Doubly Robust Semiparametric Inference Using Regularized Calibrated Estimation with High-dimensional Data

SATYAJIT GHOSH¹, ZHIQIANG TAN¹,*

¹Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, USA
E-mail: * ztan@stat.rutgers.edu

Consider semiparametric estimation where a doubly robust estimating function for a low-dimensional parameter is available, depending on two working models. With high-dimensional data, we develop regularized calibrated estimation as a general method for estimating the parameters in the two working models, such that valid Wald confidence intervals can be obtained for the parameter of interest under suitable sparsity conditions if either of the two working models is correctly specified. We propose a computationally tractable two-step algorithm and provide rigorous theoretical analysis which justifies sufficiently fast rates of convergence for the regularized calibrated estimators in spite of sequential construction and establishes a desired asymptotic expansion for the doubly robust estimator. As concrete examples, we discuss applications to partially linear, log-linear, and logistic models and estimation of average treatment effects. Numerical studies in the former three examples demonstrate superior performance of our method, compared with debiased Lasso.

Keywords: Average treatment effect; Calibration estimation; Debiased Lasso; Double robustness; High-dimensional data; Lasso penalty; Partially linear model; Semiparametric estimation

1. Introduction

Semiparametric modeling and estimation aims to draw inference about low-dimensional parameters of interest, while allowing flexible specification for nuisance parameters, which are often in the form of smooth functions of covariates (Bickel et al. [3]). With low-dimensional covariates, various methods and theory have been developed, using nonparametric smoothing techniques to estimate those unknown functions. There are increasing difficulties, as the complexity of functions increases with a fixed number of covariates, or the number of covariates increases with parametric specifications for the unknown functions. These two problems are fundamentally related. For concreteness, we focus on the latter setting, where the number of covariates is large, while the unknown functions are modeled using known basis functions, for example, main effects or interactions. This setting also allows connections to high-dimensional statistics (Bühlmann & Van de Geer [5]).

In this article, we study a broad class of semiparametric problems, where a doubly robust estimating function \( \tau(U; \theta, \alpha, \gamma) \) for the parameter of interest \( \theta \) is available as follows. Here \( U \) denotes a data vector including a possibly high-dimensional covariate vector \( X \), and \((\alpha, \gamma)\) are two nuisance parameters defined through working models \( g(x; \alpha) \) and \( f(x; \gamma) \) for unknown functions \( g^*(x) \) and \( f^*(x) \). The estimating function \( \tau \) is assumed to be unbiased, \( E\{\tau(U; \theta, \alpha, \gamma)\} = 0 \), when \( \theta \) is set to the true value \( \theta^* \), and either \( \alpha \) or \( \gamma \), but not necessarily both, is set to the true value \( \alpha^* \) or \( \gamma^* \) defined respectively such that \( g(x; \alpha^*) = g^*(x) \)
or $f(x; \gamma^*) \equiv f^*(x)$ if model $g(\cdot; \alpha)$ or $f(\cdot; \gamma)$ is correctly specified. In general, doubly robust estimation using $\tau$ consists of two stages: some estimators $(\hat{\alpha}, \hat{\gamma})$ are first defined, and then $\hat{\theta}$ is defined by solving the estimating equation $E\{\tau(U; \theta, \hat{\alpha}, \hat{\gamma})\} = 0$, where $E()$ denotes a sample average. Conventionally, the estimators $(\hat{\alpha}, \hat{\gamma})$ are derived by maximum likelihood or variations associated with models $g(\cdot; \alpha)$ and $f(\cdot; \gamma)$ for $g^*$ and $f^*$.

While such doubly robust estimation is perhaps most extensively studied in missing-data problems and estimation of average treatment effects (Scharfstein et al. [30]; Kang & Schafer [16]; Tan [35]), doubly robust methods have been developed in various semiparametric problems, including partially linear and log-linear models (Robins & Rotnitzky [27]), instrumental variable analysis (Tan [34]; Okui et al. [25]), mediation analysis (Tchetgen Tchetgen & Shpitser [40]), and dimension reduction (Ma & Zhu [18]) among others. As a somewhat under-appreciated result, we point out that the familiar least-squares estimator for each individual coefficient in linear regression is doubly robust in the context of a partially linear model. This result is also closely related to debiased Lasso estimation in high-dimensional linear regression (Zhang & Zhang [46]; Van de Geer et al. [41]; Javanmard & Montanari [15]). See Examples 5 and 10 for further discussion.

The main contribution of our work can be summarized as follows. Given a doubly robust estimating function $\tau$, we develop a general method as an alternative to maximum likelihood for constructing estimators $(\hat{\alpha}, \hat{\gamma})$ of nuisance parameters, which are used to define an estimator $\hat{\theta}$ as a solution to $E\{\tau(U; \theta, \hat{\alpha}, \hat{\gamma})\} = 0$. For this method, the limit values $(\bar{\alpha}, \bar{\gamma})$ of $(\hat{\alpha}, \hat{\gamma})$ are designed to satisfy a pair of population estimating equations, called calibration equations. If either model $g(\cdot; \alpha)$ or $f(\cdot; \gamma)$ is correctly specified, then the resulting estimator $\hat{\theta}$ can be shown to be not only consistent for $\theta^*$, but also achieve an asymptotic expansion in the following manner under suitable conditions with a sample size $n$.

- In low-dimensional settings, the expansion of $\hat{\theta}$ is in the usual order $O_p(n^{-1/2})$, but not affected by the variation of $(\hat{\alpha}, \hat{\gamma})$, which is also of order $O_p(n^{-1/2})$.
- In high-dimensional settings, the expansion of $\hat{\theta}$ remains in the order $O_p(n^{-1/2})$, even though the convergence of $(\hat{\alpha}, \hat{\gamma})$ to $(\bar{\alpha}, \bar{\gamma})$ is slower than $O_p(n^{-1/2})$.

In fact, with high-dimensional data, we propose a computationally tractable two-step algorithm using Lasso regularized estimation. We provide rigorous theoretical analysis which justifies sufficiently fast convergence rates for $(\hat{\alpha}, \hat{\gamma})$ in spite of sequential construction and establishes the desired asymptotic expansion and variance estimation for $\hat{\theta}$. Doubly robust Wald confidence intervals can be obtained, based on $\hat{\theta}$ and consistent variance estimation. As concrete examples, we discuss applications to partially linear, log-linear, and logistic models and a missing-response problem related to estimation of average treatment effects.

**Related work.** There is an extensive literature related to our work. In low-dimensional settings, estimating equations similar to our calibration equations are proposed by Vermeulen & Vansteelandt [43], where a similar asymptotic expansion similar as described above is obtained. The two methods are equivalent in some problems such as estimation of average treatment effects, where a similar method is also proposed in Kim & Haziza [17]. However, there exists a technical difference: estimating equations in Vermeulen & Vansteelandt [43] are defined from the influence function of a doubly robust estimator, whereas our calibration equations are defined more generally from a doubly robust estimating function, which includes but not limited to the influence function. This difference partly reflects different motivations. The approach of Vermeulen & Vansteelandt [43] is motivated to locally minimize the first-order bias, defined directly as the expectation of the influence function. Our approach aims to derive a doubly...
robust estimator which achieves a desired asymptotic expansion, insensitive to the variations from estimation of nuisance parameters. For instance, see Examples 6 and 7 for differences of the two methods in partially linear and log-linear models.

In high-dimensional settings, doubly robust estimating functions are used with regularized likelihood (or quasi-likelihood) estimators of \((\alpha, \gamma)\) in Belloni et al. [2] and Farrell [12]. Valid confidence intervals are established under suitable sparsity conditions, when both models \(g(\cdot; \alpha)\) and \(f(\cdot; \gamma)\) are correctly specified. For inference about average treatment effects, doubly robust confidence intervals are obtained in Tan [37] if either a propensity score model or a linear outcome model is correctly specified. In this case, regularized calibration estimators of \(\alpha\) and \(\gamma\) are determined sequentially, independent of \(\Theta\). For a nonlinear outcome model, only model-assisted confidence intervals are established, being valid when a propensity score model is correctly specified but the outcome model may be misspecified. In this case and other problems (see Examples 7–9), there are computational and theoretical complications due to coupled calibration equations. To tackle these issues, we develop the two-step algorithm and appropriate high-dimensional analysis, to obtain doubly robust confidence intervals which are not only computationally tractable but also theoretically justified in general settings where doubly robust estimating functions are available.

For estimating average treatment effects, Avagyan & Vansteelandt [1] proposed a regularized version of estimating equations in Vermeulen & Vansteelandt [43]. But their theoretical analysis appears to presume standard convergence rates for the estimators of \((\alpha, \gamma)\) in Belloni et al. [2] and Farrell [12]. Valid confidence intervals are established under suitable sparsity conditions, when both models \(g(\cdot; \alpha)\) and \(f(\cdot; \gamma)\) are correctly specified. For inference about average treatment effects, doubly robust confidence intervals are obtained in Tan [37] if either a propensity score model or a linear outcome model is correctly specified. In this case, regularized calibration estimators of \(\alpha\) and \(\gamma\) are determined sequentially, independent of \(\Theta\). For a nonlinear outcome model, only model-assisted confidence intervals are established, being valid when a propensity score model is correctly specified but the outcome model may be misspecified. In this case and other problems (see Examples 7–9), there are computational and theoretical complications due to coupled calibration equations. To tackle these issues, we develop the two-step algorithm and appropriate high-dimensional analysis, to obtain doubly robust confidence intervals which are not only computationally tractable but also theoretically justified in general settings where doubly robust estimating functions are available.

For estimating average treatment effects, Avagyan & Vansteelandt [1] proposed a regularized version of estimating equations in Vermeulen & Vansteelandt [43]. But their theoretical analysis appears to presume standard convergence rates for the estimators of \((\alpha, \gamma)\) without handling additional data-dependency in loss functions. Ning et al. [24] proposed doubly robust confidence intervals, but their methods achieve model double robustness: an estimator of \(\Theta\) is correctly specified but the outcome model may be misspecified. In this case and other problems (see Examples 7–9), there are computational and theoretical complications due to coupled calibration equations. To tackle these issues, we develop the two-step algorithm and appropriate high-dimensional analysis, to obtain doubly robust confidence intervals which are not only computationally tractable but also theoretically justified in general settings where doubly robust estimating functions are available.

For estimating average treatment effects, Avagyan & Vansteelandt [1] proposed a regularized version of estimating equations in Vermeulen & Vansteelandt [43]. But their theoretical analysis appears to presume standard convergence rates for the estimators of \((\alpha, \gamma)\) without handling additional data-dependency in loss functions. Ning et al. [24] proposed doubly robust confidence intervals, but their methods achieve model double robustness: an estimator of \(\Theta\) is correctly specified but the outcome model may be misspecified. In this case and other problems (see Examples 7–9), there are computational and theoretical complications due to coupled calibration equations. To tackle these issues, we develop the two-step algorithm and appropriate high-dimensional analysis, to obtain doubly robust confidence intervals which are not only computationally tractable but also theoretically justified in general settings where doubly robust estimating functions are available.

For partially linear and log-linear models (as in Examples 1–2), Dukes et al. [10] and Dukes & Vansteelandt [11] developed score tests and associated confidence intervals which are model-assisted: being valid when model \(f(\cdot; \gamma)\) is correctly specified but model \(g(\cdot; \alpha)\) may be misspecified. In the case of linear model \(g(\cdot; \alpha)\), their methods achieve model double robustness
as mentioned above, and with an additional location-shift condition, remain valid under the sparsity condition $s_5 \max(s_5, s_4) = o(n)$, up to a $\log(p)$ term, without sampling splitting.

Finally, our work is also connected to debiased Lasso mentioned earlier and extensions (Neykov et al. [22]) to obtain confidence intervals and tests for low-dimensional coefficients in high-dimensional models. These methods in general do not achieve double robustness. See Examples 10–12 on partially linear models for further discussion.

2. Double robustness and calibrated estimation

2.1. Doubly robust estimation

Let $\{U_i : i = 1, \ldots, n\}$ be independent and identically distributed observations as $U$, which is assumed to include a covariate vector $X$ taking values $x$ in a space $\mathcal{X}$. Consider semiparametric estimation based on an estimating equation

$$ 0 = \bar{E}\{\tau(U; \theta, g, f)\} = \frac{1}{n} \sum_{i=1}^{n} \tau(U_i; \theta, g, f), \quad (1) $$

where $\bar{E}(\cdot)$ denotes a sample average, $\tau(U; \theta, g, f)$ is an estimating function, $\theta$ is a scalar parameter of interest in $\Theta$, and $g$ and $f$ are two variation-independent nuisance parameters, defined in some function spaces $\mathcal{G}$ and $\mathcal{F}$ on $\mathcal{X}$. Denote by $(\theta^*, g^*, f^*)$ the true values (i.e., data-generating values) of $(\theta, g, f)$. Assume that the estimating function $\tau(U; \theta, g, f)$ is doubly robust in satisfying the following two properties:

$$ 0 = E\{\tau(U; \theta^*, g^*, f)\} \text{ for any } f \in \mathcal{F}, \quad (2) $$

$$ 0 = E\{\tau(U; \theta^*, g, f^*)\} \text{ for any } g \in \mathcal{G}. \quad (3) $$

In other words, $\tau(U; \theta, g, f)$ is unbiased for estimation of $\theta^*$ if either $g = g^*$ or $f = f^*$. Several examples of doubly robust estimating functions are as follows. Construction of doubly robust estimating functions is problem-dependent and not discussed here. See Robins & Rotnitzky [27], Tchetgen Tchetgen et al. [39] and Tan [36] among others.

Example 1. Suppose that an outcome $Y$ is related to a covariate $Z$ and additional covariates $X$ in a partially linear model

$$ E(Y|Z,X) = \theta^* Z + g^*(X), \quad (4) $$

where $\theta^*$ is the true value of a coefficient $\theta$ and $g^*(x)$ is the true value of a function $g(x)$. In addition to $g(\cdot)$, define a nuisance parameter $f(\cdot)$ such that $f^*(X) = E(Z|X)$. Then the following estimating function is doubly robust (Robins & Rotnitzky [27]),

$$ \tau(U; \theta, g, f) = \{Y - \theta Z - g(X)\}\{Z - f(X)\}, \quad (5) $$

where $U = (Y, Z, X)$. The true value $\theta^*$ can be regarded as a homogeneous additive treatment effect, in the setting where $Z$ is a treatment variable.

Example 2. Consider a partially log-linear model

$$ E(Y|Z,X) = \exp\{\theta^* Z + g^*(X)\}, \quad (6) $$
Doubly robust semiparametric inference

where $\theta^*$ is the true value of a coefficient $\theta$ and $g^*(x)$ is the true value of a function $g(x)$. The nuisance parameter $f(x)$ is still defined such that $f^*(x) = E(Z|X)$. Then the following estimating function is doubly robust (Robins & Rotnitzky [27]),

$$
\tau(U; \theta, g, f) = \{Ye^{-\theta Z} - e^{g(X)}\}\{Z - f(X)\},
$$

where $U = (Y, Z, X)$. The true value $\theta^*$ can be regarded as a homogeneous multiplicative treatment effect, in the setting where $Z$ is a treatment variable.

Example 3. Consider a partially logistic model with binary $Y$,

$$
E(Y|Z, X) = \expit(\theta^* Z + g^*(X)),
$$

where $\expit(c) = (1 + e^{-c})^{-1}$, $\theta^*$ is the true value of $\theta$ and $g^*(x)$ is the true value of $g(x)$. In contrast with Examples 1–2, define a nuisance parameter $f(\cdot)$ such that $f^*(X) = E(Z|Y = 0, X)$. Then a doubly robust estimating function is (Tan [36])

$$
\tau(U; \theta, g, f) = e^{-\theta Z Y}\{Y - \expit(g(X))\}\{Z - f(X)\},
$$

where $U = (Y, Z, X)$. The true value $\theta^*$ can be regarded as a homogeneous treatment effect in the scale of log odds, in the setting where $Z$ is a treatment variable.

Example 4. Let $Y$ be an outcome variable, $X$ a covariate vector, and $Z$ a binary variable such that $Z = 1$ or 0 if $Y$ is observed or missing respectively. Assume that the missing data mechanism is ignorable: $Y$ and $Z$ are conditionally independent given $X$ (Rubin [29]). It is of interest to estimate the mean $\theta^* = E(Y)$. The nuisance parameters $g(\cdot)$ and $f(\cdot)$ are defined such that the true values are

$$
g^*(X) = E(Y|Z = 1, X), \quad f^*(X) = P(Z = 1|X),
$$

which are called outcome regression function and propensity score. Then the following estimating function is doubly robust (Scharfstein et al. [30]),

$$
\tau(U; \theta, g, f) = \frac{ZY}{f(X)} - \left\{ \frac{Z}{f(X)} - 1 \right\} g(X) - \theta,
$$

where $U = (ZY, Z, X)$. The true value $\theta^*$ represents the mean of a potential outcome associated with a treatment when $Z$ encodes the receipt of the treatment.

Typically, estimating equation (1) is used in the form of two-stage semiparametric estimation, depending on some modeling restrictions, $g(x; \alpha)$ and $f(x; \gamma)$ with parameters $\alpha$ and $\gamma$, postulated on $(g^*, f^*)$. For concreteness, consider the following two models,

$$
g^*(x) = g(x; \alpha) = \psi_g\{\alpha^T \xi(x)\},
$$

$$
f^*(x) = f(x; \gamma) = \psi_f\{\gamma^T \xi(x)\},
$$

where $\psi_g$ and $\psi_f$ are inverse link functions similarly as in generalized linear models (McCullagh & Nelder [20]), $\xi(x)$ is a $p \times 1$ vector of known functions on $X$ such as $\xi(x) = (1, x^T)^T$, and $\alpha$ and $\gamma$ are $p \times 1$ vectors of unknown coefficients. Throughout, the same “basis” vector $\xi(x)$ is assumed to be used in both models (11) and (12); otherwise, the union of the basis functions from the two
models can be formed. While alternative strategies can also be studied, this condition seems inconsequential in high-dimensional, sparse settings, especially when the relative smoothness of \(g^*\) and \(f^*\) is unknown. In addition, models (11) and (12) may be misspecified. We say that model (11) is correctly specified if there exists a true value \(\gamma^*\) such that \(g^*(x) \equiv g(x; \alpha^*)\), or misspecified otherwise. Similarly, model (12) is correctly specified if there exists a true value \(\gamma^*\) such that \(f^*(x) \equiv f(x; \gamma^*)\), or misspecified otherwise. By definition, a true value \(\alpha^*\) or \(\gamma^*\) exists only if model (11) or (12) is correctly specified.

Given working models (11)–(12), the first-stage estimation involves constructing some estimators \(\hat{\alpha}\) and \(\hat{\gamma}\) and setting \(\hat{g} = g(x; \hat{\alpha})\) and \(\hat{f} = f(x; \hat{\gamma})\). Then an estimator for \(\theta^*\), denoted as \(\hat{\theta}(\hat{\alpha}, \hat{\gamma})\), is defined as a solution to (1) with \((g, f)\) replaced by \((\hat{g}, \hat{f})\), i.e.,

\[
0 = \hat{E}\{\tau(U; \theta, \hat{g}, \hat{f})\}.
\]

Conventionally, \((\hat{\alpha}, \hat{\gamma})\) are defined by maximum likelihood (or quasi-likelihood) including least squares in generalized linear models associated with (11)–(12). Our main subject is, however, calibrated estimation as an alternative approach. To facilitate discussion in Section 2.2, we describe some general asymptotic results about \((\hat{\alpha}, \hat{\gamma})\) and \(\hat{\theta}(\hat{\alpha}, \hat{\gamma})\), based on theory of estimation with possibly misspecified models (White [45]; Manski [19]), in the classical setting where \(\alpha\) and \(\gamma\) are fixed-dimensional as the sample size \(n\) grows. To focus on main issues, assume that \(\hat{\alpha}\) is consistent for \(\alpha^*\) if model (11) is correctly specified, and \(\hat{\gamma}\) is consistent for \(\gamma^*\) if model (12) is correctly specified.

With possible model misspecification, \(\hat{\alpha}\) can be shown to converge at rate \(O_p(n^{-1/2})\) to a target value \(\bar{\alpha}\), which coincides with the true value \(\alpha^*\) (i.e., \(\hat{\alpha}\) is consistent) if model (11) is correctly specified, but remains well-defined even though \(\alpha^*\) is undefined if model (11) is misspecified. Similarly, \(\hat{\gamma}\) can be shown to converge at rate \(O_p(n^{-1/2})\) to a target value \(\bar{\gamma}\), which coincides with the true value \(\gamma^*\) (i.e., \(\hat{\gamma}\) is consistent) if model (11) is correctly specified, but remains well-defined even though \(\gamma^*\) is undefined if model (12) is misspecified. As a result, unbiasedness properties (2)–(3) can be used to show that \(\hat{\theta}(\hat{\alpha}, \hat{\gamma})\) is doubly robust, i.e., remains consistent for \(\theta^*\) if either model (11) or (12) is correctly specified. Moreover, it can be shown that if model (11) is correctly specified with \(\hat{\alpha} = \alpha^*\) or model (12) is correctly specified with \(\hat{\gamma} = \gamma^*\), then \(\hat{\theta}(\hat{\alpha}, \hat{\gamma})\) admits the asymptotic expansion,

\[
\hat{\theta}(\hat{\alpha}, \hat{\gamma}) - \theta^* = -E^{-1}\left(\frac{\partial \tau}{\partial \theta}\right) \left\{\hat{E}(\tau) + E^T\left(\frac{\partial \tau}{\partial \alpha}\right)(\hat{\alpha} - \bar{\alpha}) + E^T\left(\frac{\partial \tau}{\partial \gamma}\right)(\hat{\gamma} - \bar{\gamma})\right\} + o_p(n^{-1/2}),
\]

(14)

where \(\tau = \tau(U; \theta, g(x; \alpha), f(x; \gamma))\), and \(\tau\) and its partial derivatives \(\partial \tau / \partial \theta, \partial \tau / \partial \alpha, \partial \tau / \partial \gamma\) are evaluated at \((\theta^*, \hat{\alpha}, \hat{\gamma})\). The preceding expansion (14) indicates how the asymptotic behavior of \(\hat{\theta}(\hat{\alpha}, \hat{\gamma})\) is affected by the estimators \((\hat{\alpha}, \hat{\gamma})\) through the second and third terms in the curly brackets. In fact, removing these two terms in (14) yields the asymptotic expansion of the infeasible estimator \(\hat{\theta}(\hat{\alpha}, \hat{\gamma})\), with \((\hat{\alpha}, \hat{\gamma})\) replaced by \((\bar{\alpha}, \bar{\gamma})\).

**Example 5.** We point out a somewhat under-appreciated result that the familiar least squares estimator for each individual coefficient in linear regression is doubly robust in the context of a partially linear model in Example 1. Let \(\psi_g(\cdot)\) be an identity function in model (11). For \(\tau\) in (5), the estimator \(\hat{\theta}(\hat{\alpha}, \hat{\gamma})\) as a solution to (13) is of closed form with \(\xi = \xi(X)\),

\[
\hat{\theta}(\hat{\alpha}, \hat{\gamma}) = \frac{\hat{E}\{(Y - \hat{\alpha}^T\xi)(Z - \psi_f(\hat{\gamma}^T\xi))\}}{\hat{E}\{Z(Z - \psi_f(\hat{\gamma}^T\xi))\}}.
\]
Doubly robust semiparametric inference

depending on some estimators \((\hat{\alpha}, \hat{\gamma})\). Suppose that \(\psi_f(\cdot)\) is also an identity function, i.e., a linear model is specified for \(E(Z|X)\). Let \((\hat{\theta}_0, \hat{\alpha})\) be the least-squares estimators of \((\theta, \alpha)\) in the linear regression of \(Y\) on \(Z\) and \(\xi(X)\), and \(\hat{\gamma}\) be that of \(\gamma\) in the linear regression of \(Z\) on \(\xi(X)\). Then \(\hat{\theta}(\hat{\alpha}, \hat{\gamma})\) is identical to \(\hat{\theta}_0\), the least squares estimator of \(\theta\):

\[
\hat{\theta}(\hat{\alpha}, \hat{\gamma}) - \hat{\theta}_0 = \frac{\hat{E}\{(Y - \hat{\theta}_0 Z - \hat{\alpha}^T \xi)(Z - \hat{\gamma}^T \xi)\}}{E\{Z(\hat{Z} - \hat{\gamma}^T \xi)\}} = 0,
\]

because \(\hat{E}\{(Y - \hat{\theta}_0 Z - \hat{\alpha}^T \xi) Z\} = 0\) and \(\hat{E}\{(Y - \hat{\theta}_0 Z - \hat{\alpha}^T \xi) \xi\} = 0\). Hence the least-squares estimator \(\hat{\theta}_0\) is doubly robust for \(\theta^*\) in the partially linear model (4), if either a linear model for \(g^*(x)\) or a linear model for \(f^*(x) = E(Z|X = x)\) is correctly specified. Furthermore, the sandwich variance estimator for \(\hat{\theta}_0\) (White [44]) can be written as \(n^{-1} \hat{V}\) with

\[
\hat{V} = \frac{\hat{E}\{(Y - \hat{\theta}_0 Z - \hat{\alpha}^T \xi)^2(Z - \hat{\gamma}^T \xi)^2\}}{E^2\{Z(\hat{Z} - \hat{\gamma}^T \xi)\}}.
\]

By Corollary 2 later, an asymptotic \((1 - c)\)-confidence interval for \(\theta^*\) is \(\hat{\theta}_0 \pm z_{c/2} \sqrt{\hat{V}/n}\) if either a linear model for \(E(Y|Z, X)\) or that for \(E(Z|X)\) is correctly specified. A high-dimensional version of this result is Corollary 4 later on debiased Lasso for least-squares estimation.

2.2. Calibrated estimation

We derive and discuss implications of basic mean-zero identities for a doubly robust estimating function \(\tau(U; \theta, g, f)\). In particular, we study calibrated estimation converting these identities into estimating equations in \((\alpha, \gamma)\). Here we assume the classical setting where asymptotic expansion (14) directly holds. See Section 3 for high-dimensional development.

For a function \(h(x)\) and a constant \(\delta > 0\), denote by \(h + L_2(\delta)\) the set \(\{h(x) + c(x) : E(c^2(x)) \leq \delta^2\}\). Denote by \(\partial \tau/\partial g\) and \(\partial \tau/\partial f\) the partial derivatives of \(\tau = \tau(U; \theta, g, f)\) with respect to \(G = g(x)\) and \(F = f(x)\) as free arguments. Whenever the dependency of \(\tau\) on \((\alpha, \gamma)\) is mentioned, \(\tau\) is parameterized as \(\tau(U; \theta, \alpha, \gamma) = \tau(U; \theta, g(x; \alpha), f(x; \gamma))\). For differentiation of \(\tau\) with respect to \((\alpha, \gamma)\), it is convenient to introduce linear predictors \((\eta_g, \eta_f)\) such that \(g(x) = \psi_g(\eta_g(x))\) and \(f(x) = \psi_f(\eta_f(x))\). Hence models (11) and (12) can be stated as \(\eta_g(x; \alpha) = \alpha^T \xi(x)\) and \(\eta_f(x; \gamma) = \gamma^T \xi(x)\). Denote by \(\partial \tau/\partial \eta_g\) and \(\partial \tau/\partial \eta_f\) the partial derivatives of \(\tau\) with respect to \(\eta_g(x)\) and \(\eta_f(x)\) as free arguments. By the chain rule, \(\partial \tau/\partial \alpha = (\partial \tau/\partial \eta_g) \xi = (\partial \tau/\partial g) \psi_g'(\alpha^T \xi)\) and \(\partial \tau/\partial \gamma = (\partial \tau/\partial \eta_f) \xi = (\partial \tau/\partial f) \psi_f'(\gamma^T \xi)\), where \(\psi_g'\) or \(\psi_f'\) denotes the derivative of \(\psi_g\) or \(\psi_f\).

**Proposition 1.** Under suitable regularity conditions, property (2) implies that

\[
0 = E \left\{ \frac{\partial \tau}{\partial f}(U; \theta^*, g^*, f)|X \right\},
\]

for any \(f\) such that \(f + L_2(\delta_1) \subset \mathcal{F}\) for some \(\delta_1 > 0\). Similarly, property (3) implies that

\[
0 = E \left\{ \frac{\partial \tau}{\partial g}(U; \theta^*, g, f^*)|X \right\},
\]

for any \(g\) such that \(g + L_2(\delta_2) \subset \mathcal{G}\) for some \(\delta_2 > 0\).
**Proof.** For $f$ such that $f + L_2(\delta_1) \subset \mathcal{F}$, (2) implies that for any $h \in L_2(1)$ and $a \in [-\delta_1, \delta_1]$, 
\[
0 = E\{f(U; \theta^*, g^*, f + ah)\}.
\]
Taking the derivative of the above with respect to $a$ with $f$ and $h$ fixed, and assuming the differentiation and expectation are interchangeable, we have 
\[
0 = E\left\{\frac{\partial \tau}{\partial f}(U; \theta^*, g^*, f)h(X)\right\}.
\]
Hence (15) follows because $h \in L_2(1)$ is arbitrary. Similarly, (16) can be proved.

Similar reasoning as above can be applied to the derivatives of $\tau$ with respect to $(\alpha, \gamma)$, given models (11)–(12). Differentiation of (2) or (3) with respect to $\gamma$ or $\alpha$ respectively and interchanging differentiation and expectation shows that for any $(\alpha, \gamma)$, 
\[
0 = E\left\{\frac{\partial \tau}{\partial \gamma}(U; \theta^*, g^*, f(x; \gamma))\right\} = E\left\{\xi(X)\frac{\partial \tau}{\partial \eta_f}(U; \theta^*, g^*, f(x; \gamma))\right\}, \quad (17)
\]
\[
0 = E\left\{\frac{\partial \tau}{\partial \alpha}(U; \theta^*, g(x; \alpha), f^*)\right\} = E\left\{\xi(X)\frac{\partial \tau}{\partial \eta_g}(U; \theta^*, g(x; \alpha), f^*)\right\}. \quad (18)
\]
Equivalently, (17)–(18) can also be deduced from the more general identities (15)–(16), which involve conditional expectations given $X$. Model (11) with $g(x; \alpha)$ may be misspecified in (18), and model (12) with $f(x; \gamma)$ may be misspecified in (17).

We stress that identities (15)–(16) and (17)–(18) are derived from double-robustness properties (2)–(3) in a general manner. To some extent, identities (17)–(18) are intriguingly reminiscent of the score identity in likelihood inference with a parametric model: the expectation of the gradient of the log-likelihood, evaluated at the true parameter value, is zero. However, $\tau$ is an estimating function in $\theta$, not a log-likelihood function in $\alpha$ or $\gamma$.

There are various implications of basic identities (17)–(18). First, these identities show that $E(\partial \tau/\partial \gamma)$ or $E(\partial \tau/\partial \alpha)$ reduces to $0$ in asymptotic expansion (14) for $\hat{\theta}(\hat{\alpha}, \hat{\gamma})$, depending on whether model (11) or (12) is correctly specified. If model (11) with $g(x; \alpha)$ is correctly specified and $\hat{\alpha}$ is consistent, then, by (17), asymptotic expansion (14) reduces to 
\[
\hat{\theta}(\hat{\alpha}, \hat{\gamma}) - \theta^* = -E^{-1}\left(\frac{\partial \tau}{\partial \theta}\right)\left\{\hat{E}(\tau) + E^T\left(\frac{\partial \tau}{\partial \alpha}\right)(\hat{\alpha} - \alpha^*)\right\} + o_p(n^{-1/2}), \quad (19)
\]
where $\tau$ and its partial derivatives are evaluated at $(\theta, \alpha, \gamma) = (\theta^*, \alpha^*, \hat{\gamma})$. As the term associated with $\hat{\gamma} - \gamma$ vanishes in (19), the asymptotic behavior of $\hat{\theta}(\hat{\alpha}, \hat{\gamma})$ does not depend on the definition of $\hat{\gamma}$, as long as model (11) is correctly specified and $\hat{\alpha}$ is consistent. Similarly, if model (12) with $f(x; \gamma)$ is correctly specified and $\hat{\gamma}$ is consistent, then, by (18), the asymptotic behavior of $\hat{\theta}(\hat{\alpha}, \hat{\gamma})$ does not depend on the definition of $\hat{\alpha}$:
\[
\hat{\theta}(\hat{\alpha}, \hat{\gamma}) - \theta^* = -E^{-1}\left(\frac{\partial \tau}{\partial \theta}\right)\left\{\hat{E}(\tau) + E^T\left(\frac{\partial \tau}{\partial \gamma}\right)(\hat{\gamma} - \gamma^*)\right\} + o_p(n^{-1/2}), \quad (20)
\]
where $\tau$ and its partial derivatives are evaluated at $(\theta, \alpha, \gamma) = (\theta^*, \hat{\alpha}, \gamma^*)$. Combining the preceding arguments leads to Corollary 1: if both models (11) and (12) are correctly specified, then the asymptotic behavior of $\hat{\theta}(\hat{\alpha}, \hat{\gamma})$ remains the same for all consistent estimators $(\hat{\alpha}, \hat{\gamma})$. This result, related to local efficiency in specific examples (e.g., Robins et al. [28]; Tan [33]), is obtained here as a general consequence of double robustness of $\tau$. 
Corollary 1. If both models (11) and (12) are correctly specified and \((\hat{\alpha}, \hat{\gamma})\) are consistent, then as \(p\) is fixed and \(n \to \infty\), asymptotic expansion (14) reduces to

\[
\hat{\theta}(\hat{\alpha}, \hat{\gamma}) - \theta^* = -E^{-1}\left(\frac{\partial \tau}{\partial \theta}\right) \bar{E}(\tau) + o_p(n^{-1/2}),
\]

where \(\tau\) and \(\partial \tau / \partial \theta\) are evaluated at \((\theta, \alpha, \gamma) = (\theta^*, \alpha^*, \gamma^*)\).

Second, methodologically, identities (17)–(18) can also be exploited to construct specific estimators \((\hat{\alpha}, \hat{\gamma})\), for which the simple expansion (21) is valid with the true values \((\alpha^*, \gamma^*)\) replaced by target values \((\hat{\alpha}, \hat{\gamma})\) if either model (11) or (12), but not necessarily both, is correctly specified. Suppose that estimators \((\hat{\alpha}_{\text{CAL}}, \hat{\gamma}_{\text{CAL}})\) are defined such that they converge in probability to target values \((\bar{\alpha}_{\text{CAL}}, \bar{\gamma}_{\text{CAL}})\) satisfying the simultaneous equations

\[
0 = E\left\{\frac{\partial \tau}{\partial \gamma}(U; \theta^*, \alpha, \gamma)\right\} = E\left\{\xi \frac{\partial \tau}{\partial \eta_f}(U; \theta^*, \alpha, \gamma)\right\}, \tag{22}
\]

\[
0 = E\left\{\frac{\partial \tau}{\partial \alpha}(U; \theta^*, \alpha, \gamma)\right\} = E\left\{\xi \frac{\partial \tau}{\partial \eta_g}(U; \theta^*, \alpha, \gamma)\right\}, \tag{23}
\]

that is, the coefficients of \(\hat{\gamma} - \hat{\gamma}\) and \(\hat{\alpha} - \hat{\alpha}\) are set to 0 in expansion (14) for \(\hat{\theta}(\hat{\alpha}, \hat{\gamma})\). Assume that there exists at most one value \(\alpha\) satisfying (22) for each fixed \(\gamma\), and at most one value \(\gamma\) satisfying (23) for each fixed \(\alpha\). From our discussion below, this implies that \((\bar{\alpha}_{\text{CAL}}, \bar{\gamma}_{\text{CAL}})\) is a unique solution to (22)–(23) if model (11) or (12) is correctly specified.

If model (11) with \(g(x; \alpha)\) is correctly specified, then by (17), \(\hat{\alpha}_{\text{CAL}}\) coincides with \(\alpha^*\) as a solution to (22) for fixed \(\gamma = \hat{\gamma}_{\text{CAL}}\), i.e., \(\hat{\alpha}_{\text{CAL}}\) is consistent. In this case, (22) can be seen as an unbiased population estimating equation for \(\alpha^*\) with fixed \(\gamma\). Similarly, if model (12) with \(f(x; \gamma)\) is correctly specified, then by comparison of (18) and (23), \(\hat{\gamma}_{\text{CAL}}\) coincides with \(\gamma^*\), i.e., \(\hat{\gamma}_{\text{CAL}}\) is consistent. In this case, (23) can be seen as an unbiased population estimating equation for \(\gamma^*\) with fixed \(\alpha\). (An interesting asymmetry is that differentiation of \(\tau\) with respect to \(\gamma\) leads to an estimating equation in \(\alpha\), whereas that of \(\tau\) with respect to \(\alpha\) leads to an estimating equation in \(\gamma\).) Combining the two cases and applying asymptotic expansion (14) leads to the Corollary 2, where, due to (22)–(23) again, the two terms associated with \(\hat{\gamma} - \hat{\gamma}\) and \(\hat{\alpha} - \hat{\alpha}\) are dropped from the expansion (14). Alternatively, to help understanding, asymptotic expansion (24) for \(\hat{\theta}(\hat{\alpha}_{\text{CAL}}, \hat{\gamma}_{\text{CAL}})\) can also be obtained from expansion (19) with \(E(\partial \tau / \partial \alpha) = 0\) due to (23) if model (11) is correctly specified, or from expansion (20) with \(E(\partial \tau / \partial \gamma) = 0\) due to (22) if model (12) is correctly specified.

Corollary 2. If model (11) or (12) is correctly specified, then \(\hat{\alpha}_{\text{CAL}}\) or \(\hat{\gamma}_{\text{CAL}}\) is consistent for \(\alpha^*\) or \(\gamma^*\) respectively. In either case, the estimator \(\hat{\theta}(\hat{\alpha}_{\text{CAL}}, \hat{\gamma}_{\text{CAL}})\) satisfies

\[
\hat{\theta}(\hat{\alpha}_{\text{CAL}}, \hat{\gamma}_{\text{CAL}}) - \theta^* = -E^{-1}\left(\frac{\partial \tau}{\partial \theta}\right) \bar{E}(\tau)_{\theta, \alpha, \gamma = (\theta^*, \hat{\alpha}_{\text{CAL}}, \hat{\gamma}_{\text{CAL}})} + o_p(n^{-1/2}), \tag{24}
\]

provided that expansion (14) holds for \((\hat{\alpha}, \hat{\gamma}) = (\hat{\alpha}_{\text{CAL}}, \hat{\gamma}_{\text{CAL}})\) as \(p\) is fixed and \(n \to \infty\).

We refer to equations (22)–(23) as population calibration equations and \((\hat{\alpha}_{\text{CAL}}, \hat{\gamma}_{\text{CAL}})\) as calibrated estimators for two reasons, following Tan [37]. For the missing-data problem in Example 4, related to estimation of average treatment effects, this method leads to calibrated
estimation for fitting propensity score models \( f(x;\gamma) \), which can be traced to the literature on survey calibration (Folsom [14]). See Example 9 below. More generally, as indicated by Corollaries 1–2, using estimating equations (22)–(23) can be seen as carefully choosing (or calibrating) estimators \((\hat{\alpha}, \hat{\gamma})\) for the nuisance parameters \((\alpha, \gamma)\), such that the resulting estimator \(\hat{\theta}(\hat{\alpha}, \hat{\gamma})\) behaves as if both models (11) and (12) were correctly specified, while it is only assumed that either model (11) or (12) is correctly specified.

A benefit of achieving asymptotic expansion (24) is to allow simple variance estimation for \(\hat{\theta}(\hat{\alpha}_{\text{CAL}}, \hat{\gamma}_{\text{CAL}})\), without the need to account for the variations of \((\hat{\alpha}_{\text{CAL}}, \hat{\gamma}_{\text{CAL}})\). This benefit is mainly computationally in the setting of low-dimensional \((\alpha, \gamma)\), where variance estimation can in general be performed for \(\hat{\theta}(\hat{\alpha}, \hat{\gamma})\) by using asymptotic expansion (14) and usual influence functions for \((\hat{\alpha}, \hat{\gamma})\), allowing for model misspecification (White [45]; Manski [19]). However, the influence-function based approach is not applicable in the high-dimensional setting where regularized estimation is involved. In Section 3, we develop regularized calibration estimation to achieve a simple expansion similar to (24) for the resulting estimator of \(\theta^*\), so that valid variance estimation and confidence intervals can be obtained.

Remark 1. It is important to distinguish the two expansions (21) and (24), although they appear similar to each other. The expansion (21) holds for any consistent estimators \((\hat{\alpha}, \hat{\gamma})\) provided that both models (11) and (12) are correctly specified. The two terms \(E(\partial \tau / \partial \alpha)\) and \(E(\partial \tau / \partial \gamma)\) in (14) reduce to 0 by the assumption of both models (11) and (12) being correctly specified, while appealing to the two identities (17)–(18) simultaneously. In contrast, the expansion (24) is valid for estimators \((\hat{\alpha}_{\text{CAL}}, \hat{\gamma}_{\text{CAL}})\) constructed such that (22)–(23) are satisfied, if either model (11) or (12), but not necessarily both, is correctly specified. The two terms \(E(\partial \tau / \partial \alpha)\) and \(E(\partial \tau / \partial \gamma)\) in (14) reduce to 0 by the construction of population estimating equations (22)–(23). Identity (17) is involved to show consistency of \(\hat{\alpha}_{\text{CAL}}\) if model (11) is correct or, separately, identity (18) is involved to show consistency of \(\hat{\gamma}_{\text{CAL}}\) if model (12) is correct, whereas consistency of \((\hat{\alpha}, \hat{\gamma})\) is presumed in Corollary 1.

Our preceding discussion leaves open the question how calibrated estimators \((\hat{\alpha}_{\text{CAL}}, \hat{\gamma}_{\text{CAL}})\) can be defined such that (22)–(23) are satisfied. A direct approach would be to take \((\hat{\alpha}_{\text{CAL}}, \hat{\gamma}_{\text{CAL}})\) as a solution to the sample version of calibration equations (22)–(23), where the expectation \(E(\cdot)\) is replaced by the sample average \(\bar{E}(\cdot)\). However, there are various complications for this approach even in the classical setting with low-dimensional \((\alpha, \gamma)\). First, equations (22)–(23) and the sample version may depend on \(\theta^*\) to be estimated. A preliminary doubly robust estimator can be substituted for \(\theta^*\). But the resulting sample version of (22)–(23) remains a system of nonlinear equations in \((\alpha, \gamma)\). Numerical solution of such equations with finite data may suffer the issue of no solution or multiple solutions (Small et al. [31]). Theoretical analysis of estimators from nonlinear estimating equations may require cumbersome regularity conditions which would be avoided when using conventional estimators of \((\alpha, \gamma)\). These issues can be illustrated with the following examples.

Example 6. For Example 1 with a partially linear model, let \(\psi_f(\cdot)\) be an identity function. The calibration equations (22)–(23) based on \(\tau\) in (5) are

\[
0 = E \left( \frac{\partial \tau}{\partial \gamma} \right) = -E \left\{ (Y - \theta^* Z - \alpha^T \xi) \psi_f(\gamma^T \xi) \xi \right\}, \tag{25}
\]

\[
0 = E \left( \frac{\partial \tau}{\partial \alpha} \right) = -E \left\{ (Z - \psi_f(\gamma^T \xi)) \xi \right\}, \tag{26}
\]
where $\xi = \xi(X)$ and $\tau$ is evaluated at $\theta = \theta^*$. Because (26) does not depend on $\alpha$, the sample version of simultaneous equations (25)–(26) can be solved sequentially: the sample version of (26) can be first solved, and then that of (25) be solved, provided that $\theta^*$ is replaced by a preliminary doubly robust estimator. For comparison, the influence function associated with (5) is $\tau_{\theta}(U;\theta,\alpha,\gamma) = E^{-1}\{Z(Z - \psi_f(\gamma^T \xi))\} \tau(U;\theta,\alpha,\gamma)$. The population version of estimating equations in Vermeulen & Vansteelandt [43], Section 5.2, can be stated as

$$0 = E \left[ -E\{Z(Z - \psi_f(\gamma^T \xi))(Y - \alpha^T \xi)\psi_f'(\gamma^T \xi)\xi + E\{Z\psi_f'(\gamma^T f)\}(Y - \alpha^T \xi)(Z - \psi_f(\gamma^T \xi))\xi \right]$$

for $\alpha$ and (26) for $\gamma$, which are obtained by setting $0 = E(\partial \tau_{\theta}/\partial \gamma)$ and $0 = E(\partial \tau_{\theta}/\partial \alpha)$. Compared with (25)–(26), these equations are more complex but, interestingly, free of $\theta^*$.

**Example 7.** For Example 2 with a partially log-linear model, let $\psi_y(\cdot)$ be an identity function. The calibration equations (22)–(23) based on $\tau$ in (7) are

$$0 = E \left( \frac{\partial \tau}{\partial \gamma} \right) = -E \left\{ (Ye^{-\theta^*Z} - e^{\alpha^T \xi})\psi_f'(\gamma^T \xi) \xi \right\},$$

$$0 = E \left( \frac{\partial \tau}{\partial \alpha} \right) = -E \left\{ (Z - \psi_f(\gamma^T \xi))e^{\alpha^T \xi} \xi \right\},$$

(27)

(28)

where $\xi = \xi(X)$ and $\tau$ is evaluated at $\theta = \theta^*$. Unlike (25)–(26) in Example 6, the sample version of (27)–(28) cannot be solved sequentially even after $\theta^*$ is appropriately estimated. Therefore, algorithms for solving nonlinear equations need to be used. The calibration equations (27)–(28) are simpler than estimating equations in Vermeulen & Vansteelandt [43], Section 5.2, which are defined as $0 = E(\partial \tau_{\theta}/\partial \gamma)$ and $0 = E(\partial \tau_{\theta}/\partial \alpha)$, where $\tau_{\theta}$ is the influence function, $\tau_{\theta}(U;\theta,\alpha,\gamma) = E^{-1}(\partial \tau/\partial \theta)\tau(U;\theta,\alpha,\gamma)$, evaluated at $\theta = \theta^*$.

**Example 8.** For Example 3 with a logistic partially linear model, let $\psi_y(\cdot)$ be an identity function. The calibration equations (22)–(23) based on $\tau$ in (9) are

$$0 = E \left( \frac{\partial \tau}{\partial \gamma} \right) = -E \left\{ e^{-\theta^*ZY}(Y - \expit(\alpha^T \xi))\psi_f'(\gamma^T \xi) \xi \right\},$$

$$0 = E \left( \frac{\partial \tau}{\partial \alpha} \right) = -E \left\{ e^{-\theta^*ZY}\expit_2(\alpha^T \xi))(Z - \psi_f(\gamma^T \xi)) \xi \right\},$$

(29)

(30)

where $\expit_2(c) = \expit(c)(1 - \expit(c))$ and $\tau$ is evaluated at $\theta = \theta^*$. Similarly as (27)–(28), the sample version of (29)–(28) cannot be solved sequentially, due to dependency on both $\alpha$ and $\gamma$, even after $\theta^*$ is appropriately estimated.

**Example 9.** For the missing-data problem in Example 4, the calibration equations (22)–(23) based on $\tau$ in (10) are

$$0 = E \left( \frac{\partial \tau}{\partial \gamma} \right) = -E \left\{ \frac{\psi_f'(\gamma^T \xi)}{\psi_f'(\gamma^T \xi)}Z(Y - \psi_g(\alpha^T \xi)) \xi \right\},$$

$$0 = E \left( \frac{\partial \tau}{\partial \alpha} \right) = -E \left\{ \frac{Z}{\psi_f'(\gamma^T \xi)} - 1 \right\} \psi_g'(\alpha^T \xi) \xi \right\},$$

(31)

(32)
where $\xi = \xi(X)$ and $\tau$ is evaluated at $\theta = \theta^*$. In the case where $\psi_g(\cdot)$ is an identity function, i.e., a linear model (11) is specified for $E(Y|Z = 1, X)$, the sample version of (31)–(32) can be solved sequentially, because (32) does not depend on $\alpha$. But such sequential solution is infeasible with a nonlinear function $\psi_g(\cdot)$, because equations (31)–(32) are intrinsically coupled, each depending on both $\alpha$ and $\gamma$ (Tan [37], Section 3.5).

3. Regularized calibrated estimation

We develop regularized calibrated estimation for $(\alpha, \gamma)$, such that the resulting estimator of $\theta^*$ achieves an asymptotic expansion similar to (24), hence allowing valid confidence intervals, under suitable conditions in high-dimensional settings. The estimators of $(\alpha, \gamma)$ are derived from a numerically tractable two-step algorithm. Moreover, high-dimensional analysis is provided to demonstrate the desired asymptotic expansion and consistent variance estimation, which lead to valid Wald confidence intervals.

Conceptually, regularized calibrated estimation involves constructing regularized estimators of $(\alpha, \gamma)$, which converge in probability to the target values $(\bar{\alpha}_{CAL}, \bar{\gamma}_{CAL})$ satisfying population calibration equations (22)–(23). As discussed in Section 2.2 in low-dimensional settings, there may be numerical and theoretical complications with directly using the sample version of (22)–(23) as estimating equations. With high-dimensional data, estimating equations can be regularized by generalizing the Dantzig selector (Candes & Tao [7]), which seeks to minimize $\|\alpha\|_1 + \|\gamma\|_1$ subject to

$$
\left\| \hat{E} \left\{ \frac{\partial \tau}{\partial \gamma}(U; \hat{\theta}_1, g(x; \alpha), f(x; \gamma)) \right\} \right\|_{\infty} \leq \lambda,
\left\| \hat{E} \left\{ \frac{\partial \tau}{\partial \alpha}(U; \hat{\theta}_1, g(x; \alpha), f(x; \gamma)) \right\} \right\|_{\infty} \leq \lambda,
$$

where $\hat{\theta}_1$ is a preliminary doubly robust estimator, $\lambda$ is a tuning parameter, and $\|\cdot\|_1$ or $\|\cdot\|_{\infty}$ denotes $L_1$ or $L_\infty$ norm. While theoretical analysis of generalized Dantzig selectors can be performed, this approach is not pursued here mainly because the required optimization problem seems numerically difficult to solve with complex nonlinear estimating functions. The generalized Dantzig-selector algorithm in Radchenko & James [26] can potentially be modified for the above problem, but its effectiveness seems uncertain. Further investigation of the Dantzig-selector approach can be of interest in future work.

3.1. Two-step algorithm

We propose a two-step algorithm, shown as Algorithm 1, for regularized calibrated estimation. The algorithm is facilitated by exploiting the following convexity assumption, which is satisfied in various settings including Examples 1–4 as shown in Section 4. In principle, our approach can also be applied without the convexity assumption, provided that a solution to equation (22) or (23) is unique in $\alpha$ or $\gamma$, while $\gamma$ or $\alpha$ is fixed respectively. Such an assumption is used earlier in the discussion leading to Corollary 2.

Assumption 1. There exist two loss functions $\ell_1(U; \theta, \alpha, \gamma)$ and $\ell_2(U; \theta, \alpha, \gamma)$ such that $E\{\ell_1(U; \theta, \alpha, \gamma)\}$ is strictly convex in $\alpha$, $E\{\ell_2(U; \theta, \alpha, \gamma)\}$ is strictly convex in $\gamma$, and

$$
\frac{\partial \ell_1}{\partial \alpha} = \frac{\partial \tau}{\partial \gamma}, \quad \frac{\partial \ell_2}{\partial \gamma} = \frac{\partial \tau}{\partial \alpha},
$$

where $\tau$ is parameterized as $\tau(U; \theta, \alpha, \gamma) = \tau(U; \theta, g(x; \alpha), f(x; \gamma))$. 
Algorithm 1 Two-step algorithm

1: procedure INITIAL ESTIMATION
2: \hspace{1em} Compute \((\hat{\alpha}_1, \hat{\gamma}_1)\) as model-based estimators of \((\alpha, \gamma)\);
3: \hspace{1em} Compute \(\hat{\theta}_1 = \theta(\hat{\alpha}_1, \hat{\gamma}_1)\) as a solution to \(E\{\tau(U; \theta, \hat{\alpha}_1, \hat{\gamma}_1)\} = 0\).
4: end procedure

5: procedure CALIBRATED ESTIMATION
6: \hspace{1em} Compute \(\hat{\gamma}_2 = \arg\min_{\gamma} E\{\ell_2(U; \hat{\theta}_1, \alpha, \gamma)\} + \lambda_1 \|\gamma\|_1\), also denoted as \(\hat{\gamma}_{RCAL}\);
7: \hspace{1em} Compute \(\hat{\alpha}_2 = \arg\min_{\alpha} E\{\ell_1(U; \hat{\theta}_1, \alpha, \hat{\gamma}_2)\} + \lambda_2 \|\alpha\|_2\), also denoted as \(\hat{\alpha}_{RCAL}\);
8: \hspace{1em} Compute \(\hat{\theta}_2 = \theta(\hat{\alpha}_2, \hat{\gamma}_2)\) as a solution to \(E\{\tau(U; \theta, \hat{\alpha}_2, \hat{\gamma}_2)\} = 0\), also denoted as \(\hat{\theta}_{RCAL}\).
9: end procedure

From Assumption 1, various equations in Section 2.2 can be restated in terms of minimization of convex loss functions. The basic identities (17)–(18) can be translated to minimization properties. If model (11) with \(g(x; \alpha)\) is correctly specified, then (17) amounts to \(E\{\partial/\partial\alpha\ell_1(U; \theta^*, \alpha, \gamma)\}_{\alpha=\alpha^*} = 0\) and hence for fixed \(\gamma\), the expected loss \(E\{\ell_1(U; \theta^*, \alpha, \gamma)\}\), convex in \(\alpha\), attains a minimum at \(\alpha^*\) with zero gradient under interchangeability of the differentiation and expectation. Similarly, if model (12) with \(f(x; \gamma)\) is correctly specified, then (18) amounts to \(E\{\partial/\partial\gamma\ell_2(U; \theta^*, \alpha, \gamma)\}_{\gamma=\gamma^*} = 0\) and hence for fixed \(\alpha\), the expected loss \(E\{\ell_2(U; \theta^*, \alpha, \gamma)\}\), convex in \(\gamma\), is minimized at \(\gamma^*\).

The population calibration equations (22)–(23) can be expressed in the form of alternating minimization: \(E\{\ell_1(U; \theta^*, \alpha, \gamma)\}\) is minimized at \(\alpha = \hat{\alpha}_{CAL}\) for fixed \(\gamma = \hat{\gamma}_{CAL}\), and \(E\{\ell_2(U; \theta^*, \alpha, \gamma)\}\) is minimized at \(\gamma = \hat{\gamma}_{CAL}\) for fixed \(\alpha = \hat{\alpha}_{CAL}\). This reasoning would suggest the following iterative algorithm for computing \((\hat{\alpha}_{CAL}, \hat{\gamma}_{CAL})\) at a population level.

Population calibration algorithm.

- Determine initial target values \((\bar{\alpha}_1, \bar{\gamma}_1)\);
- For \(t = 2, 3, \ldots\), determine \(\bar{\gamma}_t\) as a solution to \(E\{(\partial/\partial\gamma)\tau(U; \theta^*, \bar{\alpha}_{t-1}, \gamma)\} = 0\) or a minimizer of \(E\{\ell_2(U; \theta^*, \bar{\alpha}_{t-1}, \gamma)\}\) in \(\gamma\), and then determine \(\bar{\alpha}_t\) as a solution to \(E\{(\partial/\partial\alpha)\tau(U; \theta^*, \bar{\alpha}_t, \gamma)\} = 0\) or a minimizer of \(E\{\ell_1(U; \theta^*, \alpha, \bar{\gamma}_t)\}\) in \(\alpha\).

The limit \((\hat{\alpha}_\infty, \hat{\gamma}_\infty) = \lim_{t \to \infty} (\bar{\alpha}_t, \bar{\gamma}_t)\), if exists, can be shown to satisfy (22)–(23). However, remarkably, we show in Proposition 2 that if the initial target values \((\bar{\alpha}_1, \bar{\gamma}_1)\) are determined from model-based estimators of \((\alpha, \gamma)\) which are consistent in the case of either model (11) or (12) being correctly specified, then the iterative process can be terminated by the second step (i.e., by \(t = 2\)), as far as doubly robust estimation is concerned. It should also be mentioned that if both models (11) and (12) are misspecified, then the second-step target values \((\bar{\alpha}_2, \bar{\gamma}_2)\) may in general not satisfy calibration equations (22)–(23).

**Proposition 2.** If model (11) is correctly specified and \(\bar{\alpha}_1 = \alpha^*\) but \(\bar{\gamma}_1\) is arbitrary, or if model (12) is correctly specified and \(\bar{\gamma}_1 = \gamma^*\) but \(\bar{\alpha}_1\) is arbitrary, then \(\bar{\alpha}_2 = \alpha^*\) or \(\bar{\gamma}_2 = \gamma^*\) respectively, and \((\bar{\alpha}_2, \bar{\gamma}_2)\) jointly satisfy calibration equations (22)–(23).

**Proof.** By definition, \((\bar{\alpha}_2, \bar{\gamma}_2)\) satisfy the equations

\[
E\left\{\frac{\partial\tau}{\partial\alpha}(U; \theta^*, \bar{\alpha}_1, \bar{\gamma}_2)\right\} = 0, \tag{34}
\]

\[
E\left\{\frac{\partial\tau}{\partial\gamma}(U; \theta^*, \bar{\alpha}_2, \bar{\gamma}_2)\right\} = 0. \tag{35}
\]
If model (11) is correctly specified and $\hat{\alpha}_1 = \alpha^*$, then by comparison of (17) and (35), $\hat{\alpha}_2 = \alpha^*$, and hence (35) and (34) yield (22) and (23) respectively for $(\hat{\alpha}_2, \hat{\gamma}_2)$. If model (12) is correctly specified and $\hat{\gamma}_1 = \gamma^*$, then by comparison of (18) and (34), $\hat{\gamma}_2 = \gamma^*$, and by (18),

$$E \left\{ \frac{\partial \tau}{\partial \alpha}(U; \theta^*, \hat{\alpha}_2, \gamma^*) \right\} = 0.$$  \hfill (36)

In this case, (35) and (36) lead to (22) and (23) respectively for $(\hat{\alpha}_2, \hat{\gamma}_2)$. \hfill $\Box$

Algorithm 1 is a sample version of the population calibration algorithm with two steps, using regularized estimation with Lasso penalties to deal with high-dimensional data. The initial estimators $(\hat{\alpha}_1, \hat{\gamma}_1)$ can be Lasso-regularized maximum likelihood (or quasi-likelihood) estimators in generalized linear models associated with (11)–(12). The two-step estimators, $(\hat{\alpha}_{RCAL}, \hat{\gamma}_{RCAL}) = (\hat{\alpha}_2, \hat{\gamma}_2)$, serves as an adjustment to the usual estimators $(\hat{\alpha}_1, \hat{\gamma}_1)$, such that calibration equations (22)–(23) are satisfied if either model (11) or (12) is correct.

**Remark 2.** There are interesting issues which warrant further study. While calibration equations (22)–(23) are symmetric in $(\alpha, \gamma)$, Algorithm 1 involves choosing one parameter to be estimated and then the other in the second step. Nevertheless, the sparsity conditions required for our main result, stated as Proposition 3, appear to be symmetric in the target values $(\hat{\alpha}_1, \hat{\alpha}_2)$ and $(\hat{\gamma}_1, \hat{\gamma}_2)$, even though the definitions of $(\hat{\alpha}_2, \hat{\gamma}_2)$ depend on the updating order in Algorithm 1. In our applications (Section 4), our choices are influenced by previous work (Tan [37]) on estimation of the average treatment effect related to Example 4, where in the special case of an identity link $\psi_0(\cdot)$, calibration equation (32) for propensity score parameter $\gamma$ is independent of $\alpha$ and $\theta^*$ and hence updating $\gamma$ first becomes numerically simpler than otherwise. In more complex settings, we then choose to first update estimation of parameters similar to the propensity score in Example 4. Further research is desired to fully understand implications of different choices. In addition, although Algorithm 1 is terminated after two steps, it is interesting to study possible benefits or drawbacks from additional iterations.

### 3.2. Theoretical analysis

We provide high-dimensional analysis of the two-step estimators $(\hat{\alpha}_{RCAL}, \hat{\gamma}_{RCAL}) = (\hat{\alpha}_2, \hat{\gamma}_2)$ and the resulting estimator $\hat{\theta}_{RCAL} = \hat{\theta}(\hat{\alpha}_2, \hat{\gamma}_2)$ in Algorithm 1. Throughout this section, we assume that either model (11) or (12), but not necessarily both, is correctly specified.

Our main result, summarized as Proposition 3, can be deduced from Theorems 1–3 later. For initial estimators $(\hat{\alpha}_1, \hat{\gamma}_1)$ defined as Lasso-regularized maximum likelihood (or quasi-likelihood) estimators, the rates of convergence in Assumption 2(iv) later are satisfied under suitable conditions with $M_0 = O(1)(|S_{\hat{\alpha}_1}| + |S_{\hat{\gamma}_1}|)$, where $|S_{\hat{\alpha}_1}|$ or $|S_{\hat{\gamma}_1}|$ denotes the number of nonzero coefficients of the target value $\hat{\alpha}_1$ or $\hat{\gamma}_1$ respectively (Bühlmann & Van de Geer [5]; Negahban et al. [21]). For the two-step estimators $(\hat{\alpha}_2, \hat{\gamma}_2)$, denote by $|S_{\hat{\alpha}_2}|$ or $|S_{\hat{\gamma}_2}|$ denotes the number of nonzero coefficients of the target value $\hat{\alpha}_2$ or $\hat{\gamma}_2$ respectively. Suppose that the Lasso tuning parameters are specified as $\lambda_1 = A_1^1 r_0$ and $\lambda_2 = A_2^2 r_0$ for sufficiently large constants $A_1^1$ and $A_2^2$, where $r_0 = \log(ep)/n^{1/2}$ and $e = \exp(1)$, i.e., Euler’s constant.

**Proposition 3.** Suppose that Assumptions 1–7 hold, and $(M_0 + |S_{\hat{\alpha}_2}| + |S_{\hat{\gamma}_2}|) r_0^2 = o(n^{-1/2})$, i.e., $(M_0 + |S_{\hat{\alpha}_2}| + |S_{\hat{\gamma}_2}|) \log(ep) = o(n^{1/2})$. If model (11) with $g(x; \alpha)$ is correctly specified or
model (12) is correctly specified with \( f(x; \gamma) \), then \( \hat{\theta}_{\text{RCAL}} = \hat{\theta}(\hat{\alpha}_2, \hat{\gamma}_2) \) satisfies

\[
\hat{\theta}_{\text{RCAL}} - \theta^* = -E^{-1}\left( \frac{\partial \tau}{\partial \theta} \right) \left( \overline{E}(\tau) \right) \bigg|_{(\theta, \alpha, \gamma) = (\theta^*, \hat{\alpha}_2, \hat{\gamma}_2)} + o_p(n^{-1/2}). \tag{37}
\]

Furthermore, the following results hold in either case:

(i) \( \sqrt{n}(\hat{\theta}_{\text{RCAL}} - \theta^*) \xrightarrow{D} \mathcal{N}(0, V) \), where \( V = \operatorname{var}(\tau)/E^2(\partial \tau/\partial \theta) \bigg|_{(\theta, \alpha, \gamma) = (\theta^*, \hat{\alpha}_2, \hat{\gamma}_2)} \);

(ii) A consistent estimator \( \hat{V} \) of \( V \) is

\[
\hat{V} = \overline{E}(\tau^2)/\overline{E}^2(\partial \tau/\partial \theta) \bigg|_{(\theta, \alpha, \gamma) = (\hat{\theta}_{\text{RCAL}}, \hat{\alpha}_2, \hat{\gamma}_2)}
\]

(iii) An asymptotic \((1 - c)\) confidence interval for \( \theta^* \) is \( \hat{\theta}_{\text{RCAL}} \pm z_{c/2} \sqrt{V/n} \), where \( z_{c/2} \) is the \((1 - c/2)\) quantile of \( \mathcal{N}(0, 1) \).

Hence a doubly robust confidence interval for \( \theta^* \) is obtained.

**Remark 3.** The preceding result is obtained, depending on the exact sparsity of the target (or limit) values, that is, the numbers of nonzero coefficients in \((\hat{\alpha}_1, \hat{\gamma}_1)\) and \((\hat{\alpha}_2, \hat{\gamma}_2)\). As discussed in Dukes & Vansteelandt [11], such sparsity conditions with possibly misspecified models can be satisfied, whether or not the data-generating models are sparse. Moreover, our theoretical analysis can be extended to approximately sparse settings, where the sparsity of \((\hat{\alpha}_j, \hat{\gamma}_j)\) is measured by \( R_q(\hat{\alpha}_j) \) and \( R_q(\hat{\gamma}_j) \) for some \( q \in [0, 1], \) and \( R_q(b) = \sum_{j=1}^{p} |b_j|^q \) for a vector \( b = (b_1, \ldots, b_p)^T \) (e.g., Negahban et al. [21]). We expect that the conclusions in Proposition 3 remain valid under the more general sparsity condition, \( \sum_{j=1}^{2}(R_q(\hat{\alpha}_j) + R_q(\hat{\gamma}_j))(\log(ep)/n)^{1-q/2} = o(n^{-1/2}) \), among others. See Smucler et al. [32] for related results.

In the remainder of Section 3.2, we present several formal results underlying Proposition 3. Our analysis of the estimators \((\hat{\alpha}_2, \hat{\gamma}_2)\), while building on the existing literature on Lasso penalized M-estimation (Bühlmann & Van de Geer [5]; Negahban et al. [21]), needs to tackle the dependency of \( \hat{\gamma}_2 \) on \((\hat{\theta}_1, \hat{\alpha}_1)\) and subsequently that of \( \hat{\alpha}_2 \) on \((\hat{\theta}_1, \hat{\gamma}_2)\). The situation is more general and more complicated than studied in Tan [37]. We develop a technical strategy to control such dependency through use of the \( L_1 \) norm, so that the usual rates of convergence are obtained. See Lemma S6 in the Supplement.

We first discuss theoretical analysis of \( \hat{\gamma}_2 \), with the Lasso tuning parameter \( \lambda_1 = A_1 \lambda_0 \) for a constant \( A_1 \), where \( \lambda_0 = \{\log(p/\epsilon)/n\}^{1/2} \) with \( \epsilon \in (0, 1) \) a tail probability. The loss function for defining \( \hat{\gamma}_2 \) is \( L_2(\gamma; \hat{\theta}_1, \hat{\alpha}_1) = \overline{E}\{\ell_2(U; \hat{\theta}_1, \hat{\alpha}_1, \gamma)\} \), where \( \ell_2 \) is from Assumption 1. As \( L_2(\gamma; \hat{\theta}_1, \hat{\alpha}_1) \) is convex in \( \gamma \), the corresponding Bregman divergence is defined as \( D_2(\gamma'; \gamma; \hat{\theta}_1, \hat{\alpha}_1) = L_2(\gamma'; \hat{\theta}_1, \hat{\alpha}_1) - L_2(\gamma; \hat{\theta}_1, \hat{\alpha}_1) - (\gamma' - \gamma)^T \frac{\partial L_2}{\partial \gamma}(\gamma; \hat{\theta}_1, \hat{\alpha}_1) \). The symmetrized Bregman divergence is easily shown to be

\[
D^s_2(\gamma'; \gamma; \hat{\theta}_1, \hat{\alpha}_1) = (\gamma' - \gamma)^T \left\{ \frac{\partial L_2}{\partial \gamma}(\gamma'; \hat{\theta}_1, \hat{\alpha}_1) - \frac{\partial L_2}{\partial \gamma}(\gamma; \hat{\theta}_1, \hat{\alpha}_1) \right\}
\]

\[
= (\gamma' - \gamma)^T \overline{E} \left\{ \frac{\partial \tau}{\partial \gamma}(U; \hat{\theta}_1, \hat{\alpha}_1, \gamma') - \frac{\partial \tau}{\partial \gamma}(U; \hat{\theta}_1, \hat{\alpha}_1, \gamma) \right\}.
\tag{38}
\]

The target value \( \hat{\gamma}_2 \) is defined as a solution to \( E\{(\partial \tau/\partial \alpha)|(U; \theta^*, \hat{\alpha}_1, \gamma)\} = 0 \) or equivalently a minimizer of the expected loss \( E\{\ell_2(U; \theta^*, \hat{\alpha}_1, \gamma)\} \), where \((\theta^*, \hat{\alpha}_1)\) are the target values (i.e.,
probability limits) of the initial estimators \((\hat{\theta}_1, \hat{\alpha}_1)\). After statement of the assumptions required, Theorem 1 establishes the convergence of \(\hat{\gamma}_2\) to \(\gamma_2\) in the both L1 norm \(||\hat{\gamma}_2 - \gamma_2||_1\) and the symmetrized Bregman divergence \(D^k_2(\gamma_2, \hat{\gamma}_2; \hat{\theta}_1, \hat{\alpha}_1)\).

A variable \(Y\) is said to be sub-exponential with parameter \((B_0, B_2)\) if \(E(|Y - E(Y)|^k) \leq \frac{k!}{2} (B_0^k + B_2^{k-2})\) for each \(k \geq 2\). For a \(p \times p\) matrix \(\Sigma\), a compatibility condition (Bühlmann & Van de Geer [5]) is said to hold with a subset \(S \in \{1, \ldots, p\}\) and constants \(\mu_1 > 1\) if \(\mu^2 \gamma \leq |S| (b^T \Sigma b)\) for any vector \(b \in \{b_1, \ldots, b_p\}^T \in \mathbb{R}^p\) satisfying \(\sum_{j \in S} |b_j| \leq \mu_1 \sum_{j \in S} |b_j|\). Throughout, \(|S|\) denotes the size of a set \(S\).

**Assumption 2.** Suppose that the following conditions are satisfied.

(i) \(\max_{j=1, \ldots, p} |\xi_j(X)| \leq C_0\) almost surely for a constant \(C_0 > 0\).

(ii) The variable \(\frac{\partial^2}{\partial \theta \partial \gamma} U(\theta^*, \alpha_1, \gamma_2)\) is sub-exponential with parameter \((B_0, B_2)\).

(iii) The compatibility condition holds for \(\Sigma_\gamma = E\{\xi \xi^T \frac{\partial^2}{\partial \theta \partial \gamma} U(\theta^*, \alpha_1, \gamma_2)\}\) with the subset \(S_\gamma = \{j : (\gamma_2)_j \neq 0, j = 1, \ldots, p\}\) and some constants \(\mu_1 > 0\) and \(\mu_1 > 1\).

(iv) For some constants \(c_0 > 0\) and \(M_0 \geq 1\), possibly depending on \((\alpha_1, \gamma_1)\), and any small \(\epsilon > 0\), it holds with probability at least \(1 - c_0\epsilon\) that

\[
(\alpha_1 - \hat{\alpha}_1)^T \hat{\Sigma}_0 (\alpha_1 - \hat{\alpha}_1) \leq M_0 \lambda_0^2, \quad \|\hat{\alpha}_1 - \alpha_1\|_1 \leq M_0 \lambda_0, \quad |\hat{\theta}_1 - \theta^*| \leq M_0^{1/2} \lambda_0,
\]

where \(\hat{\Sigma}_0 = \hat{E}((\xi \xi^T))\), \(\lambda_0 = (\log(p)/n)^{1/2}\), and \(\alpha_1 = \gamma^*\) if model (11) is correctly specified or \(\gamma_1 = \gamma^*\) if model (12) is correctly specified.

**Assumption 3.** There exist positive constants \(c_1, c_2, B_{11}, B_{12}, C_1, C_2, g_0, g_1\) such that the following conditions are satisfied, where \(N_1 = \{\theta, \alpha : |\theta - \theta^*| \leq c_1, ||\alpha - \alpha_1||_1 \leq c_1\}\).

(i) The variables \(T^{(1)}(U; \theta^*, \alpha_1, \gamma_2) = \sup_{\theta, \alpha \in N_1} \frac{\partial^2}{\partial \theta \partial \gamma} U(\theta, \alpha, \gamma_2)\) and \(T^{(1)}(U; \theta^*, \alpha_1, \gamma_2)\) are sub-exponential with parameter \((B_{11}, B_{12})\), and \(E\{T^{(1)}(U; \theta^*, \alpha_1, \gamma_2)|X\} \leq C_1\) almost surely.

(ii) The variable \(\frac{\partial^2}{\partial \theta \partial \gamma} U(\theta^*, \alpha_1, \gamma_2)\) is sub-exponential with parameter \((B_{11}, B_{12})\), and \(E\{\frac{\partial^2}{\partial \theta \partial \gamma} U(\theta^*, \alpha_1, \gamma_2)|X\} \geq c_2\) almost surely.

(iii) For any \((\theta, \alpha) \in N_1\) and \(\gamma \in \mathbb{R}^p\), it holds that \(\frac{\partial^2}{\partial \theta \partial \gamma} U(\theta, \alpha, \gamma) \geq \frac{\partial^2}{\partial \theta \partial \gamma} U(\theta^*, \alpha_1, \gamma_2)\)

\[e^{-C_2 ((\theta - \theta^*)^2 + (\alpha - \alpha_1)^2 + (\gamma - \gamma_2)^2)}\]

almost surely.

(iv) \(M_0 \lambda_0 \leq g_0 \leq c_1\) and \(|S_{\gamma_2}| \leq g_1\) such that \(g_2 = \nu_1^{-2} (1 + \mu_2^2) g_1 B_{15} < 1\), \(g_3 = C_0 C_2 A_{14}^2 \times \mu_2^2 g_1^2 g_6 e_n < 1\), and \(g_4 = C_0 C_2 A_{14}^{-1} \mu_2 C_1 B_{12} B_{10} g_6 e_n < 1\), where \(g_5 = C_2 (1 + C_0) g_0, A_{11} = A_{11} - B_0 - C_{13} = 1 - 2A_{11} \{A_{11} \in [0, 1]\}, \mu_2 = (\mu_1 + 1) A_{11}, \nu_1 = \nu_1 (1 - g_1)^{1/2}, b_0 = B_0 + \sqrt{2} B_{10}\), \(B_{15}\) is defined in Lemma S2 depending on \((C_0, C_1, B_{11}, B_{12})\), and \((C_{12}, C_{13})\) are defined in Lemma S6 depending on \((g_0, c_2, C_0, C_1, B_{11}, B_{12})\).

**Theorem 1.** Suppose that Assumptions 1–3 hold and \(\lambda_0 \leq 1\). Then for \(A_1 = A_1 \lambda_0 < A_1 > (B_0 + C_{13})/(\mu_1 - 1)\), we have that with probability at least \(1 - (c_0 + 10)\epsilon\),

\[
D^k_2(\gamma_2, \hat{\gamma}_2; \theta_1, \hat{\alpha}_1) + A_1 \lambda_0 \|\gamma_2 - \hat{\gamma}_2\|_1
\leq \{e^{1/2} (1 - g_3)^{-1} \mu_2^2 (S_{\gamma_2} [\lambda_0^2]) \bigvee \{e^{1/2} (1 - g_4) \mu_1^{-1} C_1 (M_0 \lambda_0^2)\}\},
\] (39)
where \( \lor \) denotes the maximum between two numbers, and \( (\mu_{11}, \mu_{12}, \nu_{11}, \varrho_3, \varrho_4, \varrho_5, A_{11}, B_0, C_{12}, C_{13}) \) are defined in Assumption 3(iv).

**Remark 4.** Assumptions 2(iii) and 3(iii) are standard in high-dimensional analysis of M-estimation (e.g., Bühlmann & Van de Geer [5]; Tan [38]). Assumptions 3(i)–(ii) are used to control the deviation of \( \hat{\theta}_1, \hat{\alpha}_1 \) from \( (\theta^*, \alpha_1^*) \) in the basic inequality. Given Assumption 3(ii), the compatibility condition on \( \Sigma_\gamma \) in Assumption 2(iii) can be equivalently replaced by a compatibility condition on the matrix \( \Sigma_0 = E(\xi \xi^T) \), independent of \( (\theta^*, \alpha_1^*, \gamma)^2 \).

**Remark 5.** Assumption 2(iv) is concerned with the convergence of the initial estimators \( \hat{\theta}_1, \hat{\alpha}_1, \hat{\gamma}_1 \). In fact, \( \hat{\theta}_1 \) is required to converge to \( \theta^* \) at rate \( M_0^{1/2} \lambda_0 \) if model (11) or (12) is correctly specified. Hence \( \hat{\theta}_1 \) is pointwise doubly robust, although it does not in general admit doubly robust confidence intervals. For \( \hat{\theta}_1 = \hat{\theta}(\hat{\alpha}_1, \hat{\gamma}_1) \) in Algorithm 1, the required convergence for \( \hat{\theta}_1 \) can be deduced from the stated rates of convergence for \( (\hat{\alpha}_1, \hat{\gamma}_1) \) under suitable conditions, similar to Assumptions 6–7 for Theorem 3 later. For simplicity, the convergence of \( \hat{\theta}_1 \) is included as part of Assumption 2(iv). This formulation also allows Theorem 1 to be applied with other possible choices of \( \hat{\theta}_1 \). See the proof of Corollary 4.

The following corollary provides a bound on the prediction \( L_2 \) norm (in the scale of linear predictors \( \eta_f \) ), \( \hat{E}(\{\hat{\gamma}_2 - \hat{\gamma}_2\}^2) = (\hat{\gamma}_2 - \hat{\gamma}_2)^T \hat{\Sigma}_0 (\hat{\gamma}_2 - \hat{\gamma}_2) \), where \( \hat{\Sigma}_0 = \hat{E}(\xi \xi^T) \).

**Corollary 3.** In the setting of Theorem 1, with probability at least \( 1 - (c_0 + 10)\epsilon \), we have that, in addition to (39),

\[
(\hat{\gamma}_2 - \hat{\gamma}_2)^T \hat{\Sigma}_0 (\hat{\gamma}_2 - \hat{\gamma}_2) 
\leq \left\{ c_2^{-1} \epsilon^2 (1 - \varrho_3 \lor \varrho_4)^{-1} + (1 + c_2^{-1}) B_1 A_{11}^{-2} C_3 (\varrho_0 \lor \varrho_1) \right\} C_3 (|S_{\gamma_2}| \lor M_0) \lambda_0^2, \quad (40)
\]

where \( B_1 = (4C_2^2) \lor B_{15} \), and \( C_3 \) is a constant such that the right hand side of (39) is upper bounded by \( C_3 (|S_{\gamma_2}| \lor M_0) \lambda_0^2 \).

From Theorem 1 and Corollary 3, let \( M_1 (\geq M_0) \) be a constant such that the right hand side of (39) is upper bounded by \( A_{11} M_1 \lambda_0^2 \) and that of (40) is upper bounded by \( M_1 \lambda_0^2 \). Then with probability at least \( 1 - (c_0 + 10)\epsilon \), we have

\[
(\hat{\gamma}_2 - \hat{\gamma}_2)^T \hat{\Sigma}_0 (\hat{\gamma}_2 - \hat{\gamma}_2) \leq M_1 \lambda_0^2, \quad \|\hat{\gamma}_2 - \hat{\gamma}_2\| \leq M_1 \lambda_0. \quad (41)
\]

These bounds can be used to justify a rate condition on the convergence of \( \hat{\gamma}_2 \) corresponding to Assumption 2(iv), and to obtain a similar result to Theorem 1 about the convergence of \( \hat{\alpha}_2 \) to a target value \( \bar{\alpha}_2 \), which is defined as a solution to \( E\{\partial \tau / \partial \gamma \} (U; \theta^*, \alpha, \gamma) = 0 \) or equivalently a minimizer of the expected loss \( E\{\ell_1(U; \theta^*, \alpha, \gamma)\} \).

**Assumption 4.** Suppose that the conditions (ii)–(iii) in Assumption 2 hold, with \( (\hat{\alpha}_1, \hat{\gamma}_2) \) replaced by \( (\hat{\alpha}_2, \hat{\gamma}_2) \), \( \partial \tau / \partial \theta^*_\eta \) by \( \partial \tau / \partial \eta_f \), and \( (B_0, B_{12}, \mu_1, \nu_1) \) replaced by some alternative constants throughout. For example, \( \partial \tau / \partial \eta_f \) (\( U; \theta^*, \bar{\alpha}_2, \hat{\gamma}_2) \) is sub-exponential.

**Assumption 5.** Suppose that the conditions (i)–(iv) in Assumption 3 hold, with \( (\hat{\alpha}_1, \hat{\gamma}_2) \) replaced by \( (\hat{\alpha}_2, \hat{\gamma}_2) \) (or \( (\hat{\alpha}_1, \hat{\gamma}_2), (\partial^2 \tau / \partial \eta^2_0, \partial^2 \tau / \partial \eta_0 \partial \theta), (\partial^2 \tau / \partial \eta^2_2, \partial^2 \tau / \partial \eta_2 \partial \theta), (\partial^2 \tau / \partial \eta^2_1, \partial^2 \tau / \partial \eta_1 \partial \theta) \)) by \( (\partial^2 \tau / \partial \eta^2_0, \partial^2 \tau / \partial \eta_0 \partial \theta), M_0 \) by \( M_1 \), and \( (c_1, c_2, B_{11}, B_{12}, C_1, C_2, \varrho_0, \varrho_1) \) by some alternative constants throughout.
Theorem 2. In the setting of Theorem 1, suppose that Assumptions 4–5 also hold. Then for \( \lambda_2 = A_2\lambda_0 \) and sufficiently large \( A_2 \), we have that with probability at least \( 1 - (c_0 + 18)\epsilon \), in addition to (41),
\[
(\hat{\alpha}_2 - \tilde{\alpha}_2)^T \hat{\Sigma}_0 (\hat{\alpha}_2 - \tilde{\alpha}_2) \leq M_2\lambda_0^2, \quad \|\hat{\alpha}_2 - \tilde{\alpha}_2\|_1 \leq M_2\lambda_0,
\]
where \( M_2 \) is a constant determined similarly as \( M_1 \) in (41).

With the preceding results about \((\hat{\alpha}_2, \hat{\gamma}_2)\), we are ready to study the convergence of \( \hat{\theta}_2 = \hat{\theta}(\hat{\alpha}_2, \hat{\gamma}_2) \). As convergence in probability is of main interest, the high-probability bounds (41) and (42) can be used to deduce the following in-probability statements: \((\hat{\alpha}_2 - \tilde{\alpha}_2)^T \hat{\Sigma}_0 (\hat{\alpha}_2 - \tilde{\alpha}_2) = O_p(M_2^{-2})\), \(\|\hat{\alpha}_2 - \tilde{\alpha}_2\|_1 = O_p(M_2\lambda_0)\), \((\hat{\gamma}_2 - \tilde{\gamma}_2)^T \hat{\Sigma}_0 (\hat{\gamma}_2 - \tilde{\gamma}_2) = O_p(M_2^{-2})\), and \(\|\hat{\gamma}_2 - \tilde{\gamma}_2\|_1 = O_p(M_2\lambda_0)\), where \( r_0 = (\log(ep)/n)^{1/2} \). After statement of assumptions required, Theorem 3 establishes the desired convergence result for \( \hat{\theta}_2 \).

Assumption 6. Suppose that the following conditions are satisfied.

(i) \( E\{\tau(U; \theta^*, \hat{\alpha}_2, \hat{\gamma}_2)\} = 0 \) and \( \inf_{\theta \in \Theta} |\theta - \theta^*| > \delta > 0 \).

(ii) \( E\{\sup_{\theta \in \Theta} |\tau(U; \theta, \hat{\alpha}_2, \hat{\gamma}_2)|\} < \infty \).

(iii) There exists a neighborhood \( N_2 = \{\alpha, \gamma : \|\alpha - \hat{\alpha}_2\|_1 \leq c_3, \|\gamma - \hat{\gamma}_2\|_1 \leq c_3\} \) for a constant \( c_3 > 0 \) such that \( E\{T_{\hat{n}_2}(U; \hat{\alpha}_2, \hat{\gamma}_2)\} < \infty \) and \( E\{T_{\hat{n}_2}(U; \hat{\alpha}_2, \hat{\gamma}_2)\} < \infty \), where \( T_{\hat{n}_2}(U; \hat{\alpha}_2, \hat{\gamma}_2) = \sup_{\theta \in \Theta, (\alpha, \gamma) \in N_2} |\frac{\partial \tau}{\partial \theta}(U; \theta, \alpha, \gamma)| \) and \( T_{\hat{n}_2}(U; \hat{\alpha}_2, \hat{\gamma}_2) = \sup_{\theta \in \Theta, (\alpha, \gamma) \in N_2} |\frac{\partial \tau}{\partial \theta}(U; \theta, \alpha, \gamma)| \).

Assumption 7. There exist positive constants \( c_4 \) and \( C_4 \) such that the following conditions are satisfied, where \( N_3 = \{\theta, \alpha, \gamma : |\theta - \theta^*| \leq c_4, \|\alpha - \hat{\alpha}_2\|_1 \leq c_4, \|\gamma - \hat{\gamma}_2\| \leq c_4\} \).

(i) \( E\{\sup_{\theta, \alpha, \gamma \in N_3} |\tau(U; \theta, \alpha, \gamma)|\} < \infty \).

(ii) \( H = E\{|\frac{\partial \tau}{\partial \theta}(U; \theta^*, \hat{\alpha}_2, \hat{\gamma}_2)|\} = 0 \) and \( E\{|\sup_{\theta, \alpha, \gamma \in N_3} |\frac{\partial \tau}{\partial \theta}(U; \theta, \alpha, \gamma)|\} < \infty \).

(iii) The variables \( \frac{\partial \tau}{\partial \theta}(U; \theta^*, \alpha, \gamma) \) and \( \frac{\partial \tau}{\partial \theta}(U; \theta^*, \alpha, \gamma) \) are sub-exponential.

(iv) The variables \( T_{\hat{n}_2}(U; \theta^*, \hat{\alpha}_2, \hat{\gamma}_2) = \sup_{\theta, (\alpha, \gamma) \in N_3} |\frac{\partial \tau}{\partial \theta}(U; \theta, \alpha, \gamma)| \), \( T_{\hat{n}_2}(U; \theta^*, \hat{\alpha}_2, \hat{\gamma}_2) = \sup_{\theta, (\alpha, \gamma) \in N_3} |\frac{\partial \tau}{\partial \theta}(U; \theta, \alpha, \gamma)| \), and \( T_{\hat{n}_2}(U; \theta^*, \hat{\alpha}_2, \hat{\gamma}_2) = \sup_{\theta, (\alpha, \gamma) \in N_3} |\frac{\partial \tau}{\partial \theta}(U; \theta, \alpha, \gamma)| \) are sub-exponential, and \( E\{T_{\hat{n}_2}(U; \theta^*, \hat{\alpha}_2, \hat{\gamma}_2)|X\} \leq C_4 \), \( E\{T_{\hat{n}_2}(U; \theta^*, \hat{\alpha}_2, \hat{\gamma}_2)|X\} \leq C_4 \), and \( E\{T_{\hat{n}_2}(U; \theta^*, \hat{\alpha}_2, \hat{\gamma}_2)|X\} \leq C_4 \) almost surely.

Theorem 3. In the setting of Theorem 2, suppose that Assumptions 6 and 7 hold and \( M_2\lambda_0 = o(1) \). Then \( \hat{\theta}_2 \) is consistent for \( \theta^* \) and admits the asymptotic expansion
\[
\hat{\theta}_2 - \theta^* = -H^{-1}\hat{\tau}(U; \theta^*, \hat{\alpha}_2, \hat{\gamma}_2) + O_p(M_2\lambda_0^2),
\]
where \( H = E\{|\frac{\partial \tau}{\partial \theta}(U; \theta^*, \hat{\alpha}_2, \hat{\gamma}_2)|\} \). Moreover, a consistent estimator of \( \tau(U; \theta^*, \hat{\alpha}_2, \hat{\gamma}_2) \)
\( /H^2 \) is \( \hat{\tau} = \hat{E}\{\tau(U; \hat{\theta}_2, \hat{\alpha}_2, \hat{\gamma}_2)\}/H^2 \), where \( \hat{H} = \hat{E}\{|\frac{\partial \tau}{\partial \theta}(U; \hat{\theta}_2, \hat{\alpha}_2, \hat{\gamma}_2)|\} \).

Remark 6. Assumption 6 is involved to show the consistency of \( \hat{\theta}_2 \) for \( \theta^* \). Assumptions 6(i)–(ii) are standard for showing consistency if \( \tau(U; \theta, \hat{\alpha}_2, \hat{\gamma}_2) \) were employed as an estimating
function in $\theta$ (e.g., Van der Vaart [42]). Assumption 6(iii) is used to control the deviation of $(\hat{\alpha}_2, \hat{\gamma}_2)$ from the target values, with unrestricted $\theta \in \Theta$. Moreover, Assumption 7 is involved to show the asymptotic expansion (43). Assumption 7(i)–(ii) is adapted from classical asymptotic theory for maximum likelihood estimation (e.g., Ferguson [13]). Assumption 7(iv) is used to control the deviation of $(\hat{\theta}_2, \hat{\alpha}_2, \hat{\gamma}_2)$.

Combining Theorems 1–3 leads to Proposition 3 provided $M_2p^2 = o(n^{-1/2})$, i.e., the remainder term in (43) reduces to $o_p(n^{-1/2})$. As motivated in Section 2.2 and made explicit in the proofs, the primary reason for $\hat{\theta}_2$ to achieve asymptotic expansion (43) is that the two-step estimators $(\hat{\alpha}_2, \hat{\gamma}_2)$ are constructed such that according to Proposition 2, the target values $(\hat{\alpha}_2, \hat{\gamma}_2)$ satisfy the calibration equations (22)–(23) if model (11) or (12) is correctly specified. In this case, both the linear and quadratic terms in $(\hat{\alpha}_2, \hat{\gamma}_2)$ are constructed such that according to Proposition 2, the target values $(\hat{\alpha}_2, \hat{\gamma}_2)$ are O_p($M_2r^2_0$) from a Taylor expansion argument. Otherwise, the linear term would in general be $O_p(M_2^{1/2}r_0)$, as reflected in the convergence rate for the initial estimator $\hat{\theta}_1$ in Assumption 2(iv).

4. Applications

Example 10. Return to Examples 1 and 6 with a partially linear model (4). For $g(x; \alpha) = \alpha^T \xi$ and $f(x; \gamma) = \psi_f(\gamma^T \xi)$, models (11) and (12) can be stated as

$$E(Y|Z,X) = \theta Z + \alpha^T \xi,$$

$$E(Z|X) = \psi_f(\gamma^T \xi).$$

(44)

For estimating function $\tau$ in (5) and any estimators $(\hat{\alpha}, \hat{\gamma})$, $\hat{\theta} = \hat{\theta}(\hat{\alpha}, \hat{\gamma})$ as a solution to $\hat{E}\{\tau(U; \theta, \hat{\alpha}, \hat{\gamma})\} = 0$ is of closed form:

$$\hat{\theta}(\hat{\alpha}, \hat{\gamma}) = \frac{\hat{E}\{(Y - \hat{\alpha}^T \xi)(Z - \psi_f(\hat{\gamma}^T \xi))\}}{\hat{E}\{Z(Z - \psi_f(\hat{\gamma}^T \xi))\}}.$$

(45)

For initial estimation, let $(\hat{\theta}_0, \hat{\alpha}_1)$ be Lasso regularized least-squares estimators in model (44), $\hat{\gamma}_1$ be a Lasso regularized quasi-likelihood estimator in model (45), and $\hat{\theta}_1 = \hat{\theta}(\hat{\alpha}_1, \hat{\gamma}_1)$. For second-step estimation, the regularized calibrated estimator $\hat{\gamma}_2$ is defined with a Lasso penalty and the loss function

$$L_2(\gamma) = \hat{E}\{\ell_2(U; \gamma)\} = \hat{E}\{-Z\gamma^T \xi + \Psi_f(\gamma^T \xi)\},$$

(46)

and $\hat{\alpha}_2$ is defined with a Lasso penalty and the loss function

$$L_1(\alpha; \hat{\theta}_1, \hat{\gamma}_2) = \hat{E}\{\ell_1(U; \hat{\theta}_1, \alpha, \hat{\gamma}_2)\} = \hat{E}\{\psi'_{f}(\hat{\gamma}_2^T \xi)(Y - \hat{\theta}_1 Z - \alpha^T \xi)^2\},$$

(47)

where $\Psi_f(t) = \int_0^t \psi_f(u) du$ and $\psi'_f$ is the derivative of $\psi_f$, and $\ell_1$ and $\ell_2$ are determined from (33), with $\partial \tau / \partial \alpha_1, \partial \tau / \partial \gamma_1$ in (25)–(26). The estimator $\hat{\gamma}_2$ coincides with the usual estimator $\hat{\gamma}_1$ with a canonical link in (45), whereas $\hat{\alpha}_2$ can be interpreted as a regularized weighted least squares estimator. The resulting estimator of $\theta$ is then $\hat{\theta}_2 = \hat{\theta}(\hat{\alpha}_2, \hat{\gamma}_2)$.

We stress that the loss (47) is for estimation of $\alpha$ with $(\hat{\theta}_1, \hat{\gamma}_2)$ fixed, and $\hat{\theta}_1$ is determined as $\hat{\theta}(\hat{\alpha}_1, \hat{\gamma}_1)$ and hence pointwise doubly robust (see Remark 5). In other words, for $\hat{\theta}_2$ to
admit doubly robust confidence intervals as in Proposition 3, it is in general incorrect to (i) replace \( \hat{\theta}_1 \) in (47) by \( \hat{\theta}_0 \) computed from the first step, or (ii) to redefine \((\hat{\theta}_1, \hat{\gamma}_1)\) jointly as a regularized weighted least squares estimator in model (44), with weight \( \psi_f(\hat{\gamma}_2^T \xi) \). Nevertheless, these simple options become valid in the special situation where \( \psi_f() \) is an identity function, i.e., (45) is a linear model. In this case, \( \hat{\gamma}_2 \) can be taken the same as \( \hat{\gamma}_1 \) because (46) becomes the usual least-squares loss, and then either option (i) or (ii) can be shown to yield \( \hat{\theta}_2 \) identical to the first-step estimator \( \hat{\theta}(\hat{\alpha}_1, \hat{\gamma}_1) \), provided that the same Lasso tuning parameter is used in computing \( \hat{\alpha}_2 \) as in computing \((\hat{\theta}_0, \hat{\alpha}_1)\). See the proof of Corollary 4. Moreover, \( \hat{\theta}(\hat{\alpha}_1, \hat{\gamma}_1) \) can be expressed as a debiased Lasso estimator of \( \theta \) in linear regression (44) (Zhang & Zhang [46]; Van de Geer et al. [41]; Javanmard & Montanari [15]):

\[
\hat{\theta}_{DB} = \hat{\theta}(\hat{\alpha}_1, \hat{\gamma}_1) = \frac{E\{(Y - \hat{\alpha}_1^T \xi)(Z - \hat{\gamma}_1^T \xi)\}}{E\{Z(Z - \hat{\gamma}_1^T \xi)\}} = \hat{\theta}_0 + \frac{E\{(Y - \hat{\theta}_0 Z - \hat{\alpha}_1^T \xi)(Z - \hat{\gamma}_1^T \xi)\}}{E\{Z(Z - \hat{\gamma}_1^T \xi)\}},
\]

where \((\hat{\theta}_0, \hat{\alpha}_1)\) are jointly Lasso estimators in linear regression of \( Y \langle Z, X \rangle \), and \( \hat{\gamma}_1 \) is that in linear regression of \( Z \langle X \rangle \). Suppose that the Lasso tuning parameters are sufficiently large, of order \( O((\log(ep)/n)^{1/2}) \). The following result can be deduced from Proposition 3.

**Corollary 4.** Suppose that Assumption 2(i) and a compatibility condition holds for \( \Sigma_0 = E(\xi \xi^T) \), \( Y - \theta^* Z - \hat{\alpha}_1^T \xi \) and \( Z - \hat{\gamma}_1^T \xi \) are sub-exponential, \( V = E\{(Y - \theta^* Z - \hat{\alpha}_1^T \xi)^2(Z - \hat{\gamma}_1^T \xi)^2\} < \infty \), and \( H = -E\{Z(Z - \hat{\gamma}_1^T \xi)\} \neq 0 \). If model (44) or model (45) with \( \psi_f \equiv 1 \) is correctly specified, then the conclusions in Proposition 3 are valid for \( \hat{\theta}_{DB} = \hat{\theta}(\hat{\alpha}_1, \hat{\gamma}_1) \), where

\[
\hat{V} = \hat{E}\{(Y - \hat{\theta}_{DB} Z - \hat{\alpha}_1^T \xi)^2(Z - \hat{\gamma}_1^T \xi)^2\}/E^2\{Z(Z - \hat{\gamma}_1^T \xi)\}.
\]

Hence a doubly robust confidence interval for \( \theta^* \) is obtained in partially linear model (4).

From Corollary 4, the debiased Lasso estimator \( \hat{\theta}_{DB} \) in linear regression (44) can be used to obtain doubly robust Wald confidence intervals for \( \theta^* \) in a partially linear model. This finding gives a high-dimensional extension of the double robustness (including pointwise and Wald confidence intervals) of least-squares estimation in low-dimensional settings (Example 5). Doubly robust confidence intervals based on score tests are also obtained in Dukes & Vansteelandt [11], Corollary 1. We make several additional comments. First, although \( \hat{\theta}_{DB} \) is the same point estimator, the variance estimator \( \hat{V} \) differs from those originally in debiased Lasso, in the context of linear regression with a constant error variance, which then needs to be estimated (Zhang & Zhang [46]; Van de Geer et al. [41]; Javanmard & Montanari [15]).

Second, Bühlmann & Van de Geer [6] studied debiased Lasso in possibly misspecified linear regression. They employed the same point estimator \( \hat{\theta}_{DB} \) and proposed a variance estimator similar to \( \hat{V} \),

\[
\hat{V}_{BG} = \hat{E}\{(\hat{\varepsilon} \hat{Z} - \hat{E}(\hat{\varepsilon} \hat{Z}))^2\}/E^2(Z \hat{Z}),
\]

where \( \hat{\varepsilon} = Y - \hat{\theta}_0 Z - \hat{\alpha}_1^T \xi \) and \( \hat{Z} = Z - \hat{\gamma}_1^T \xi \). Specifically, \( \hat{V}_{BG} \) can be obtained from \( \hat{V} \) by replacing \( \hat{\theta}_{DB} \) with \( \hat{\theta}_0 \) and the sample second-moment of the product \( (Y - \hat{\theta}_0 Z - \hat{\alpha}_1^T \xi)(Z - \hat{\gamma}_1^T \xi) \) with the sample variance. Bühlmann & Van de Geer [6] showed that under suitable conditions, \( \hat{\theta}_{DB} \pm z_{c/2} \sqrt{\hat{V}_{BG}} \) is a \((1 - c)\) confidence interval for \( \hat{\theta}_0 \), defined such that

\[
(\hat{\theta}_0, \hat{\alpha}_1) = \arg\min_{(\theta, \alpha)} E\{(Y - \theta Z - \alpha^T \xi)^2\},
\]
with possible misspecification of linear model (44). This result is compatible with ours, because, from the proof of Corollary 4, \( \hat{\theta}_0 \) identifies \( \theta^* \) in partially linear model (4) if linear model (44) is misspecified but a linear model for \( E(Z|X) \) is correctly specified.

Finally, for a nonlinear model (12) with \( \psi_f \) a non-identity function (for example when \( Z \) is binary or nonnegative), our estimator \( \hat{\theta}_2 \) and associated confidence intervals are distinct from debiased Lasso including Bühlmann & Van de Geer [6]. Although \( \hat{\theta}_{db} \pm z_{c/2} \sqrt{V_{bc}} \) remains a \((1-c)\) confidence interval for \( \theta_0 \) under suitable conditions, the target value \( \theta_0 \) may in general differ from \( \theta^* \) in partially linear model (4) even if model (44) is misspecified but a nonlinear model for \( E(Z|X) \) is correctly specified.

**Example 11.** Return to Examples 2 and 7 with a partially log-linear model (6). For \( g(x; \alpha) = \alpha^T \xi \) and \( f(x; \gamma) = \psi_f(\gamma^T \xi) \), models (11) and (12) can be stated as
\[
E(Y|Z, X) = \exp(\theta Z + \alpha^T \xi), \tag{48}
\]
\[
E(Z|X) = \psi_f(\gamma^T \xi). \tag{49}
\]
For estimating function \( \tau \) in (7) and any estimators \((\hat{\alpha}, \hat{\gamma})\), \( \hat{\theta}(\hat{\alpha}, \hat{\gamma}) \) is a solution to
\[
0 = \bar{E}\{\tau(U; \theta, \hat{\alpha}, \hat{\gamma})\} = \bar{E}\{(Y e^{-\theta Z} - e^{\hat{\alpha}^T \xi})(Z - \psi_f(\hat{\gamma}^T \xi))\}. \tag{50}
\]
For initial estimation, let \((\hat{\theta}_0, \hat{\alpha}_1)\) be Lasso regularized quasi-likelihood estimators in model (48), \( \hat{\gamma}_1 \) be that in model (49), and \( \hat{\theta}_1 = \hat{\theta}(\hat{\alpha}_1, \hat{\gamma}_1) \). For second-step estimation, the regularized calibrated estimator \( \hat{\gamma}_2 \) is defined with a Lasso penalty and the loss function
\[
L_2(\gamma; \hat{\alpha}_1) = \bar{E}\{\ell_2(U; \hat{\alpha}_1, \gamma)\} = \bar{E}\{e^{\hat{\alpha}^T \xi} \left(-Z \hat{\gamma}^T \xi + \psi_f(\hat{\gamma}^T \xi)\right)\}, \tag{51}
\]
and \( \hat{\alpha}_2 \) is defined with a Lasso penalty and the loss function
\[
L_1(\alpha; \hat{\theta}_1, \hat{\gamma}_2) = \bar{E}\{\ell_1(U; \hat{\theta}_1, \alpha, \hat{\gamma}_2)\} = \bar{E}\{\psi_f'(\hat{\gamma}_2^T \xi) \left(-Y e^{-\hat{\theta}_1 Z} + e^{\alpha^T \xi}\right)\}, \tag{52}
\]
where \( \ell_1 \) and \( \ell_2 \) are determined from (39), with \((\partial \tau/\partial \alpha, \partial \tau/\partial \gamma)\) in (27)–(28). Unlike (46), the loss (51) in \( \gamma \) depends on \( \alpha_1 \). The resulting estimator of \( \theta \) is then \( \hat{\theta}_2 = \hat{\theta}(\hat{\alpha}_2, \hat{\gamma}_2) \).

In contrast with Example 10, our method is distinct from debiased Lasso, even when \( \psi_f = 1 \), i.e., (49) is a linear model. Similarly as in our method, let \((\hat{\theta}_0, \hat{\alpha}_1)\) be the Lasso estimators associated with the loss \( \bar{E}\{-Y(\theta Z + \alpha^T \xi) + e^{\theta Z+\alpha^T \xi}\} \), and \( \hat{\gamma}_1 \) be that associated with the loss \( \bar{E}\{e^{\theta_0 Z+\hat{\alpha}_1^T \xi}(Z - \gamma^T \xi)^2\} \). The debiased Lasso estimator in Van de Geer et al. [41], also called the one-step estimator in Ning & Liu [23], is
\[
\hat{\theta}_{db} = \hat{\theta}_0 + \frac{\bar{E}\{(Y - e^{\theta_0 Z+\hat{\alpha}_1^T \xi})(Z - \hat{\gamma}_1^T \xi)\}}{\bar{E}\{e^{\theta_0 Z+\hat{\alpha}_1^T \xi}Z(Z - \hat{\gamma}_1^T \xi)\}}.
\]
A variation of debiased Lasso in Neykov et al. [22] is to define \( \hat{\theta}_{db2} \) as a solution to
\[
\bar{E}\{(Y - e^{\theta Z+\hat{\alpha}_1^T \xi})(Z - \hat{\gamma}_1^T \xi)\} = 0. \tag{52}
\]
Equation (52) is somewhat similar to (50) with \( \psi_f \equiv 1 \), but there is an important difference. Equation (50) is doubly robust: its limit version, with \((\hat{\alpha}_1, \hat{\gamma}_1)\) replaced by their limit values and \(\bar{E}(\cdot)\) replaced by \(E(\cdot)\), holds at \( \theta = \theta^* \) in partially log-linear model (6) if either model (48) or model (49) is correctly specified. In contrast, (52) is not doubly robust: its limit version in general holds at \( \theta = \theta_0 \), defined such that

\[
(\hat{\theta}_0, \hat{\alpha}_1) = \arg\min_{\theta, \alpha} E \left\{ -Y(\theta Z + \alpha^T \xi) + e^{\theta Z + \alpha^T \xi} \right\}.
\]

The target value \( \bar{\theta}_0 \) coincides with \( \theta^* \) if model (48) is correctly specified, but in general may differ from \( \theta^* \) otherwise including when model (49) with \( \psi_f \equiv 1 \) is correctly specified.

**Example 12.** Return to Examples 3 and 8 with a partially logistic model (8). For \( g(x; \alpha) = \alpha^T \xi \) and \( f(x; \gamma) = \psi_f(\gamma^T \xi) \), models (11) and (12) can be stated as

\[
E(Y|Z, X) = \expit(\theta Z + \alpha^T \xi),
\]

\[
E(Z|Y = 0, X) = \psi_f(\gamma^T \xi).
\]

For estimating function \( \tau \) in (9) and any estimators \((\hat{\alpha}, \hat{\gamma})\), \( \hat{\theta}(\hat{\alpha}, \hat{\gamma}) \) is a solution to

\[
0 = \bar{E}\{\tau(U; \theta, \hat{\alpha}, \hat{\gamma})\} = \bar{E}\{e^{-\theta Z Y}(Y - \expit(\hat{\alpha}^T \xi))(Z - \psi_f(\gamma^T \xi))\}.
\]

For initial estimation, let \((\hat{\theta}_0, \hat{\alpha}_1)\) be Lasso likelihood estimators in model (53), \( \hat{\gamma}_1 \) be a Lasso quasi-likelihood estimator in model (54), and \( \hat{\theta}_1 = \hat{\theta}(\hat{\alpha}_1, \hat{\gamma}_1) \). For second-step estimation, the regularized calibrated estimator \( \hat{\gamma}_2 \) is defined with a Lasso penalty and the loss function

\[
L_2(\gamma; \hat{\theta}_1, \hat{\alpha}_1) = \bar{E}\{\ell_2(U; \hat{\theta}_1, \hat{\alpha}_1, \gamma)\} = \bar{E}\left\{e^{-\hat{\theta}_1 Z Y}\expit(\hat{\alpha}_1^T \xi)\left\{-Z\gamma^T \xi + \Psi_f(\gamma^T \xi)\right\}\right\},
\]

and \( \hat{\alpha}_2 \) is defined with a Lasso penalty and the loss function

\[
L_1(\alpha; \hat{\theta}_1, \hat{\gamma}_2) = \bar{E}\{\ell_1(U; \hat{\theta}_1, \alpha, \hat{\gamma}_2)\} = \bar{E}\left\{e^{-\hat{\theta}_1 Z Y}\psi_f(\hat{\gamma}_2^T \xi)\left\{-Y\alpha^T \xi + \log(1 + e^{\alpha^T \xi})\right\}\right\},
\]

where \( \ell_1 \) and \( \ell_2 \) are determined from (33), with \((\partial \tau/\partial \alpha, \partial \tau/\partial \gamma)\) in (29)–(30). The resulting estimator of \( \theta \) is then \( \hat{\theta}_2 = \hat{\theta}(\hat{\alpha}_2, \hat{\gamma}_2) \).

Our method in Example 12 differs from debiased Lasso even more substantially than in Examples 10–11. The debiased Lasso estimator in van de Geer et al. (2014) is

\[
\hat{\theta}_{db} = \hat{\theta}_0 + \bar{E} \left\{ (Y - \expit(\hat{\theta}_0 Z + \hat{\alpha}_1^T \xi))(Z - \hat{\gamma}_1^T \xi) \right\} / \bar{E} \left\{ \expit(\hat{\theta}_0 Z + \hat{\alpha}_1^T \xi)Z(Z - \hat{\gamma}_1^T \xi) \right\},
\]

and a variation \( \hat{\theta}_{db2} \) in Neykov et al. [22] is a solution to

\[
\bar{E} \left\{ (Y - \expit(\theta Z + \hat{\alpha}_1^T \xi))(Z - \hat{\gamma}_1^T \xi) \right\} = 0,
\]

where \((\hat{\theta}_0, \hat{\alpha}_1)\) are Lasso estimators in model (53) as in our method, but \( \hat{\gamma}_1 \), different from \( \hat{\gamma}_1 \), is a Lasso estimator associated with the loss \( \bar{E}\{\expit(\hat{\theta}_0 Z + \hat{\alpha}_1^T \xi)(Z - \gamma^T \xi)^2\} \), corresponding
to a model \(E(Z|X) = \gamma^T \xi\) instead of model (54) in our method. Confidence intervals based on \(\hat{\theta}_{DB1}\) or \(\hat{\theta}_{DB2}\) would not be valid for \(\theta^*\) in partially linear model (8) if model (53) is misspecified, irrespective of whether model (54) is correctly specified.

**Example 13.** Return to Examples 4 and 9. For \(g(x; \alpha) = \psi_g(\alpha^T \xi)\) and \(f(x; \gamma) = \psi_f(\gamma^T \xi)\), models (11) and (12) can be stated as

\[
\begin{align*}
E(Y|Z = 1, X) &= \psi_g(\alpha^T \xi), \\
P(Z = 1|X) &= \psi_f(\gamma^T \xi).
\end{align*}
\]

(56)

(57)

For estimating function \(\tau\) in (10) and any estimators \((\hat{\alpha}, \hat{\gamma})\), \(\hat{\theta} = \hat{\theta}(\hat{\alpha}, \hat{\gamma})\) is of closed form

\[
\hat{\theta}(\hat{\alpha}, \hat{\gamma}) = \hat{E} \left[ \frac{Z Y}{\psi_f(\hat{\gamma}^T \xi)} - \left\{ \frac{Z}{\psi_f(\hat{\gamma}^T \xi)} - 1 \right\} \psi_g(\hat{\alpha}^T \xi) \right].
\]

(58)

For initial estimation, let \(\hat{\alpha}_1\) be a Lasso quasi-likelihood estimator in model (56), \(\hat{\gamma}_1\) be that in model (57), and \(\hat{\theta}_1 = \hat{\theta}(\hat{\alpha}_1, \hat{\gamma}_1)\). For second-step estimation, the regularized calibrated estimator \(\hat{\gamma}_2\) is defined with a Lasso penalty and the loss function

\[
L_2(\gamma; \hat{\alpha}_1) = \hat{E} \{ \ell_2(U; \hat{\alpha}_1, \gamma) \} = \hat{E} \left[ \psi_g(\hat{\alpha}_1^T \xi) \left\{ -Z \Psi_f(\gamma^T \xi) + \gamma^T \xi \right\} \right],
\]

and \(\hat{\alpha}_2\) is defined with a Lasso penalty and the loss function

\[
L_1(\alpha; \hat{\gamma}_2) = \hat{E} \{ \ell_1(U; \alpha, \hat{\gamma}_2) \} = \hat{E} \left[ \frac{\psi_f(\hat{\gamma}_2^T \xi)}{\psi_f(\hat{\gamma}_2^T \xi)} Z \left\{ -Y \alpha^T \xi + \Psi_g(\alpha^T \xi) \right\} \right],
\]

where \(\Psi_f(u) = \int_0^u \psi_f^{-1}(t) dt\), \(\Psi_g(u) = \int_0^u \psi_g(t) dt\), and \(\ell_1\) and \(\ell_2\) are determined from (33), with \((\partial \tau / \partial \alpha, \partial \tau / \partial \gamma)\) in (31)–(32). The resulting estimator of \(\theta\) is then \(\hat{\theta}_2 = \hat{\theta}(\hat{\alpha}_2, \hat{\gamma}_2)\).

By Proposition 3, valid confidence intervals based on \(\hat{\theta}_2\) can be obtained for \(\theta^* = E(Y)\) if either model (56) or (57) is correctly specified. Hence our work extends Tan [37], where doubly robust confidence intervals are obtained for \(\theta^*\) only with linear outcome model (56). With a nonlinear outcome model, valid confidence intervals are obtained in Tan [37], depending on propensity score model (57) being correctly specified.

5. Simulation studies

Consider the settings of partially linear, log-linear, and logistic models (Examples 10–12). Assume that the covariate of interest \(Z\) is binary (for example a treatment variable), and hence the coefficient \(\theta^*\) represents some homogeneous treatment effect. The link function for \(Z\) given \(X\) is taken to be logistic: \(\psi_f = \expit(\cdot)\).

We investigate the performance of our two-step estimator \(\hat{\theta}_2\), compared with the debiased Lasso estimator \(\hat{\theta}_{DB}\) and the initial estimator \(\hat{\theta}_1\) using regularized likelihood (or quasi-likelihood) estimation, as described in Section 4. For all point estimators, the Lasso tuning parameters are selected via 5-fold cross validation. Wald confidence intervals based on \(\hat{\theta}_2\) are obtained by Proposition 3. For comparison, confidence intervals based on \(\hat{\theta}_1\) are computed in
a similar manner, with \((\hat{\alpha}_1, \hat{\gamma}_1, \hat{\theta}_1)\) in place of \((\hat{\alpha}_2, \hat{\gamma}_2, \hat{\theta}_2)\). Wald confidence intervals based on \(\hat{\theta}_{DB}\) are computed using a robust variance estimator, which, for linear modeling, is defined as \(\hat{V}_{BG}\) in Section 4. See the Supplement for further implementation details.

Table 1. Summary of results for partially linear modeling \((n = 400, p = 800)\)

<table>
<thead>
<tr>
<th></th>
<th>(C1) Cor Cor</th>
<th>(C2) Cor Mis</th>
<th>(C3) Mis Cor</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{\theta}_{DB})</td>
<td>0.006</td>
<td>0.013</td>
<td>0.283</td>
</tr>
<tr>
<td>(\hat{\theta}_1)</td>
<td>0.006</td>
<td>0.012</td>
<td>0.082</td>
</tr>
<tr>
<td>(\hat{\theta}_2)</td>
<td>0.007</td>
<td>0.012</td>
<td>0.004</td>
</tr>
<tr>
<td>(\sqrt{\text{Var}})</td>
<td>0.058</td>
<td>0.057</td>
<td>0.322</td>
</tr>
<tr>
<td>(\sqrt{\text{Evar}})</td>
<td>0.057</td>
<td>0.058</td>
<td>0.334</td>
</tr>
<tr>
<td>\text{Cov95}</td>
<td>0.922</td>
<td>0.921</td>
<td>0.872</td>
</tr>
</tbody>
</table>

Note: Bias and Var are the Monte Carlo bias and variance of the point estimator, EVar is the mean of the variance estimator, and Cov95 is the coverage proportion of nominal 95% confidence intervals, based on 1000 repeated simulations.

Figure 1: QQ plots of \(t\)-statistics for partially linear modeling \((n = 400, p = 800)\)

5.1. Partially linear modeling

Consider the following data-generating configurations for \((Z, X, Y)\), where \(\theta^* = 3\).

(C1) Generate \(Z\) as Bernoulli with \(P(Z = 1) = 0.5\) and \(X\) given \(Z = 0\) or \(1\) as multivariate normal with means \(\mu_0 \neq \mu_1\) and variance matrices \(\Sigma_0 = \Sigma_1\), such that \(Z\) given \(X\) is Bernoulli with \(P(Z = 1|X) = \text{expit}(-0.4297 - 0.25X_1 + 0.5X_2 + 0.75X_3 + X_4 + 1.25X_5)\). Then \(Y\) given \((Z, X)\) is generated as normal with variance 0.5 and mean \(E(Y|Z, X) = \theta^* Z + 0.25X_1 + 1.5X_2 + 1.75X_3 + 5X_4\).

(C2) Generate \(Z\) as Bernoulli with \(P(Z = 1) = 0.5\) and \(X\) given \(Z = 0\) or \(1\) as multivariate normal with means \(\mu_0 \neq \mu_1\) and variance matrices \(\Sigma_0 \neq \Sigma_1\), such that \(Z\) given \(X\) is Bernoulli with \(P(Z = 1|X) = \text{expit}(-0.4687 + \frac{p}{2}\ln 2 - 0.25X_1 + 0.5X_2 + 0.75X_3 + X_4 + 0.5X^2)\). Then \(Y\) given \((Z, X)\) is generated as in (C1).

(C3) Generate \(Z\) given \(X\) as in (C1) and then \(Y\) given \((Z, X)\) as normal with variance 0.5 and mean \(E(Y|Z, X) = \theta^* Z + \text{expit}(0.5X_1 + X_2) + 4(X_3 - 0.75) + 2(X_4 - 1)^2\).

See the Supplement for details of \((\mu_0, \mu_1)\) and \((\Sigma_0, \Sigma_1)\) and the derivation of \(P(Z = 1|X)\) stated above, related to Fisher’s discrimination analysis.
Consider models (44) for $E(Y|Z,X)$ and (45) for $P(Z = 1|X)$, with the regressor vector $\xi = (1,X^T)$. Then the two models are both correctly specified in (C1), model (44) is correctly specified but model (45) is misspecified in (C2), and model (44) is misspecified and model (45) is correctly specified in (C3). Hence the simulation settings (C1), (C2), and (C3) are labeled as “Cor Cor”, “Cor Mis”, and “Mis Cor” respectively.

For $n = 400$ and $p = 800$, Table 1 summarizes the results for estimation of $\theta^*$ and Figure 1 shows the QQ plots of $t$-statistics. Additional results with $p = 100$ or 200 are included in the Supplement. In settings (C1) and (C2) with model (44) correctly specified for $E(Y|Z,X)$, the three methods using $\hat{\theta}_{DB}$, $\hat{\theta}_1$, and $\hat{\theta}_2$ perform similarly to each other. In theory, all the methods in such settings deliver valid confidence intervals. In setting (C3) with model (44) misspecified for $E(Y|Z,X)$ but model (45) correctly specified for $P(Z = 1|X)$, there are important differences between the three methods. The debiased Lasso estimator $\hat{\theta}_{DB}$ becomes inconsistent for $\theta^*$, as seen from a large bias and poor coverage proportion. The initial estimator $\hat{\theta}_1$ is, in theory, consistent but does not yield valid confidence intervals. Our calibrated estimator $\hat{\theta}_2$ shows the best performance, with a small bias and close to 95% coverage. The improvement of $\hat{\theta}_2$ over $\hat{\theta}_{DB}$ and $\hat{\theta}_1$ is also confirmed in Figure 1, where the QQ plot of $t$-statistics from $\hat{\theta}_2$ is much better aligned with standard normal.

### 5.2. Partially log-linear modeling

Consider the following data-generating configurations for $(Z,X,Y)$, where $\theta^* = 2$.

(C4) Generate $(Z,X)$ as in (C1) in Section 5.1 and then $Y$ given $(Z,X)$ as Poisson with mean $\exp(\theta^* Z + 0.1X_1 + 0.25X_2 + 0.5X_3 + 0.75X_4)$.

(C5) Generate $(Z,X)$ as in (C2) in Section 5.1 and then $Y$ as in (C4).

(C6) Generate $(Z,X)$ as in (C1) in Section 5.1 and then $Y$ given $(Z,X)$ as Poisson with mean $\exp(\theta^* Z + X_1 + 0.1X_2^2 + 0.2X_3^2)$.

Consider models (48) for $E(Y|Z,X)$ and (49) for $P(Z = 1|X)$, with the regressor vector $\xi = (1,X^T)$. Then the two models are both correctly specified in (C4), model (48) is correctly specified but model (49) is misspecified in (C5), and model (48) is misspecified and model (49) is correctly specified in (C6).

For $n = 600$ and $p = 800$, Table 2 summarizes the results for estimation of $\theta^*$ and Figure 2 shows the QQ plots of $t$-statistics. Additional results with $p = 100$ or 200 are included in the Supplement. The three methods perform similarly to each other in setting (C4). However, unlike in Section 5.1, our calibrated method achieves the best performance in both settings (C5) and (C6), with a smaller bias, closer to 95% coverage, and better aligned $t$-statistics with standard normal than the other methods.

### Table 2. Summary of results for partially log-linear modeling ($n = 600, p = 800$)

<table>
<thead>
<tr>
<th></th>
<th>(C4) Cor Cor</th>
<th>(C5) Cor Mis</th>
<th>(C6) Mis Cor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\theta}_{DB}$</td>
<td>$\hat{\theta}_1$</td>
<td>$\hat{\theta}_2$</td>
</tr>
<tr>
<td>Bias</td>
<td>0.008</td>
<td>0.009</td>
<td>0.005</td>
</tr>
<tr>
<td>$\sqrt{Var}$</td>
<td>0.043</td>
<td>0.046</td>
<td>0.048</td>
</tr>
<tr>
<td>$\sqrt{EVar}$</td>
<td>0.043</td>
<td>0.048</td>
<td>0.045</td>
</tr>
<tr>
<td>Cov95</td>
<td>0.938</td>
<td>0.934</td>
<td>0.941</td>
</tr>
</tbody>
</table>

Note: See the footnote of Table 1.
5.3. Partially logistic modeling

The covariates \( X = (X_1, \ldots, X_p)^T \) are generated as multivariate normal with means 0 and 
\( \text{cov}(X_j, X_k) = 2^{-|j-k|} \) for \( 1 \leq j, k \leq p \). Then \((Z, Y)\) given \( X\) are generated jointly (rather than sequentially) such that the following configurations are obtained, where \( \theta^* = 2 \).

\begin{enumerate}
\item[(C7)] \( Z \) given \( Y = 0 \) and \( X \) is Bernoulli with 
\[ P(Z = 1|Y = 0, X) = \expit(0.25 - 0.125X_1 + 0.125X_2 + 0.25X_3 + 0.375X_4) \]
and \( Y \) given \((Z, X)\) is Bernoulli with 
\[ P(Y = 1|Z, X) = \expit(\theta^*Z - 0.125X_1 + 0.125X_2 + 0.25X_3 + 0.375X_4). \]
\item[(C8)] \( Z \) given \( Y = 0 \) and \( X \) is Bernoulli with 
\[ P(Z = 1|Y = 0, X) = \expit(\theta^*Z - 0.25 + 0.25X_1 + 0.8X_2^2 + \expit(X_3)) \]
and \( Y \) given \((Z, X)\) is the same as in (C7).
\item[(C9)] \( Z \) given \( Y = 0 \) and \( X \) is the same as in (C7) and \( Y \) given \((Z, X)\) is Bernoulli with 
\[ P(Y = 1|Z, X) = \expit(\theta^*Z - 0.25 + 0.25X_1 + 0.8X_2^2 + \expit(X_3)). \]
\end{enumerate}

See the Supplement for details of data generation, related to the odds ratio model in Chen [8]. Consider models (53) for \( E(Y|Z, X) \) and (54) for \( P(Z = 1|Y = 0, X) \), with the regressor vector \( \xi = (1, X^T)^T \). Then the two models are both correctly specified in (C7), model (53) is correctly specified but model (54) is misspecified in (C8), and model (53) is misspecified and model (54) is correctly specified in (C9).

For \( n = 600 \) and \( p = 800 \), Table 3 summarizes the results for estimation of \( \theta^* \) and Figure 3 shows the QQ plots of t-statistics. Additional results with \( p = 100 \) or 200 are included in the Supplement. While the three methods perform similarly to each other in settings (C7) and (C8), our calibrated method achieves substantially better performance in setting (C9) than the other methods, similarly as in Section 5.1.

6. Conclusion

We develop regularized calibrated estimation as a general method for obtaining doubly robust confidence intervals in high-dimensional settings, provided a doubly robust estimating function is available. While various applications of the method can be pursued, there are interesting
topics which warrant further investigation. Alternatively to the two-step algorithm, it is of interest to study the generalized Dantzig selector mentioned in Section 3, including development of practical algorithms and theoretical analysis without a convex loss. This approach has a potential benefit in producing valid confidence intervals centered about the target value (i.e., the limit value of the estimator used) even if both working models are misspecified, similarly as discussed in Tan [37], Remark 9. Moreover, it can be helpful to exploit sample splitting and cross fitting for our method and study whether both rate and model double robustness can be achieved, beyond the setting of bilinear influence functions in Smucler et al. [32].

Acknowledgements

The authors would like to thank a referee for constructive comments leading to improvement of the paper. This work was supported in part by PCORI Grant ME-1511-32740.

Supplement Material

Supplementary Material available at Bernoulli online includes technical details and additional material for simulation studies.

References


