepsilon}^T\mathbf{R}^X/\|\mathbf{R}^X\|_2$ should behave roughly like a weighted sum of independent zero-mean random variables ε_i. The fact that $\hat{\mathbf{D}}^Y$ is used in the construction of the residuals $\mathbf{R}^X$, however, means they are not independent of $\boldsymbol{\varepsilon}$, and one cannot directly apply a version of the central limit theorem to the term. While some form of sample splitting could in principle help with this technical issue (see, e.g., Janková et al. (2020) where sample splitting is used in a similar context), as the dependence is weak, a normal approximation should work well in practice; indeed we show empirically in Section 4 that this is the case.

3.3.3 *Connections to the generalised covariance measure, the decorrelated score test and the debiased lasso.* An alternative to the approach for DEF inference in high-dimensional generalised linear models presented in the previous sections is based on the score test. Considering the setup of Section 2.2, the key argument that results in the DEF property for maximum likelihood estimation in low-dimensional generalised linear models is that $\beta^\dagger$ defined as the maximiser of $\mathbb{E}\ell(Z^T\beta; Y)$ over $\beta \in \mathbb{R}^p$ satisfies

$$\begin{aligned}(25)\qquad &\mathbb{E}\{XU(Z^T\beta^\dagger; Y)\} \\ &\quad = \mathbb{E}\{(X - Z^T\beta^X)U(Z^T\beta^\dagger; Y)\},\end{aligned}$$

where $\beta^X := \arg\min_{\beta\in\mathbb{R}^p} \mathbb{E}\{(X - Z^T\beta)^2\}$ is the best linear predictor of X based on Z. It is straightforward to see that if $X \perp\!\!\!\perp Y|Z$, the RHS is always zero whenever $Z^T\beta^X$ coincides with $\mathbb{E}(X|Z)$, and clearly the LHS (and hence also the RHS) is zero whenever the model (5) is well specified.

The RHS of (25) may be used as the basis of a score-type test involving linearly regressing $\mathbf{X}$ onto $\mathbf{Z}$, and forming the empirical covariance of these residuals and $(U(Z_i^T\check{\beta}^Y; Y_i))_{i=1}^n$, where $\check{\beta}^Y$ is a maximum likelihood estimate of β^Y. Given that both regressions of $\mathbf{X}$ and $\mathbf{Y}$ on $\mathbf{Z}$ are performed to produce such a test statistic, it is more intuitively clear that this would have a DEF property. The $\mathbf{X}$ on $\mathbf{Z}$ regression is, however, redundant as the stationarity conditions of $\check{\beta}^Y$ dictate that $(U(Z_i^T\check{\beta}^Y; Y_i))_{i=1}^n$ is orthogonal to the column space of $\mathbf{Z}$. Thus a regular score test would have the DEF property for a linear regression model of X on Z.

In high-dimensional settings the estimate $\check{\beta}^Y$ will necessarily only yield approximate orthogonality to $\mathbf{Z}$, and so the regression of $\mathbf{X}$ on $\mathbf{Z}$ is crucial. In a setting where the regression for Y-model is a generalised linear model with canonical link, this leads to a test statistic of the form

$$(26)\qquad \frac{1}{\hat{\tau}_D}\sum_{i=1}^n (X_i - Z_i^T\check{\beta}^X)^T\{Y_i - \mu_Y(Z_i^T\check{\beta}^Y)\},$$

where $\hat{\tau}_D$ is a normalisation term that ensures an asymptotically unit variance under the null. This is the form of the generalised covariance measure (GCM) (Shah and Peters, 2020), the decorrelated score test (Ning and Liu, 2017), and, to a first-order Taylor approximation, the debiased Lasso (van de Geer et al., 2014); however, they differ primarily in their choice of estimates $\check{\beta}^X$ and $\check{\beta}^Y$. Both the GCM and the decorrelated score construct $\check{\beta}^Y$ through only regressing on $\mathbf{Z}$, similarly to our DEF approach, whereas the debiased Lasso involves a regression on $(\mathbf{X}, \mathbf{Z})$. Like our approach, the $\mathbf{X}$ on $\mathbf{Z}$ regression in the GCM is performed without using $\mathbf{Y}$ and can be tailored to a specified X-model, whereas both the decorrelated score test and the debiased Lasso aim to construct $\check{\beta}^X$ so that the residuals $\mathbf{X} - \mathbf{Z}\check{\beta}^X$ are orthogonal to the bias in the residuals from $\mathbf{Y}$ regression, were the Y-model to be correct. Our DEF approach instead employs an orthogonalisation step using the square-root Lasso corresponding to each of X and Y after initial $\mathbf{X}$ and $\mathbf{Y}$ regressions have been performed. A further difference is that whereas (26) involves an empirical covariance between raw residuals, our DEF approach uses Pearson residuals. This is so that the square-root Lasso orthogonalisation corresponding to the true model is performed on data with (approximately) homoscedastic errors, which permits (24) to hold. We have, however. found that a version of the test with raw residuals performs very similarly in terms of power and type I error control.

4. NUMERICAL EXPERIMENTS

In this section, we explore the empirical properties of our proposed DEF methodology set out in Section 3.

4.1 Partially Linear Models

Here we investigate the empirical performance of our DEF confidence interval construction described in Section 3.2, and compare it with the debiased Lasso. We consider partially linear regression models of the form

$$Y_i = \theta X_i + f(Z_i, \varepsilon_i),$$

where the goal is to provide a confidence interval for θ. The nuisance function f, parameter θ and data $(Y_i, X_i, Z_i, \varepsilon_i) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^p\times\mathbb{R}$ for $i = 1, \ldots, n$ with $n = 100$ are generated as follows. We use the publicly available gene expression data of Bacillus Subtilis (Dezeure et al., 2014), which has 71 observations and 4088 predictors. We first select the $p + 1 = 500$ predictors with the highest empirical variances, and then centre and scale these so the empirical variances are 1. We then fit a Gaussian copula model to these predictors to give a 500-dimensional multivariate distribution P from which we can generate independent realisations of (X_i, Z_i). This distribution is non-Gaussian and has some large pairwise correlations and thus is helpful for assessing how our methods may perform in challenging and realistic settings.

To form $(X_i, Z_i)_{i=1}^n$ we first generate $(W_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} P$ and then consider 12 settings taking each of the first 12 components of W_i as the variable X_i of interest, and collecting the remaining components into Z_i. For each of the 12 settings, we generate a new $\theta \sim U[0, 2]$, and look at 3 forms for the nuisance function f.

(a) *Linear.* We set

$$f(Z_i, \varepsilon_i) = \sum_{j=1}^{11} Z_{ij}\beta_j + \varepsilon_i,$$

where the $(\beta_j)_{j=1}^{11}$ are generated independently and follow a $U[0, 2]$. distribution.

(b) *Slightly nonlinear.* We set

$$f(Z_i, \varepsilon_i) = \sum_{j=1}^{11} \tilde{Z}_{ij}\beta_j + \varepsilon_i$$

with $(\beta_j)_{j=1}^{11}$ as in (a) and $\tilde{Z}_{ij} := 2e^{Z_{ij}}/(1+e^{Z_{ij}}) - 1$.

(c) *Highly nonlinear.* We first form

$$\eta_i := \sum_{j=1}^{11} \tilde{Z}_{ij}\beta_j + \sum_{j=1}^{11}\sum_{k=1}^{11} \tilde{Z}_{ij}\tilde{Z}_{ik}\theta_{jk} + \varepsilon_i,$$

where the $(\tilde{Z}_{ij})_{j=1}^{11}$ and $(\beta_j)_{j=1}^{11}$ are as above and $(\theta_{jk})_{j,k=1}^{11}$. We then set $f(Z_i, \varepsilon_i) = e^{\eta_i}/(1+e^{\eta_i})$.

In all cases the errors $(\varepsilon_i)_{i=1}^n$ are taken to be i.i.d. standard normal. In our implementation of the debiased Lasso and DEF confidence intervals, we use the square-root Lasso with parameters λ_X and λ_Y chosen according to the method of Sun and Zhang (2013). Figures 4, 5 and 6 show the results. We see that the DEF 95% confidence intervals have significantly better coverage compared to those based on the debiased Lasso. This is even true in the linear setting where one might have expected the performances to be similar, suggesting that the strategy of inverting hypothesis tests may also be useful when applied in conjunction with debiased Lasso-based tests. The improved coverage we observe is partly due to the DEF confidence intervals being wider, but they also seem to have slightly better centring around the true parameter values; in contrast the debiased Lasso confidence intervals display a substantial bias towards zero in several cases.

Note that the nonlinear settings (b) and (c) do not quite satisfy the conditions for our theory (see Theorem 5) as the non-Gaussianity of the Z_i would mean that the X-models are unlikely to be sparse linear models. Nevertheless, the coverage is reasonable if not perfect in these more challenging settings. Results for analogous scenarios to those studied here but with P replaced by a multivariate Gaussian with a Toeplitz covariance matrix Σ where $\Sigma_{jk} = 0.9^{|j-k|}$ are shown in Section D of the Appendix. In these settings, the X-model is a highly sparse linear model, and as a result the coverage properties of both methods are improved; however, the debiased Lasso still undercovers while the DEF confidence intervals reach a coverage of closer to 95%. We have observed a very similar pattern of results for other settings of (n, p).

4.2 Generalised Linear Models

Here we present some simple experiments to investigate the performance of the DEF statistic $T_{\text{GLM-DEF}}$ for generalised linear models (Section 3.3) where we take the X and Y-models to be logistic regression models. We generate data $(Y_i, X_i, Z_i) \in \{0,1\} \times \{0,1\} \times \mathbb{R}^p$ for $i = 1, \dots, n$ with $(n, p) = (250, 100)$ in the following way. We first construct a multivariate distribution P as in Section 4.1, but take $p = 250$. We then simulate $Z_i \overset{\text{i.i.d.}}{\sim} P$, and independently generate $Y_i \sim \text{Bern}(\pi_i^Y)$ and

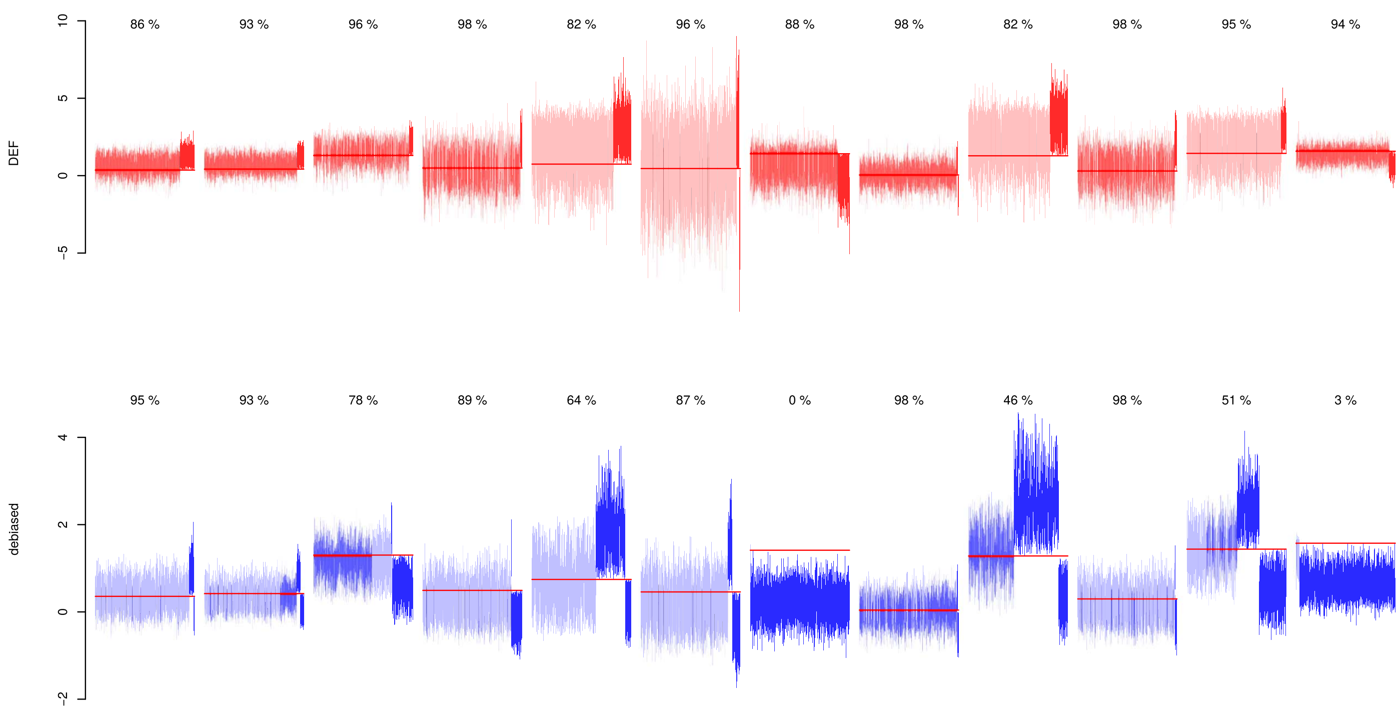

FIG. 4. *DEF (top row) and debiased Lasso (bottom row) 95% confidence intervals from 500 simulations of each of the 12 linear settings (a). The light red and blue vertical lines depict those confidence intervals that covered their target parameter θ shown the red horizontal lines. Darker vertical lines are confidence intervals that failed to cover their target and are grouped into those whose endpoints were too high, and too low. Coverage proportions are reported above each of the plots.*

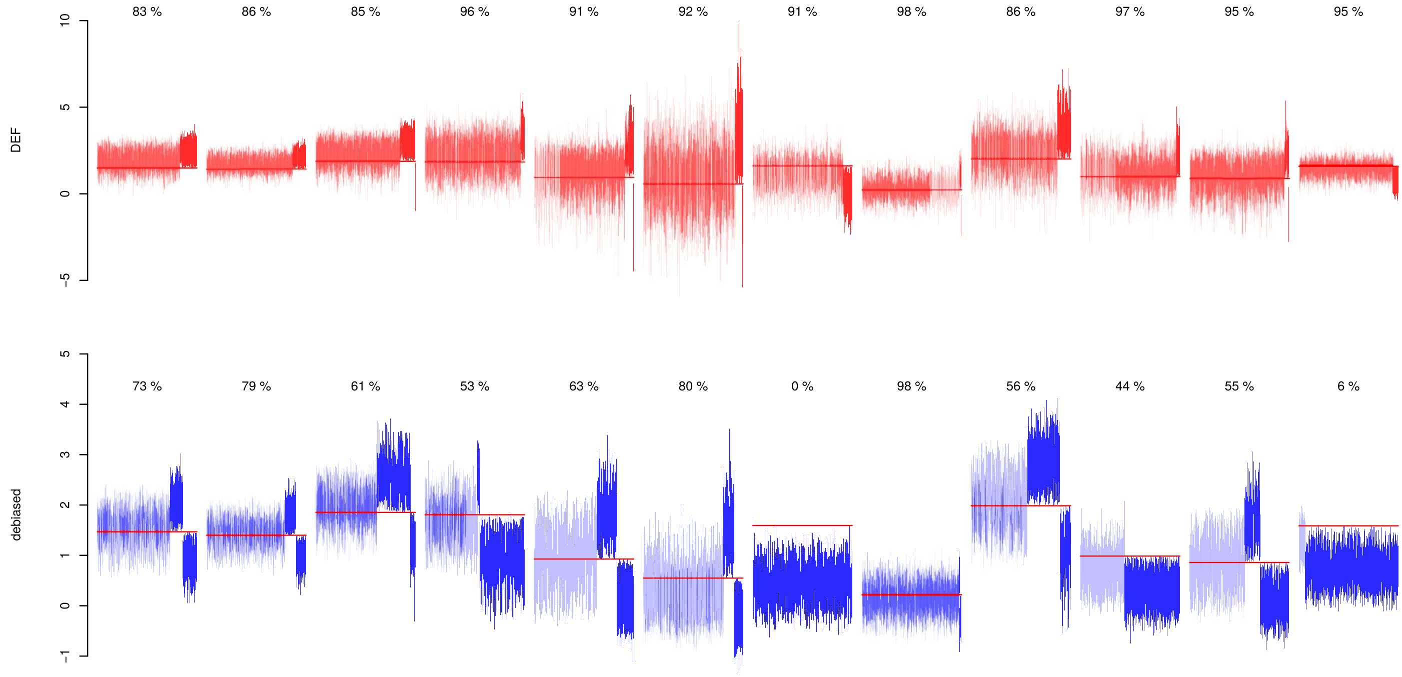

FIG. 5. *The slightly nonlinear setting* (*b*); *the interpretation is similar to that of Figure* 4.

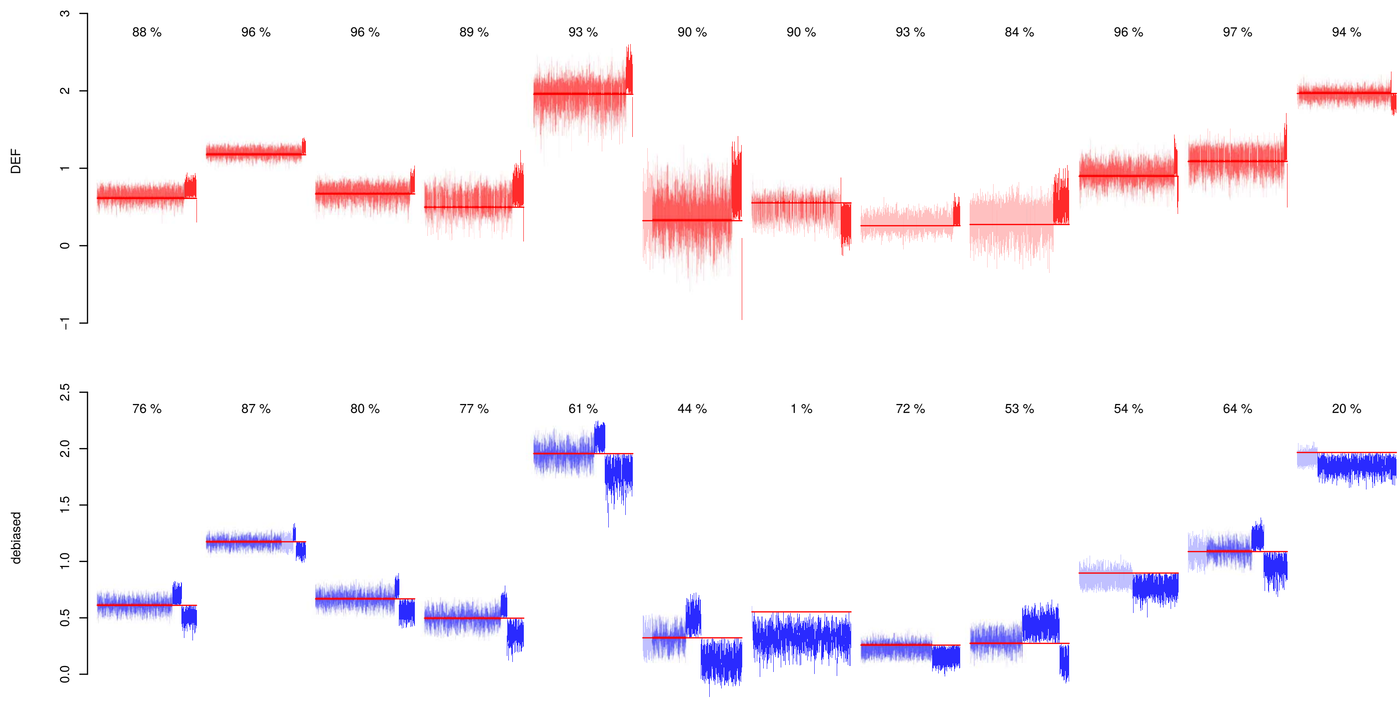

FIG. 6. *The highly nonlinear setting* (*c*); *the interpretation is similar to that of Figure* 4.

$X_i \sim \text{Bern}(\pi_i^X)$ where probabilities π_i^Y and π_i^X satisfy

$$\begin{aligned} \text{logit}(\pi_i^Y) &= \sum_{j=1}^{24} a_j Z_{ij} \beta_j, \\ \text{logit}(\pi_i^X) &= \sum_{j=1}^{4} a_j Z_{ij} \beta_j, \end{aligned} \tag{27}$$

with $\beta_j \overset{\text{i.i.d.}}{\sim} U[0,1]$ and $a_j = 1 - (j-1)/24$. Note that $X_i \perp\!\!\!\perp Y_i | Z_i$; however, the X_i and Y_i are positively correlated, making control of the type I error when performing the conditional independence test challenging.

We generate 6 sets of $(\beta, \mathbf{Z})$ pairs, and for each of these simulate 250 realisations of $\mathbf{X}$ and $\mathbf{Y}$. To each of the 6×250 datasets, we apply our DEF methodology positing logistic regression models for the X and Y-models,

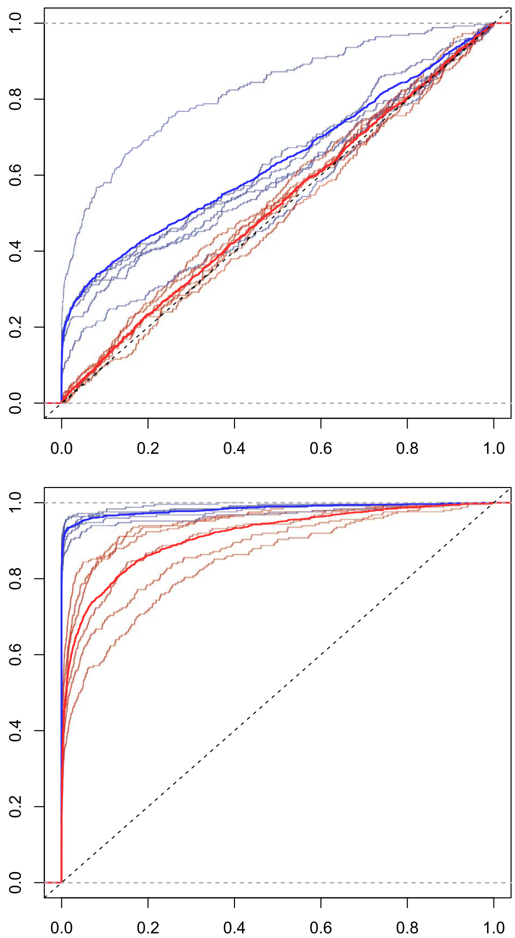

FIG. 7. *Empirical distribution functions (ECDFs) of p-values constructed via the DEF (red) and debiased Lasso (blue) approaches for null (top) and alternative (bottom) settings described in Section 4.2. In each panel, the fainter and thinner lines correspond to the* 6 *setups with different* $(\beta, \mathbf{Z})$ *while the thicker solid lines are aggregate ECDFs.*

and also the debiased Lasso for generalised linear models via weighted least squares (see Section 3.2 of Dezeure et al., 2015). The results are given in the top plot of Figure 7. We see that that DEF approach is able to control the type I error by exploiting the fact that the X-model, being highly sparse, is relatively easy to estimate. On the other hand, the debiased Lasso requires accurate estimation of all 24 components of β in the Y-model, and as a consequence is highly anti-conservative here.

To assess the power of the methods, we consider an identical setup as just described, but X_i is added to the right-hand side of (27) to induce dependence. The bottom plot in Figure 7 presents the corresponding results. We see that while the p-values for $T_{\text{GLM-DEF}}$ are sub-uniform, power is reduced compared to the debiased Lasso as expected; this is the price of the additional robustness offered by the DEF approach.

5. DISCUSSION

In recent years, there has been growing interest in understanding the performance of statistical procedures when the models they have been designed for are misspecified; see, for example, Buja et al. (2019a), Buja et al. (2019b). In this work, we consider regression models with response Y, a single predictor of interest X, and additional covariates $Z \in \mathbb{R}^p$. Our goal is assessing the significance of X after controlling for Z, a problem which may be equivalently framed as testing for the null hypothesis H_0 of conditional independence $Y \perp\!\!\!\perp X|Z$. If either the Y or the X-model is linear or generalised linear, the situation is favourable for DEF inference.

The DEF property holds for a test statistic T if the following is true. Under $H_0 : X \perp\!\!\!\perp Y|Z$ we have $T \xrightarrow{d} \mathcal{N}(0, 1)$ when at least one among the Y and X-model holds. Examples of such test statistics include the following ones:

(i) T_{OLS}, the standard t-statistic for testing significance of the parameter corresponding to X as laid out in Theorem 1;

(ii) $T_{\text{GLM},1}$ and $T_{\text{GLM},2}$, the modified Wald statistics with correction factors (see (9)) for the standard error as discussed in Section 2.2;

(iii) T_{DEF} in (15) based on a symmetrised version of the debiased Lasso in a high-dimensional linear model as discussed in Theorem 5;

(iv) $T_{\text{GLM-DEF}}$ based on a symmetrised version of the debiased Lasso in high-dimensional generalised linear models as discussed in Section 3.3.

In cases (iii) and (iv), we explicitly model both the X and Y regressions, and also explicitly build in symmetry into the test statistics to reflect the symmetry of the null hypothesis. On the other hand, the first two examples, which relate to low-dimensional settings, are not obviously engineered to have the DEF property. An interesting finding here is that the these classical test statistics implicitly use a linear X-model. We may speculate that this hidden robustness of classical significance tests to potentially severe Y-model misspecification has in some way contributed to their popularity and usefulness given that all models—but as we have established here, *not* all inferential tools—are wrong (Box, 1976).

As a separate point of interest, we argue that confidence intervals in high-dimensional settings should be constructed via inversion of tests instead of relying directly on asymptotic distribution theory for the relevant pivot. Supporting empirical evidence is given in Section 3.2.

Our work also offers a number of potentially fruitful directions for further research. For example, it would be interesting to investigate the power properties of our DEF procedures. In addition, lower bounds on the power that can be achieved subject to a DEF property holding would be worth exploring. Finally, the analogue of the method proposed for confidence interval construction via inverting tests seems not to have the DEF property in the context of generalised linear models. It would be very useful to develop DEF confidence intervals for this setting, or indeed prove that it is in some sense not possible.

APPENDIX A: PROOFS

A.1 Proof of Theorem 1

The relationship (3) between the t-statistic T and the partial correlation $\hat{\rho}$ follows easily from the following observations:

$$\hat{\theta} = \frac{\mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{X}}{\|(\mathbf{I}-\mathbf{P})\mathbf{X}\|_2^2},$$

$$\{(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\}_{11} = \|(\mathbf{I}-\mathbf{P})\mathbf{X}\|_2^{-2},$$

$$\begin{aligned}\|(\mathbf{I}-\tilde{\mathbf{P}})\mathbf{Y}\|_2^2 &= \|(\mathbf{I}-\mathbf{P})\mathbf{Y}\|_2^2 - \frac{\{\mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{X}\}^2}{\|(\mathbf{I}-\mathbf{P})\mathbf{X}\|_2^2} \\ &= \|(\mathbf{I}-\mathbf{P})\mathbf{Y}\|_2^2(1-\hat{\rho}^2).\end{aligned}$$

Thus, it suffices to show that $\sqrt{n}\hat{\rho} \xrightarrow{d} \mathcal{N}(0,1)$ since this implies that $\hat{\rho} \xrightarrow{p} 0$. As $\hat{\rho}$ is symmetric in $\mathbf{X}$ and $\mathbf{Y}$, we need only show these facts hold assuming (Y1) and (Y2). Note then we have

$$\hat{\rho} = \frac{\mathbf{X}^T(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}}{\|(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}\|_2\|(\mathbf{I}-\mathbf{P})\mathbf{X}\|_2},$$

where since $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$, the properties of $(\varepsilon_i)_{i=1}^n$ hold conditionally on $(\mathbf{Z},\mathbf{X})$. We first show $\|(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}\|_2/\sqrt{n} \xrightarrow{p} \sigma$. We have

$$\frac{1}{n}\|(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}\|_2^2 = \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2 - \frac{1}{n}\boldsymbol{\varepsilon}^T\mathbf{P}\boldsymbol{\varepsilon}. \tag{28}$$

By the weak law of large numbers, the first term converges in probability to σ^2. For the second term, note that due to (Y1), using the cyclic property of the trace operator,

$$\begin{aligned}\mathbb{E}\boldsymbol{\varepsilon}^T\mathbf{P}\boldsymbol{\varepsilon} &= \mathbb{E}\,\mathrm{tr}(\boldsymbol{\varepsilon}^T\mathbf{P}\boldsymbol{\varepsilon}) = \mathrm{tr}(\mathbb{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{P}) \\ &= \mathrm{tr}\,\mathbb{E}\{\underbrace{\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T|\mathbf{P})}_{=\sigma^2\mathbf{I}}\mathbf{P}\}] = \sigma^2\,\mathrm{tr}(\mathbf{P}) \le \sigma^2 p.\end{aligned}$$

Thus the final term in (28) has expectation tending to 0 as $p/n \to 0$. By Markov's inequality, this must therefore go to 0 in probability, and so $\|(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}\|_2/\sqrt{n} \xrightarrow{p} \sigma$ as required.

Next, we claim that

$$A_n := \frac{\mathbf{X}^T(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}}{\|(\mathbf{I}-\mathbf{P})\mathbf{X}\|_2} = \|\mathbf{R}\|_2^{-1}\sum_{i=1}^n R_i\varepsilon_i \xrightarrow{d} \mathcal{N}(0,\sigma^2).$$

Note that conditional on $(\mathbf{X},\mathbf{Z})$, the ε_i are i.i.d. with variance σ^2 and third moment bounded by M. Lemma 8 below with $\mathbf{R}_n = (\mathbf{I}-\mathbf{P})\mathbf{X}$ then shows that $A_n \xrightarrow{d} \mathcal{N}(0,\sigma^2)$. Combining with the previous result and applying Slutsky's lemma gives $\sqrt{n}\hat{\rho} \xrightarrow{d} \mathcal{N}(0,1)$ as required

LEMMA 8. *Let $(\varepsilon_{in})_{i\le n}$ and $(R_{in})_{i\le n}$ be triangular arrays of random variables and define $\mathbf{R}_n = (R_{1n},\dots,R_{nn})$ for all n. Assume these random variables satisfy the following conditions*:

(i) *$\varepsilon_{1n},\dots,\varepsilon_{nn}$ are independent conditional on $\mathbf{R}_n$*;
(ii) *for all $i = 1,\dots,n$ and some $\delta, M > 0$,*

$$\mathbb{E}(\varepsilon_{in}|\mathbf{R}_n) = 0,$$

$$\mathbb{E}(\varepsilon_{in}^2|\mathbf{R}_n) = \sigma^2 > 0,$$

$$\mathbb{E}(|\varepsilon_{in}|^{2+\delta}|\mathbf{R}_n) < M;$$

(iii) $\mathbb{P}(\mathbf{R}_n = \mathbf{0}) \to 0$;
(iv) *for some $\delta > 0$,*

$$A_n := \begin{cases}\dfrac{1}{\|\mathbf{R}_n\|_2^{2+\delta}}\displaystyle\sum_{i=1}^n |R_{in}|^{2+\delta} & \textit{if } \mathbf{R}_n \ne \mathbf{0},\\ 0 & \textit{if } \mathbf{R} = \mathbf{0},\end{cases}$$

satisfies $A_n \xrightarrow{p} 0$, and

$$\frac{1}{\|\mathbf{R}_n\|_2^{2+\delta}}\sum_{i=1}^n |R_{in}|^{2+\delta}\mathbb{1}_{\{\mathbf{R}_n\ne\mathbf{0}\}} \xrightarrow{p} 0.$$

Then

$$B_n := \|\mathbf{R}_n\|_2^{-1}\sum_{i=1}^n R_{in}\mathbb{1}_{\{\mathbf{R}_n\ne 0\}}\varepsilon_{in} \xrightarrow{d} \mathcal{N}(0,\sigma^2).$$

PROOF. Let the random sequences above be defined on a common probability space $(\Omega,\mathcal{F},\mathbb{P})$. Let $(n_k)_{k=1}^\infty \subseteq \mathbb{N}$ be an arbitrary subsequence. Then we know there exists a further subsequence $(n_{k(l)})_{l=1}^\infty$ on which the following occur:

(a) the convergence in (iv) above happens almost surely, that is, the probability that

$$\lim_{l\to\infty}\frac{1}{\|\mathbf{R}_{n_{k(l)}}\|_2^{2+\delta}}\sum_{i=1}^{n_{k(l)}} |R_{in_{k(l)}}|^{2+\delta}\mathbb{1}_{\{\mathbf{R}_{n_{k(l)}}\ne\mathbf{0}\}} = 0$$

equals one.
(b) $\sum_{l=1}^\infty \mathbb{P}(\mathbf{R}_{n_{k(l)}} = \mathbf{0}) < \infty$.

By the first Borel–Cantelli lemma, we have that the sequence of events $\Omega_l := \mathbf{R}_{n_{k(l)}} \ne \mathbf{0}$ satisfies $\mathbb{P}(\liminf_{l\to\infty}\Omega_l) = 1$. Let Ω_2 be the intersection of the event in (a) above and $\liminf_{l\to\infty}\Omega_l$. Note that $\mathbb{P}(\Omega_2) = 1$.

Now observe that for each $\omega \in \Omega_2$, writing $\mathbf{r} := \mathbf{R}_n(\omega)$, we have $\mathbf{r} \ne \mathbf{0}$ and

$$\begin{aligned}C(\omega,n) &:= \frac{\sum_{i=1}^n \mathbb{E}(|R_{in}\varepsilon_{in}|^{2+\delta}|\mathbf{R}_n = \mathbf{r})}{(\sum_{i=1}^n \mathbb{E}(R_{in}^2\varepsilon_{in}^2|\mathbf{R}_n = \mathbf{r}))^{1+\delta/2}} \\ &= \frac{\sum_{i=1}^n |r_{in}|^{2+\delta}\mathbb{E}(|\varepsilon_{in}|^{2+\delta}|\mathbf{R}_n = \mathbf{r})}{(\sum_{i=1}^n r_{in}^2\mathbb{E}(\varepsilon_{in}^2|\mathbf{R}_n = \mathbf{r}))^{1+\delta/2}} \\ &< \frac{M}{\sigma^2}\frac{\sum_{i=1}^n |r_{in}|^{2+\delta}}{(\sum_{i=1}^n r_{in}^2)^{1+\delta/2}}.\end{aligned}$$

Thus

$$\lim_{l\to\infty} B(\omega, n_{k(l)}) = 0$$

for all $\omega \in \Omega_2$.

For each n, let $\tilde{P}_n : \mathbb{R}^n \times \mathcal{F} \to [0, 1]$ be a regular conditional probability given $\mathbf{R}_n$, and for $\omega \in \Omega$, let $P_{n,\omega} : \mathcal{F} \to [0, 1]$ be given by $P_{n,\omega}(A) = \tilde{P}_n(\mathbf{R}_n(\omega), A)$. Denoting expectation with respect to $P_{n,\omega}$ by $E_{n,\omega}$, note that

$$C(\omega, n) = \frac{\sum_{i=1}^n E_{n,\omega}(|R_{in}\varepsilon_{in}|^{2+\delta})}{(\sum_{i=1}^n E_{n,\omega}(R_{in}^2\varepsilon_{in}^2))^{1+\delta/2}}.$$

From the above, for each $\omega \in \Omega_2$, we can apply the Lindeberg–Feller central limit theorem for triangular arrays (van der Vaart, 1998, Prop. 2.27) along the sequence of probability measures given by $(P_{n_{k(l)},\omega})_{l=1}^{\infty}$, noting that Lyapunov's condition implies the Lindeberg–Feller condition. Writing $Z \sim \mathcal{N}(0, \sigma^2)$, we have that for any $\omega \in \Omega_2$ and any continuous bounded function $g : \mathbb{R} \to \mathbb{R}$,

$$\begin{aligned}\lim_{l\to\infty} E_{n_{k(l)},\omega}\{g(B_{n_{k(l)}})\} &= \lim_{l\to\infty} \mathbb{E}\{g(B_{n_{k(l)}})|\mathbf{R}_{n_{k(l)}}\}(\omega)\\ &= \mathbb{E}g(Z).\end{aligned}$$

Now as $\mathbb{P}(\Omega_2) = 1$, we have

$$\lim_{l\to\infty} \mathbb{E}\{g(B_{n_{k(l)}})|\mathbf{R}_{n(k(l)}}\} = \mathbb{E}g(Z)$$

almost surely. Then as the subsequence $(n_k)_{k=1}^{\infty} \subseteq \mathbb{N}$ was arbitrary, we see that in fact

$$\mathbb{E}\{g(B_n)|\mathbf{R}_n\} \xrightarrow{p} \mathbb{E}g(Z).$$

Finally, note that as g is bounded, we may apply dominated convergence theorem to show that

$$\mathbb{E}\{g(B_n)\} \to \mathbb{E}g(Z).$$

As this holds for every continuous bounded g, we have the result. □

A.2 Proof and Regularity Conditions for Theorem 2

A.2.1 *Regularity conditions*. Assume the following regularity conditions.

(i) $L(xt + z^T\beta; y) > 0$ almost everywhere and $\mathbb{E}|\ell(Xt + Z^T\beta; Y)| < \infty$ for all $(t, \beta) \in \Theta$.

(ii) The ratio $L(xt_1 + z^T\beta_1; y)/L(xt_2 + z^T\beta_2; y)$ is not almost everywhere equal to 1 when $(t_1, \beta_1) \neq (t_2, \beta_2)$.

(iii) There exists an open set $K \subset \Theta$ containing $(0, \beta^\dagger)$ such that $\mathbb{E}\ell(Xt + Z^T\beta)$ is partially differentiable with respect to t and with respect to β_j for all j, with integrable derivatives given by $\mathbb{E}\{XU(Xt + Z^T\beta; Y)\}$ and $\mathbb{E}\{Z_jU(Xt + Z^T\beta; Y)\}$, respectively.

A.2.2 *Proof of Theorem 2*. Suppose first that the Y-model is well specified. Then as $X \perp\!\!\!\perp Y|Z$, we know from (ii) that $\theta = 0$. Standard arguments show that then $(0, \beta^Y)$ maximises $\mathbb{E}\ell(tX + Z^T\beta; Y)$ over $(t, \beta) \in \Theta$ and satisfies the score equations. Thus, $\beta^Y = \beta^\dagger$.

Let us now consider the case where the X-model is linear. We first show that $(\theta, \beta) = (0, \beta^\dagger)$ satisfies (7). By optimality of $\beta^\dagger$, we must have

$$\mathbb{E}\{ZU(Z^T\beta^\dagger; Y)\} = 0, \tag{29}$$

so $(t, \beta) = (0, \beta^\dagger)$ satisfies (7). It suffices to check that this also satisfies (6). We have

$$\begin{aligned}\mathbb{E}\{XU(Z^T\beta^\dagger; Y)\} &= \mathbb{E}[\mathbb{E}\{XU(Z^T\beta^\dagger; Y)|Z\}]\\ &= \mathbb{E}[\mathbb{E}\{(Z^T\beta^X + \varepsilon)U(Z^T\beta^\dagger; Y)|Z\}]\\ &= \mathbb{E}[\mathbb{E}\{\varepsilon U(Z^T\beta^\dagger; Y)|Z, Y\}], \qquad (30)\\ &= \mathbb{E}[U(Z^T\beta^\dagger; Y)\mathbb{E}\{\varepsilon|Z, Y\}] = 0. \qquad (31)\end{aligned}$$

using property (29) of $\beta^\dagger$ in (30) and that $\mathbb{E}\{\varepsilon|Z, Y\} = \mathbb{E}\{\varepsilon|Z\}$ due to the conditional independence $X \perp\!\!\!\perp Y|Z$ in the final line.

A.3 Proof and Regularity Conditions for Theorem 3

A.3.1 *Regularity conditions*. In addition to the regularity conditions laid out in Section A.2.1, we assume the following.

(i) When the Y-model holds, differentiation and integration can be interchanged such that the variance of the score is equal to the Fisher information matrix, that is, $H = V$, and moreover $\mathbb{E}L''(Z^T\beta^*; Y) = 0$.

(ii) The solution (θ^*, β^*) to the score equations is unique.

(iii) $\mathbb{E}U'(Z^T\beta^*) \neq 0$.

(iv) $\mathbb{E}|U'(Z^T\beta^*)| < \infty$, $\mathbb{E}\|Z\|_2^2|U'(Z^T\beta^*)| < \infty$ and $\mathbb{E}\{X^2|U'(Z^T\beta^*)|\} < \infty$.

A.3.2 *Proof of Theorem 3*. When the Y-model holds, we have $H^{-1}VH^{-1} = H^{-1}$, and

$$\frac{\mathbb{E}\{U^2(Z^T\beta^*)\}}{\mathbb{E}\{U'(Z^T\beta^*)\}} = 1$$

as $\mathbb{E}L''(Z^T\beta^*; Y) = 0$. We now turn to the case where the X-model holds. Let $\varepsilon = X - Z^T\beta^X$ and note that $\mathbb{E}(\varepsilon|Z) = 0$. We know from Theorem 2 that $\theta^* = 0$. Let us first compute H. We have

$$-H_{1,j+1} = \mathbb{E}\{XZ_jU'(Z^T\beta^*)\}.$$

Now

$$\begin{aligned}&\mathbb{E}\{XZ_jU'(Z^T\beta^*)|Y, Z\}\\ &\quad = Z_jU'(Z^T\beta^*)\mathbb{E}(Z^T\beta^X + \varepsilon|Y, Z)\\ &\quad = Z_jU'(Z^T\beta^*)Z^T\beta^X.\end{aligned}$$

Here we have used the fact that as $Y \perp\!\!\!\perp Y|Z$, $\mathbb{E}(\varepsilon|Y,Z) = \mathbb{E}(\varepsilon|Z) = 0$. Considering now H_{11}, we have

$$\begin{aligned}&\mathbb{E}\{X^2U'(Z^T\beta^*)|Y,Z\}\\&\quad= \{\mathbb{E}(\varepsilon^2|Y,Z) + (Z^T\beta^X)^2\}U'(Z^T\beta^*).\end{aligned}$$

Thus, writing $A = \mathbb{E}\{ZZ^TU'(Z^T\beta^*)\} \in \mathbb{R}^{p\times p}$, we have

$$H = -\begin{pmatrix}(\beta^X)^TA\beta^X + \mathbb{E}\{\varepsilon^2U'(Z^T\beta^*)\} & (\beta^X)^TA\\ A\beta^X & A\end{pmatrix}.$$

Using standard formulas for the blockwise inverse of matrices in terms of Schur complements, we have that the first column h of H^{-1} satisfies

$$h = \begin{pmatrix}-1\\ \beta^X\end{pmatrix}[\mathbb{E}\{\varepsilon^2U'(Z^T\beta^*)\}]^{-1}.$$

Thus

$$\begin{aligned}(H^{-1}VH^{-1})_{11} &= h^TVh\\ &= \frac{\mathbb{E}\{\varepsilon^2U^2(Z^T\beta^*)\}}{[\mathbb{E}\{\varepsilon^2U'(Z^T\beta^*)\}]^2}.\end{aligned} \tag{32}$$

Now as $\mathbb{E}(\varepsilon^2|Z) = \mathrm{Var}(X|Z) = \mathrm{Var}(X) = \mathbb{E}(\varepsilon^2)$, we have that for any measurable function f of Z with $\mathbb{E}|f(Z)|, \mathbb{E}(|f(Z)|\varepsilon^2) < \infty$,

$$\begin{aligned}\mathbb{E}\{\varepsilon^2f(Z)\} &= \mathbb{E}[f(Z)\mathbb{E}\{\varepsilon^2|Z\}]\\ &= \mathbb{E}\{f(Z)\}\mathbb{E}\{\varepsilon^2\}.\end{aligned}$$

Thus, we have that the quantity in (32) is equal to

$$\frac{\mathbb{E}\{U^2(Z^T\beta^*)\}}{\mathbb{E}(\varepsilon^2)\{\mathbb{E}U'(Z^T\beta^*)\}^2} = -(H^{-1})_{11}\frac{\mathbb{E}\{U^2(Z^T\beta^*)\}}{\mathbb{E}\{U'(Z^T\beta^*)\}}.$$

A.4 Proof and Regularity Conditions for Theorem 4

A.4.1 *Regularity conditions*. In addition to the regularity conditions laid out in Sections A.2.1 and A.3.1, we assume that Θ is compact and that there exists functions $f_1, f_2 : \mathbb{R}^{p+2} \to [0,\infty)$ such that for all $(t,\beta) \in \Theta$,

$$\begin{aligned}|U'(tX + Z^T\beta;Y)| &\le f_1(X,Y,Z),\\ U^2(tX + Z^T\beta;Y) &\le f_2(X,Y,Z)\end{aligned}$$

with $\mathbb{E}f_j(X,Y,Z) < \infty$ for $j = 1,2$. We further assume that $\mathbb{E}U^2(Z^T\beta^*) > 0$ and that U' is continuous.

A.4.2 *Proof of Theorem* 4. From Theorems 2 and 3, it suffices by Slutsky's lemma and the continuous mapping theorem to show that

$$\hat{C}_j \xrightarrow{p} \frac{\mathbb{E}\{U^2(Z^T\beta^*)\}}{\mathbb{E}\{U'(Z^T\beta^*)\}}.$$

Let us consider $j = 2$; the arguments are similar for $j = 1$. By Slustky's lemma, it suffices to show that

$$\frac{1}{n}\sum_{i=1}^n U'(\hat\theta X_i + Z_i^T\beta) \xrightarrow{p} \mathbb{E}U'(tX + Z^T\beta), \tag{33}$$

$$\frac{1}{n}\sum_{i=1}^n U^2(\hat\theta X_i + Z_i^T\beta) \xrightarrow{p} \mathbb{E}U^2(tX + Z^T\beta). \tag{34}$$

Theorem 2 of Jennrich (1969) shows that

$$\sup_{(\beta,t)\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n U'(tX_i + Z_i^T\beta) - \mathbb{E}U'(tX + Z^T\beta)\right| \to 0,$$

$$\sup_{(\beta,t)\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n U^2(tX_i + Z_i^T\beta) - \mathbb{E}U^2(tX + Z^T\beta)\right| \to 0$$

almost surely. By assumption, $(\hat\theta, \hat\beta^Y) \xrightarrow{p} (\theta^*, \beta^*) = (0, \beta^*)$, using Theorem 2 for the final equality. Thus for any subsequence $(m(n))_{n=1}^\infty$, there exists a further subsequence $(l_{m(n)})_{n=1}^\infty$ on which the above convergence is almost sure. Let us write $f(t,\beta) = \mathbb{E}U'(tX + Z^T\beta)$. Then given $\epsilon > 0$, there exists N_1 such that for all $n \ge N_1$,

$$\begin{aligned}&\left|\frac{1}{l_{m(n)}}\sum_{i=1}^{l_{m(n)}} U'(\hat\theta X_i + Z_i^T\hat\beta^Y) - f(\hat\theta, \hat\beta^Y)\right|\\ &\quad\le \sup_{(\beta,t)\in\Theta}\left|\frac{1}{l_{m(n)}}\sum_{i=1}^{l_{m(n)}} U'(tX_i + Z_i^T\beta) - f(t,\beta)\right| < \frac{\epsilon}{2}.\end{aligned}$$

Note that $\hat\theta$ and $\hat\beta^Y$ depend on the sample size, though we have suppressed this in the notation. Meanwhile, by continuity of U' and the continuous mapping theorem, on along $(l_{m(n)})_{n=1}^\infty$ we have

$$U'(\hat\theta X + Z^T\hat\beta^Y) \to U'(Z^T\beta^*)$$

almost surely. Thus by dominated convergence, we have that

$$f(\hat\theta, \hat\beta^Y) \to f(0, \beta^*)$$

along the same subsequence, and so there exists $N_2 \ge N_1$ such that for all $n \ge N_2$

$$|f(\hat\theta, \hat\beta^Y) - f(0,\beta^*)| < \epsilon/2;$$

note that $\hat\theta$ and $\hat\beta^Y$ above are evaluated at sample sizes $l_{m(n)}$ for $n \ge N_2$. Putting things together, we see that on the subsequence $(l_{m(n)})_{n=1}^\infty$, we have

$$\frac{1}{l_{m(n)}}\sum_{i=1}^{l_{m(n)}} U'(\hat\theta X_i + Z_i^T\hat\beta^Y) \to \mathbb{E}U'(Z^T\hat\beta^Y)$$

almost surely. As the original subsequence $(m(n))_{n=1}^\infty$ was arbitrary, we see that (33) holds. The argument to show (34) proceeds similarly.

A.5 Proof of Theorem 5

By symmetry, it is enough to show the result when (Y1) and (Y3)–(Y6) hold. On the event where $\mathbf{R} \neq \mathbf{0}$, we have

$$\begin{aligned}&\frac{(\mathbf{Y} - \mathbf{Z}\hat\beta^Y)^T(\mathbf{X} - \mathbf{Z}\hat\beta^X)}{\|\mathbf{X} - \mathbf{Z}\hat\beta^X\|_2}\\ &\quad= \frac{\mathbf{R}^T}{\|\mathbf{R}\|_2}\mathbf{Z}(\beta^Y - \hat\beta^Y) + \frac{\mathbf{R}^T}{\|\mathbf{R}\|_2}\boldsymbol{\varepsilon}.\end{aligned}$$

The KKT conditions of the Lasso regression of $\mathbf{X}$ on $\mathbf{Z}$ imply $\|\mathbf{Z}^T\mathbf{R}\|_\infty/\|\mathbf{R}\|_2 \le \sqrt{n}\lambda_X$. Thus by Hölder's inequality and (Y4), we have that

$$\begin{aligned}|\mathbf{R}^T\mathbf{Z}(\boldsymbol{\beta}^Y - \hat{\boldsymbol{\beta}}^Y)|/\|\mathbf{R}\|_2\mathbb{1}_{\{\mathbf{R}\neq\mathbf{0}\}} &\le \|\mathbf{R}^T\mathbf{Z}\|_\infty\|\boldsymbol{\beta}^Y - \hat{\boldsymbol{\beta}}^Y\|_1/\|\mathbf{R}\|_2\mathbb{1}_{\{\mathbf{R}\neq\mathbf{0}\}}\\ &= O_{\mathbb{P}}(\sqrt{\log(p)} \times s_Y\sqrt{\log(p)/n}).\end{aligned}$$

From (Y3) we see that $|\mathbf{R}^T\mathbf{Z}(\boldsymbol{\beta}^Y - \hat{\boldsymbol{\beta}}^Y)|/\|\mathbf{R}\|_2\mathbb{1}_{\{\mathbf{R}\neq\mathbf{0}\}} \xrightarrow{p} 0$.

The proof that $\mathbf{R}^T\boldsymbol{\varepsilon}/\|\mathbf{R}\|_2\mathbb{1}_{\{\mathbf{R}\neq\mathbf{0}\}} \xrightarrow{d} \mathcal{N}(0,\sigma^2)$ is identical to the argument used in the proof of Theorem 1 and uses Lemma 8 (see Section A.1). Slutsky's lemma and (Y5) then yield the desired result.

A.6 Proof of Theorem 7

Let $\check{\mathbf{Y}} = \mathbf{D}^Y\tilde{\mathbf{Y}}$ and let $\check{\mathbf{Z}} = (\mathbf{D}^Y\mathbf{Z}\Lambda^Y, \mathbf{D}^X\mathbf{Z}\Lambda^X)$. Note that $\check{\mathbf{Y}} = \check{\mathbf{Z}}\vartheta + \boldsymbol{\varepsilon}^Y$, where $\vartheta \in \mathbb{R}^{2p}$ with $\vartheta_j = (\beta^Y - \hat{\beta}^Y)_j$ for $j \le p$ and $\vartheta_j = 0$ for $j > p$. We seek to bound $\|\check{\vartheta}\|_1$ where

$$\check{\vartheta} \in \underset{b\in\mathbb{R}^{2p}}{\arg\min}\{\|\check{\mathbf{Y}} - \check{\mathbf{Z}}b\|_2/\sqrt{n} + \lambda\|b\|_1\}.$$

Now writing $\check{\sigma} = \|\check{\mathbf{Y}} - \check{\mathbf{Z}}\check{\vartheta}\|_2/\sqrt{n}$, we have that

$$\check{\vartheta} \in \underset{\vartheta\in\mathbb{R}^{2p}}{\arg\min}\{\|\check{\mathbf{Y}} - \check{\mathbf{Z}}b\|_2^2/(2n) + \lambda\check{\sigma}\|b\|_1\}.$$

This may be seen from examining the KKT conditions of each of the optimisations, which are identical, and take the form

$$\frac{1}{n}\check{\mathbf{Z}}^T(\check{\mathbf{Y}} - \check{\mathbf{Z}}\check{\vartheta}) = \lambda\check{\sigma}\nu,$$

where $\|\nu\|_\infty \le 1$ and $\nu_j = \mathrm{sgn}(\check{\vartheta}_j)$ for all j such that $\check{\vartheta}_j \neq 0$. Dotting both sides with $\vartheta - \check{\vartheta}$, we obtain

$$\begin{aligned}&\frac{1}{n}\|\check{\mathbf{Z}}(\vartheta - \check{\vartheta})\|_2^2 + \lambda\check{\sigma}\|\check{\vartheta}\|_1\\ &\quad\le \lambda\check{\sigma}\|\vartheta\|_1 + \frac{1}{n}\|\vartheta - \check{\vartheta}\|_1\|\check{\mathbf{Z}}^T\boldsymbol{\varepsilon}^Y\|_\infty,\end{aligned} \tag{35}$$

where we have used Hölder's inequality to bound $|\nu^T\vartheta| \le \|\nu\|_\infty\|\vartheta\|_1 \le \|\vartheta\|_1$ and $|\check{\mathbf{Z}}(\vartheta - \check{\vartheta})^T\boldsymbol{\varepsilon}^Y| \le \|\vartheta - \check{\vartheta}\|_1 \times \|\check{\mathbf{Z}}^T\boldsymbol{\varepsilon}^Y\|_\infty$, and also the fact that $\check{\vartheta}^T\nu = \|\check{\vartheta}\|_1$. We now aim to show that with high probability,

$$\frac{\|\check{\mathbf{Z}}^T\boldsymbol{\varepsilon}^Y\|_\infty}{n\check{\sigma}} < a\lambda \tag{36}$$

for a constant $a < 1$, where recall that $\lambda = A\sqrt{2\log(p)/n}$ with $A > 1$. We would then have from (35) that on the event in question,

$$\|\check{\vartheta}\|_1 \le \|\vartheta\|_1 + a\|\vartheta - \check{\vartheta}\|_1 \le (1+a)\|\vartheta\|_1 + a\|\check{\vartheta}\|_1,$$

by the triangle inequality, whence

$$\|\check{\vartheta}\|_1 \le \frac{1+a}{1-a}\|\vartheta\|_1 = \frac{1+a}{1-a}\|\beta^Y - \hat{\beta}^Y\|_1,$$

giving the result.

We first observe that by Lemma 2 of Belloni, Chernozhukov and Wang (2011) and also equation (13) therein, for any $B > 1$,

$$\mathbb{P}\biggl(\frac{\|\check{\mathbf{Z}}^T\boldsymbol{\varepsilon}^Y\|_\infty/n}{\|\boldsymbol{\varepsilon}^Y\|_2/\sqrt{n}} \le B\sqrt{\frac{2\log p}{n}}\biggr) \to 1. \tag{37}$$

Now by Lemma 3.1 of van de Geer (2016), writing

$$\hat{\delta} := 2\sqrt{\biggl(\frac{\lambda\|\beta^Y - \hat{\beta}^Y\|_1}{2\|\boldsymbol{\varepsilon}^Y\|_2/\sqrt{n}} + 1\biggr)^2 - 1},$$

we have that the event

$$\Omega_{1n} := \biggl\{\frac{\|\check{\mathbf{Z}}^T\boldsymbol{\varepsilon}^Y\|_\infty/n}{\|\boldsymbol{\varepsilon}^Y\|_2/\sqrt{n}} \le (1-\hat{\delta})\lambda\biggr\}$$

satisfies $\Omega_{1n} \subseteq \Omega_{2n}$ given by

$$\Omega_{2n} := \biggl\{\frac{\|\boldsymbol{\varepsilon}^Y\|_2/\sqrt{n}}{\check{\sigma}} \le \frac{1}{1-\hat{\delta}}\biggr\}.$$

Next for any $1 \ge a > 1/A$, we have writing

$$\Omega_{3n} := \biggl\{\frac{\|\check{\mathbf{Z}}^T\boldsymbol{\varepsilon}^Y\|_\infty/n}{\|\boldsymbol{\varepsilon}^Y\|_2/\sqrt{n}} \le a(1-\hat{\delta})\lambda\biggr\},$$

that $\Omega_{3n} \subseteq \Omega_{1n}$. Thus on Ω_{3n} we have that (36) holds.

Now by the weak law of large numbers and the continuous mapping theorem, $\|\boldsymbol{\varepsilon}^Y\|_2/\sqrt{n} \xrightarrow{p} \sigma$. Moreover $\lambda\|\beta^Y - \hat{\beta}^Y\|_1 \xrightarrow{p} 0$ due to (Y3) and (Y4). Thus from (37), we see that $\mathbb{P}(\Omega_{3n}) \to 1$, proving the first part of the result. The second part of the result is an easy consequence of the first and follows from the same arguments as used to prove Theorem 5.

APPENDIX B: COMPUTATION OF THE SQUARE-ROOT LASSO

Here we explain how the square-root Lasso

$$\hat{\beta}_\lambda^{\mathrm{sq}} := \underset{\beta\in\mathbb{R}^p}{\arg\min}\biggl\{\frac{1}{\sqrt{n}}\|\mathbf{Y} - \mathbf{Z}\beta\|_2 + \lambda\|\beta\|_1\biggr\} \tag{38}$$

may be computed easily given regular Lasso solutions

$$\hat{\beta}_\gamma^{\mathrm{re}} := \underset{\beta\in\mathbb{R}^p}{\arg\min}\biggl\{\frac{1}{2n}\|\mathbf{Y} - \mathbf{Z}\beta\|_2^2 + \gamma\|\beta\|_1\biggr\}. \tag{39}$$

As we will see, a square-root Lasso solution path may be derived from any Lasso solution path via a nondecreasing reparametrisation of the tuning parameter.

Now the minimisers $\hat{\beta}^{\mathrm{sq}}(\lambda)$ and $\hat{\beta}^{\mathrm{re}}(\gamma)$ need not be unique, but the fitted values $\mathbf{Z}\hat{\beta}^{\mathrm{re}}(\gamma)$ of the regular Lasso are always unique. To see this, observe that fixing $\gamma \ge 0$

and taking $\beta^{(1)}$ and $\beta^{(2)}$ as two solutions to (39) achieving minimum value c^*, we have due to the triangle inequality and strict convexity of $\|\cdot\|_2^2$ that

$$\begin{aligned} c^* &\le \frac{1}{2n}\|\mathbf{Y}-\mathbf{Z}(\beta^{(1)}+\beta^{(2)})/2\|_2^2 \\ &\quad + \lambda(\|(\beta^{(1)}+\beta^{(2)})/2\|_1) \\ &\le \frac{1}{2}\frac{1}{2n}(\|\mathbf{Y}-\mathbf{Z}\beta^{(1)}\|_2^2+\|\mathbf{Y}-\mathbf{Z}\beta^{(2)}\|_2^2) \\ &\quad + \frac{\lambda}{2}(\|\beta^{(1)}\|_1+\|\beta^{(2)}\|_1)=c^*. \end{aligned} \tag{40}$$

Thus equality must hold throughout, which can only be the case if $\beta^{(1)}=\beta^{(2)}$.

Let us write

$$\hat{\sigma}_\lambda^{\mathrm{sq}} := \frac{1}{\sqrt{n}}\|\mathbf{Y}-\mathbf{Z}\hat{\beta}_\lambda^{\mathrm{sq}}\|_2,$$
$$\hat{\sigma}_\gamma^{\mathrm{re}} := \frac{1}{\sqrt{n}}\|\mathbf{Y}-\mathbf{Z}\hat{\beta}_\gamma^{\mathrm{re}}\|_2;$$

note that as the fitted values are unique, the latter is uniquely defined though the former may not be.

To establish the relationship between the Lasso and square-root Lasso solutions, observe that the KKT conditions of (38) and (39) are given by

$$\frac{1}{n\hat{\sigma}_\lambda^{\mathrm{sq}}}\mathbf{Z}^T(\mathbf{Y}-\mathbf{Z}\hat{\beta}_\lambda^{\mathrm{sq}}) = \lambda\hat{\nu}_\lambda^{\mathrm{sq}},$$
$$\frac{1}{n}\mathbf{Z}^T(\mathbf{Y}-\mathbf{Z}\hat{\beta}_\gamma^{\mathrm{re}}) = \gamma\hat{\nu}_\gamma^{\mathrm{re}},$$

where $\|\hat{\nu}_\lambda^{\mathrm{sq}}\|_\infty \le 1$, and $\hat{\nu}_\lambda^{\mathrm{sq}}$ agrees in sign with $\hat{\beta}_\lambda^{\mathrm{sq}}$ on its active set (and similarly for $\hat{\nu}_\gamma^{\mathrm{re}}$), provided $\hat{\sigma}_\lambda^{\mathrm{sq}} > 0$. Comparing the KKT conditions above, we see that any Lasso solution $\hat{\beta}_\gamma^{\mathrm{re}}$ is a square-root Lasso solution $\hat{\beta}_\lambda^{\mathrm{sq}}$ with $\lambda=\gamma/\hat{\sigma}_\gamma^{\mathrm{re}}$ (provided $\hat{\sigma}_\gamma^{\mathrm{re}}>0$). Conversely, any square-root Lasso solution $\hat{\beta}_\lambda^{\mathrm{sq}}$ is equal to a Lasso solution $\hat{\beta}_\gamma^{\mathrm{re}}$ with $\gamma=\lambda\hat{\sigma}_\lambda^{\mathrm{sq}}$, provided $\hat{\sigma}_\lambda^{\mathrm{sq}}>0$.

LEMMA 9. *Let γ^* be maximal such that $\hat{\sigma}_{\gamma^*}^{re}=0$. The function $\gamma \mapsto \gamma/\hat{\sigma}_\gamma^{re}$ defined on (γ^*,∞) is nondecreasing.*

The result above shows that given a square-root Lasso tuning parameter λ, we may find via a bisection search the Lasso tuning parameter γ such that $\gamma/\hat{\sigma}_\gamma^{\mathrm{re}}=\lambda$ and thereby obtain a square-root Lasso solution.

B.1 Proof of Lemma 9

The conclusion is equivalent to the following: for any $\gamma_1,\gamma_2\in(\gamma^*,\infty)$ with

$$\lambda_1 := \frac{\gamma_1}{\hat{\sigma}_{\gamma_1}^{\mathrm{re}}} < \frac{\gamma_2}{\hat{\sigma}_{\gamma_2}^{\mathrm{re}}} =: \lambda_2,$$

we have $\gamma_1<\gamma_2$. Let us write $\beta^{(1)}=\hat{\beta}_{\gamma_1}^{\mathrm{re}}$ and $\beta^{(2)}=\hat{\beta}_{\gamma_2}^{\mathrm{re}}$, noting that while these need not be unique, the corresponding fitted values and ℓ_1-norms are. Then as $\beta^{(1)}$ and $\beta^{(2)}$ are square-root Lasso solutions at λ_1 and λ_2 respectively, we have that

$$\begin{aligned} &\frac{1}{\sqrt{n}}\|\mathbf{Y}-\mathbf{Z}\beta^{(1)}\|_2+\lambda_1\|\beta^{(1)}\|_1 \\ &\qquad \le \frac{1}{\sqrt{n}}\|\mathbf{Y}-\mathbf{Z}\beta^{(2)}\|_2+\lambda_1\|\beta^{(2)}\|_1, \\ &\frac{1}{\sqrt{n}}\|\mathbf{Y}-\mathbf{Z}\beta^{(2)}\|_2+\lambda_2\|\beta^{(2)}\|_1 \\ &\qquad \le \frac{1}{\sqrt{n}}\|\mathbf{Y}-\mathbf{Z}\beta^{(1)}\|_2+\lambda_2\|\beta^{(1)}\|_1. \end{aligned} \tag{41}$$

Adding these inequalities, we deduce that

$$\lambda_1\|\beta^{(1)}\|_1+\lambda_2\|\beta^{(2)}\|_1 \le \lambda_1\|\beta^{(2)}\|_1+\lambda_2\|\beta^{(1)}\|_1.$$

Rearranging, we obtain

$$(\lambda_2-\lambda_1)(\|\beta^{(2)}\|_1-\|\beta^{(1)}\|_1)\ge 0,$$

and so dividing by $\lambda_2-\lambda_1>0$ we conclude that $\|\beta^{(2)}\|_1\ge\|\beta^{(1)}\|_1$. Substituting this into (41), we see that $\hat{\sigma}_{\gamma_1}^{\mathrm{re}}\le\hat{\sigma}_{\gamma_2}^{\mathrm{re}}$, so $\gamma_1=\hat{\sigma}_{\gamma_1}^{\mathrm{re}}\lambda_1<\hat{\sigma}_{\gamma_2}^{\mathrm{re}}\lambda_2=\gamma_2$ as required.

APPENDIX C: CONFIDENCE REGIONS FOR $w^T\beta^0$

In this section, we consider a linear model $\mathbf{Y}=\mathbf{Z}\beta^0+\boldsymbol{\varepsilon}$ and consider the problem of finding a confidence interval for $w^T\beta^0$ for a given $w\in\mathbb{R}^p$. When $w=e_j$ for a standard basis vector $e_j\in\mathbb{R}^p$, the methodology set out in Section 3.2 may be used to obtain a confidence region even in the case where only a partially linear model holds. For more general w, these methods must be adapted and here we will need to assume the linear model above holds with β^0 sufficiently sparse. We describe these modifications below.

First, consider testing a null hypothesis H_0: $w^T\beta^0=0$. Let $P=ww^T/\|w\|_2^2$. Note $w^T\beta^0=0$ if and only if $(I-P)\beta^0=\beta^0$, so the null model may be expressed as

$$\mathbf{Y}=(I-P)\mathbf{Z}\beta^0+\boldsymbol{\varepsilon}. \tag{42}$$

Let

$$\hat{\beta}=\operatorname*{arg\,min}_{\beta\in\mathbb{R}^p}\{\|\mathbf{Y}-\mathbf{Z}(I-P)\beta\|_2/\sqrt{n}+\lambda\|\beta\|_1\}. \tag{43}$$

Note that under H_0 we should have

$$\|\hat{\beta}-\beta^0\|_1=O_{\mathbb{P}}(s\sqrt{\log(p)/n})$$

for $\lambda=A\sqrt{2\log(p)/n}$ with $A>1$ and where $s=|\{j:\beta_j^0\ne 0\}|$. Also let $\mathbf{R}\in\mathbb{R}^n$ be the vector of residuals from the regression

$$\underset{\beta \in \mathbb{R}^p}{\arg\min}\{\|\mathbf{Z}w - \mathbf{Z}(I-P)\beta\|_2/\sqrt{n} + \lambda\|\beta\|_1\}. \tag{44}$$

Note that $\mathbf{R}$ thus defined enjoys a near-orthogonality property of the form $(I-P)\mathbf{Z}^T\mathbf{R}/\|\mathbf{R}\|_2 \leq \sqrt{n}\lambda$. The reason for aiming to orthogonalise $\mathbf{Z}w$ is that were we to have $w^T\beta^0 \neq 0$, the residuals from the regression (43) should have expectation close to $\mathbf{Z}P\beta^0 \propto \mathbf{Z}w$. Thus a test statistic involving dotting these residuals with something close to the direction of $\mathbf{Z}w$ should be large in magnitude under an alternative.

We thus consider the test statistic given by

$$T = \sqrt{n}\frac{\mathbf{R}^T\{\mathbf{Y} - \mathbf{Z}(I-P)\hat{\beta}\}}{\|\mathbf{R}\|_2\|\mathbf{Y} - \mathbf{Z}(I-P)\hat{\beta}\|_2}. \tag{45}$$

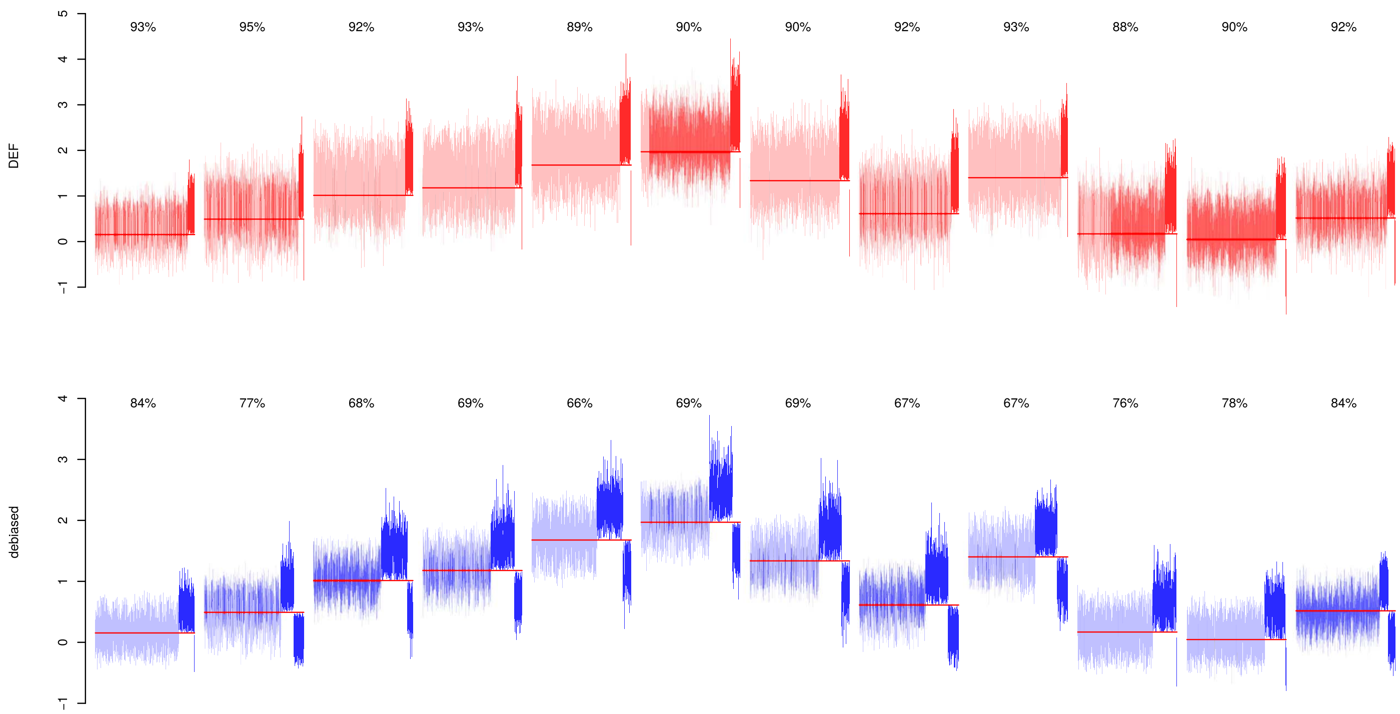

FIG. 8. *The linear setting (a) with Toeplitz design; the interpretation is similar to that of Figure 4.*

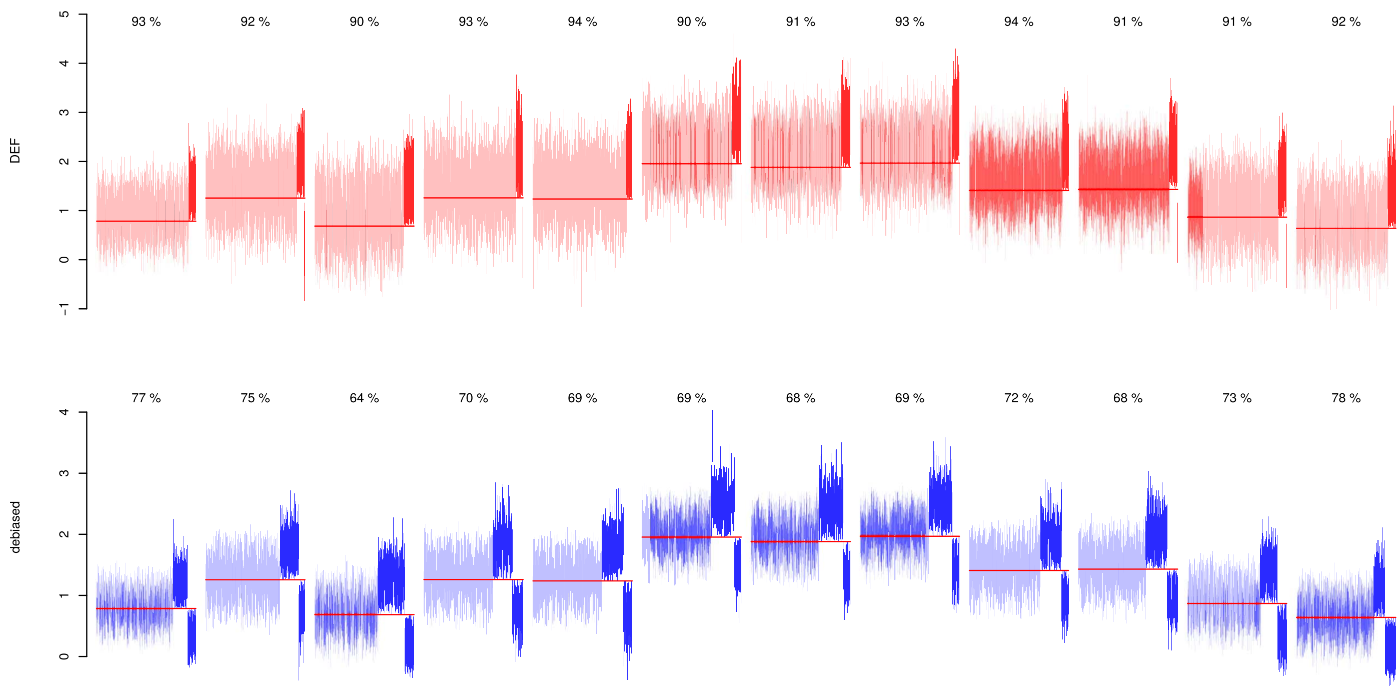

FIG. 9. *The slightly nonlinear setting (b) with Toeplitz design; the interpretation is similar to that of Figure 4.*

Writing $\hat{\sigma} = \|\mathbf{Y} - \mathbf{Z}(I - P)\hat{\beta}\|_2/\sqrt{n}$, we have

$$T = \frac{1}{\hat{\sigma}} \frac{\mathbf{R}^T}{\|\mathbf{R}\|_2} \boldsymbol{\varepsilon} + \frac{1}{\hat{\sigma}} (\hat{\beta} - \beta^0)^T (I - P)\mathbf{Z}^T \frac{\mathbf{R}}{\|\mathbf{R}\|_2}$$
$$=: \text{(i)} + \text{(ii)}.$$

Term (i) will be well approximated by a standard normal under reasonable conditions, and term (ii) may be bounded in absolute value using an argument similar to that presented in Section 3.1. Thus under appropriate conditions, we will have $T \xrightarrow{d} \mathcal{N}(0, 1)$.

Now consider testing $H_0(t)$: $w^T \beta^0 = t$. Observe that

$$\mathbf{Y} - t\mathbf{Z}w/\|w\|_2^2 = \mathbf{Z}\beta^0 - \mathbf{Z}P\beta^0 + \boldsymbol{\varepsilon} =: \mathbf{Y}^{(t)},$$

so the new response $\mathbf{Y}^{(t)}$ respects the null model (42). We may thus test $H_0(t)$ using test statistic T_t defined as in (45) but computed using the response $\mathbf{Y}^{(t)}$ in place of $\mathbf{Y}$.

Then to form a $1 - \alpha$ confidence region for $w^T \beta^0$ we can simply invert the tests as in Section 3.2:

$$R_\alpha := \{t \in \mathbb{R} : |T_t| \geq z_\alpha\}.$$

Provided $\mathbb{P}(H_0(w^T \beta^0) \text{ rejected}) \geq 1 - \alpha$, the confidence region R_α will satisfy $\mathbb{P}(w^T \beta^0 \in R_\alpha) \geq 1 - \alpha$; see Corollary 6.

We note that compared to the confidence regions constructed in Cai and Guo (2017), which are introduced primarily for theoretical purposes, our confidence region does not require prior knowledge of the the inverse covariance of $\mathbf{Z}$, the sparsity of β^0, or the noise level $\text{Var}(\varepsilon_1)$. Our construction is related to that in Zhu and Bradic (2018b), but where we require sparsity of β^0, Zhu and Bradic (2018b) instead require sparsity of a projection of the quantity 'estimated' by the minimiser in (44). In fact, with such an assumption, it is straightforward to see that we can still expect T to have an asymptotically normal distribution regardless of the sparsity of β^0 by reversing the roles of (43) and (44): we only use the former to establish approximate orthogonality while we exploit assumed small estimation error of the latter. An additional difference is that the approach in Zhu and Bradic (2018b) requires solving a family of large-scale linear programs, whereas our region requires only standard software for computing the Lasso.

APPENDIX D: ADDITIONAL NUMERICAL RESULTS

Here we present the results of analogous numerical experiments to those described in Section 4.1, but with the multivariate distribution P used for generating predictors (X_i, Z_i) replaced with a multivariate Gaussian distribution $\mathcal{N}_p(0, \Sigma)$. We take the covariance matrix Σ to have a Toeplitz design with $\Sigma_{jk} = 0.9^{|j-k|}$. Note that the inverse of Σ is tridiagonal and so the X-model is a sparse linear model (with sparsity level $s_X = 1$). The settings considered here thus satisfy the conditions of Theorem 5.

We see that compared to the more challenging settings investigated in Section 4.1, the coverage properties of both confidence interval construction methods are improved; however, the debiased Lasso still undercovers while the DEF confidence intervals reach a coverage of closer to 95%.

ACKNOWLEDGMENTS

The authors would like to thank Nicolai Meinshausen for many helpful conversations, and also coining the term "double estimation friendly".

FUNDING

The first author was supported by an EPSRC Programme Grant EP/N031938/1 and an EPSRC First Grant EP/R013381/1.

The second author was supported by the European Research Council under the Grant Agreement No 786461 (CausalStats—ERC-2017-ADG).

REFERENCES

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189 https://doi.org/10.1111/j.1541-0420.2005.00377.x

BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324 https://doi.org/10.1093/biomet/asr043

BOX, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Assoc.* **71** 791–799. MR0431440

BRADIC, J., WAGER, S. and ZHU, Y. (2019). Sparsity double robust inference of average treatment effects. Preprint. Available at arXiv:1905.00744.

BRILLINGER, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann*. *Wadsworth Statist./Probab. Ser.* 97–114. Wadsworth, Belmont, CA. MR0689741

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

BÜHLMANN, P. and VAN DE GEER, S. (2015). High-dimensional inference in misspecified linear models. *Electron. J. Stat.* **9** 1449–1473. MR3367666 https://doi.org/10.1214/15-EJS1041

BUJA, A., BROWN, L., BERK, R., GEORGE, E., PITKIN, E., TRASKIN, M., ZHANG, K. and ZHAO, L. (2019a). Models as approximations I: Consequences illustrated with linear regression. *Statist. Sci.* **34** 523–544. MR4048582 https://doi.org/10.1214/18-STS693

BUJA, A., BROWN, L., KUCHIBHOTLA, A. K., BERK, R., GEORGE, E. and ZHAO, L. (2019b). Models as approximations II: A model-free theory of parametric regression. *Statist. Sci.* **34** 545–565. MR4048583 https://doi.org/10.1214/18-STS694

CAI, T. T. and GUO, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395 https://doi.org/10.1214/16-AOS1461

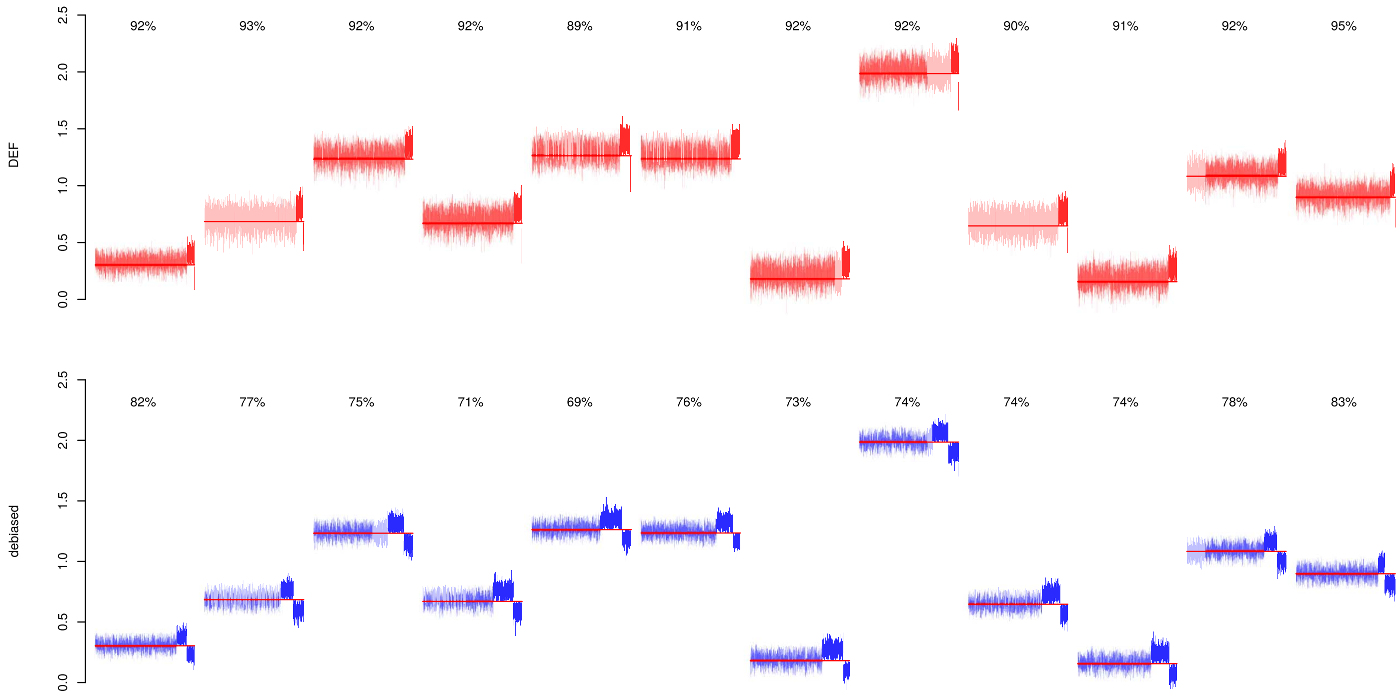

FIG. 10. *The highly nonlinear setting* (*c*) *with Toeplitz design*; *the interpretation is similar to that of Figure* 4.

CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265

CAO, W., TSIATIS, A. A. and DAVIDIAN, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96** 723–734. MR2538768 https://doi.org/10.1093/biomet/asp033

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. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

BÜHLMANN, P., KALISCH, M. and MEIER, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annu. Rev. Stat. Appl.* **1** 255–278.

DEZEURE, R., BÜHLMANN, P., MEIER, L. and MEINSHAUSEN, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statist. Sci.* **30** 533–558. MR3432840 https://doi.org/10.1214/15-STS527

DUAN, N. and LI, K.-C. (1991). Slicing regression: A link-free regression method. *Ann. Statist.* **19** 505–530. MR1105834 https://doi.org/10.1214/aos/1176348109

DUKES, O., AVAGYAN, V. and VANSTEELANDT, S. (2020). Doubly robust tests of exposure effects under high-dimensional confounding. *Biometrics* **76** 1190–1200. MR4186835 https://doi.org/10.1111/biom.13231

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166 https://doi.org/10.1214/009053604000000067

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity*: *The Lasso and Generalizations*. *Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. MR3616141

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley*, *Calif.*, 1965/66), *Vol. I*: *Statistics* 221–233. Univ. California Press, Berkeley, CA. MR0216620

JANKOVÁ, J., SHAH, R. D., BÜHLMANN, P. and SAMWORTH, R. J. (2020). Goodness-of-fit testing in high dimensional generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 773–795. MR4112784 https://doi.org/10.1111/rssb.12371

JENNRICH, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Stat.* **40** 633–643. MR0238419 https://doi.org/10.1214/aoms/1177697731

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458 https://doi.org/10.1214/07-STS227

LI, K.-C. and DUAN, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1052. MR1015136 https://doi.org/10.1214/aos/1176347254

MACKINNON, J. G. and WHITE, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J. Econometrics* **29** 305–325.

NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45** 158–195. MR3611489 https://doi.org/10.1214/16-AOS1448

REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. MR3346695 https://doi.org/10.1214/14-AOS1286

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** 479–495. MR1173493 https://doi.org/10.2307/2532304

ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90** 122–129. MR1325119

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

ROTNITZKY, A., LEI, Q., SUED, M. and ROBINS, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99** 439–456. MR2931264 https://doi.org/10.1093/biomet/ass013

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478 https://doi.org/10.2307/2669923

SHAH, R. D. and BÜHLMANN, P. (2018). Goodness-of-fit tests for high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 113–135. MR3744714 https://doi.org/10.1111/rssb.12234

SHAH, R. D. and PETERS, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.* **48** 1514–1538. MR4124333 https://doi.org/10.1214/19-AOS1857

SMUCLER, E., ROTNITZKY, A. and ROBINS, J. M. (2019). A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. Preprint. Available at arXiv:1904.03737.

SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166 https://doi.org/10.1093/biomet/ass043

SUN, T. and ZHANG, C.-H. (2013). Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.* **14** 3385–3418. MR3144466

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VAN DE GEER, S. (2016). *Estimation and Testing Under Sparsity*. *Lecture Notes in Math.* **2159**. Springer, Cham. MR3526202 https://doi.org/10.1007/978-3-319-32774-7

VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316 https://doi.org/10.1214/09-EJS506

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics*: *A Non-asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163 https://doi.org/10.2307/1912526

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

ZHU, Y. and BRADIC, J. (2018a). Significance testing in non-sparse high-dimensional linear models. *Electron. J. Stat.* **12** 3312–3364. MR3861831 https://doi.org/10.1214/18-EJS1443

ZHU, Y. and BRADIC, J. (2018b). Linear hypothesis testing in dense high-dimensional linear models. *J. Amer. Statist. Assoc.* **113** 1583–1600. MR3902231 https://doi.org/10.1080/01621459.2017.1356319