Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data

By Z. TAN

Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, New Jersey 08854, U.S.A.
ztan@stat.rutgers.edu

Summary

Propensity scores are widely used with inverse probability weighting to estimate treatment effects in observational studies. We study calibrated estimation as an alternative to maximum likelihood estimation for fitting logistic propensity score models. We show that, with possible model misspecification, minimizing the expected calibration loss underlying the calibrated estimators involves reducing both the expected likelihood loss and a measure of relative errors between the limiting and true propensity scores, which governs the mean squared errors of inverse probability weighted estimators. Furthermore, we derive a regularized calibrated estimator by minimizing the calibration loss with a lasso penalty. We develop a Fisher scoring descent algorithm for computing the proposed estimator and provide a high-dimensional analysis of the resulting inverse probability weighted estimators, leveraging the control of relative errors of propensity scores for calibrated estimation. We present a simulation study and an empirical application to demonstrate the advantages of the proposed methods over maximum likelihood and its regularization. The methods are implemented in the R package RCAL.

Some key words: Calibrated estimation; Causal inference; Fisher scoring; Inverse probability weighting; Lasso penalty; Model misspecification; Propensity score; Regularized M-estimation.

1. Introduction

Inverse probability weighted estimation is one of the main methods of using propensity scores to estimate treatment effects in observational studies. In practice, propensity scores are unknown and typically estimated from observed data by fitting logistic regression using maximum likelihood. However, inverse probability weighted estimation may suffer from large weights, caused by propensity scores being estimated close to 0 for a few treated subjects or close to 1 for a few untreated subjects. It has been argued that such methods may perform poorly even when the propensity score model appears to be nearly correct (e.g., Kang & Schafer, 2007).

As an alternative to maximum likelihood estimation, calibrated estimation has been proposed for fitting propensity score models. The basic idea is to use a system of estimating equations such that the weighted averages of the covariates in the treated subsample are equal to the simple averages in the overall sample. Subsequently, the fitted propensity scores can be used as usual in inverse probability weighted estimators or extensions. Such ideas and similar methods have been studied, and sometimes independently (re)derived, in various contexts of causal inference, missing-data problems and survey sampling (e.g., Folsom, 1991; Tan, 2010; Graham et al., 2012; Hainmueller, 2012; Imai & Ratkovic, 2014; Kim & Haziza, 2014; Vermeulen & Vansteelandt, 2015; Chan et al., 2016; Yiu & Su, 2018).
This work aims to address two questions. Previously, calibrated estimation was studied either through numerical experiments or by establishing properties such as local efficiency and double robustness. The first question is whether, in the presence of possible model misspecification, any advantage can be formally established for calibrated estimation over maximum likelihood estimation when fitting propensity score models for inverse probability weighting, regardless of outcome regression models. In addition, calibrated estimation was previously analyzed with the number of covariates \( p \) either fixed or growing slowly under strong enough smoothness conditions (e.g., Chan et al., 2016). The second question is how to extend and analyze calibrated estimation when the number of covariates is close to or greater than the sample size.

We develop theory and methods to address the foregoing questions, with a logistic propensity score model. First, we establish a new relationship between the loss functions underlying calibrated and maximum likelihood estimation. From this result, we show that minimizing the expected calibration loss involves reducing both the expected likelihood loss and, separately, a measure of relative errors of the target, or limiting, propensity score, which then governs the mean squared errors of inverse probability weighted estimators. Such control of relative errors of propensity scores cannot be achieved by minimizing the expected likelihood loss alone.

Second, we derive a regularized calibrated estimator by minimizing the calibration loss with a lasso penalty (Tibshirani, 1996), in which calibration equations are relaxed to box constraints. We develop a novel algorithm for computing the proposed estimator, exploiting quadratic approximation (Friedman et al., 2010), Fisher scoring (McCullagh & Nelder, 1989) and the majorization-minimization technique (Wu & Lange, 2010). We also perform a high-dimensional analysis of the regularized calibrated estimator and the resulting inverse probability weighted estimators of population means, allowing possible model misspecification. Compared with previous results based on regularized maximum likelihood estimation (e.g., Belloni et al., 2014), our results are established under weaker conditions, by carefully leveraging the control of relative errors of propensity scores in calibrated estimation.

In a companion paper (Tan, 2019a), we build on the work on propensity score estimation and develop regularized calibrated estimation using both propensity score and outcome regression models. Both the computational algorithm and the high-dimensional analysis developed here are directly used in Tan (2019a). In addition, the theoretical comparison presented here between calibrated and maximum likelihood estimation of propensity scores is of independent interest.

2. BACKGROUND: CAUSAL INFERENCE

Suppose that the observed data consist of independent and identically distributed observations \( \{(Y_i, T_i, X_i) : i = 1, \ldots, n\} \) of \((Y, T, X)\), where \( Y \) is an outcome variable, \( T \) is a treatment variable taking value 0 or 1, and \( X = (X_1, \ldots, X_p) \) is a vector of measured covariates. In the potential outcomes framework for causal inference (Neyman, 1923; Rubin, 1974), let \((Y^0, Y^1)\) be potential outcomes that would be observed under treatments 0 and 1, respectively. For consistency, assume that \( Y \) is either \( Y^0 \) if \( T = 0 \) or \( Y^1 \) if \( T = 1 \), that is, \( Y = (1 - T) Y^0 + T Y^1 \). There are two causal parameters commonly of interest: the average treatment effect, defined as \( E(Y^1 - Y^0) = \mu^1 - \mu^0 \) with \( \mu^t = E(Y^t) \) \((t = 0, 1)\), and the average treatment effect on the treated, defined as \( E(Y^1 - Y^0 | T = 1) = v^1 - v^0 \) with \( v^t = E(Y^t | T = 1) \) \((t = 0, 1)\). For concreteness, we mainly discuss estimation of \((\mu^0, \mu^1)\) until \( \S \), where we discuss \((v^0, v^1)\).

For identification of \((\mu^0, \mu^1)\), we make the following two assumptions throughout: (i) \( T \perp Y^0 | X \) and \( T \perp Y^1 | X \) (Rubin, 1976); (ii) \( 0 < \text{pr}(T = 1 | X = x) < 1 \) for all \( x \). There are two broad approaches to estimating \((\mu^0, \mu^1)\), depending on additional modelling assumptions on the outcome regression function \( m^*(t, X) = E(Y | T = t, X) \) or the propensity score (Rosenbaum & Rubin, 1983) \( \pi^*(X) = \text{pr}(T = 1 | X) \); see Tan (2007) for further discussion.
Calibrated estimation of propensity scores

For the propensity score approach, consider a regression model of the form

$$\text{pr}(T = 1 \mid X) = \pi(X; \gamma) = \Pi[\gamma^T f(X)],$$ \hfill (1)

where $\Pi(\cdot)$ is an inverse link function, $f(x)$ is a vector of known functions, and $\gamma$ is a vector of unknown parameters. Typically, logistic regression is used with $\pi(X; \gamma) = [1 + \exp\{-\gamma^T f(X)\}]^{-1}$. Let $\hat{\gamma}_{ML}$ be the maximum likelihood estimator of $\gamma$, which for logistic regression minimizes the average negative loglikelihood

$$\ell_{ML}(\gamma) = \bar{E}(\log[1 + \exp\{-\gamma^T f(X)\}] - T \gamma^T f(X)).$$ \hfill (2)

Throughout, $\bar{E}(\cdot)$ denotes the sample average. The fitted propensity score, $\hat{\pi}_{ML}(X) = \pi(X; \hat{\gamma}_{ML})$, can be used in various ways to estimate $(\mu^0, \mu^1)$. In particular, we focus on inverse probability weighting, which is central to the semiparametric theory of estimation in missing-data problems. Two commonly used inverse probability weighted estimators for $\mu^1$ are

$$\hat{\mu}_{IPW}^1(\hat{\pi}_{ML}) = \bar{E}\{TY\hat{\pi}_{ML}^{-1}(X)\}, \quad \hat{\mu}_{IPW}^1(\hat{\pi}_{ML}) = \hat{\mu}_{IPW} / \bar{E}\{T\hat{\pi}_{ML}^{-1}(X)\}.$$

Similarly, two inverse probability weighted estimators for $\mu^0$ are defined by replacing $T$ with $1 - T$ and $\hat{\pi}_{ML}(X)$ with $1 - \hat{\pi}_{ML}(X)$. If model (1) is correctly specified, then the above estimators are consistent under standard regularity conditions as $n \to \infty$ with the dimension of $\gamma$ fixed.

3. Calibrated estimation

3.1. Preliminaries

For concreteness, we assume that propensity score model (1) is a logistic regression model:

$$\text{pr}(T = 1 \mid X) = \pi(X; \gamma) = [1 + \exp\{-\gamma^T f(X)\}]^{-1},$$ \hfill (3)

where $f(x) = \{f_1(x), \ldots, f_p(x)\}^T$ is a vector of known functions including a constant and $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_p)^T$ is a vector of unknown parameters. Let $\hat{\gamma}_{CAL}^1$ be an estimator of $\gamma$ that solves

$$\bar{E}[(T\pi^{-1}(X; \gamma) - 1)f(X)] = 0.$$ \hfill (4)

The fitted propensity score is $\hat{\pi}_{CAL}^1(X) = \pi(X; \hat{\gamma}_{CAL}^1)$. Then $\mu^1$ can be estimated by $\hat{\mu}_{IPW}^1(\hat{\pi}_{CAL}^1)$ or equivalently $\hat{\mu}_{IPW}^1(\hat{\pi}_{CAL}^1)$, with $\hat{\pi}_{ML}(X)$ replaced by $\hat{\pi}_{CAL}^1(X)$. These two estimators are identical because $\bar{E}\{T/\hat{\pi}_{CAL}^1(X)\} = 1$ by (4) with a constant included in $f(X)$.

Similarly, let $\hat{\gamma}_{CAL}^0$ be an estimator of $\gamma$ solving

$$\bar{E}[(1 - T)(1 - \pi(X; \gamma))^{-1} - 1]f(X)] = 0,$$ \hfill (5)

and let $\hat{\pi}_{CAL}^0(X) = \pi(X; \hat{\gamma}_{CAL}^0)$. Then $\mu^0$ can be estimated by $\hat{\mu}_{IPW}^0(\hat{\pi}_{CAL}^0)$ or equivalently $\hat{\mu}_{IPW}^0(\hat{\pi}_{CAL}^0)$, with $\hat{\pi}_{ML}(X)$ replaced by $\hat{\pi}_{CAL}^0(X)$, where the equivalence of these two estimators follows from the fact that $\bar{E}[(1 - T)/[1 - \hat{\pi}_{CAL}^0(X)]] = 1$ by (5) with a constant included in $f(X)$. See §7 for remarks on the unusual fact that two different sets of fitted propensity scores, $\hat{\pi}_{CAL}^1(X)$ and $\hat{\pi}_{CAL}^0(X)$, are used for estimating $\mu^1$ and $\mu^0$, respectively.

Following the survey literature (Folsom, 1991), (4) is referred to as the calibration equation for the treated (i.e., treatment 1), because the inverse probability weighted average of $f(X_i)$ over the treated group $\{i: T_i = 1, \ i = 1, \ldots, n\}$ is calibrated to the average of $f(X_i)$ over the entire sample including the treated and untreated. Similarly, (5) is referred to as the calibration equation...
for the untreated (i.e., treatment 0). The resulting estimators \( \hat{\gamma}^1_{\text{CAL}} \) and \( \hat{\gamma}^0_{\text{CAL}} \) are called calibrated estimators of \( \gamma \), in contrast to the maximum likelihood estimator \( \hat{\gamma}_{\text{ML}} \). The fitted values \( \hat{\pi}^1_{\text{CAL}}(X) \) and \( \hat{\pi}^0_{\text{CAL}}(X) \) are also called calibrated propensity scores.

Calibration equations have been proposed in various contexts. The idea seems to have been explored first by Folsom (1991). The equations in (4) can be deduced from Tan (2010, § 4.4) with a degenerate propensity score model. Moreover, (4) has been used by Graham et al. (2012), Kim & Haziza (2014) and Vermeulen & Vansteelandt (2015) to develop locally efficient and doubly robust estimators, and was obtained by Chan et al. (2016) in a dual formulation. As shown in § 7, (5) leads to the same estimator as in entropy balancing (Hainmueller, 2012).

3.2. Comparison between calibrated and maximum likelihood estimation

We compare the loss functions for calibrated and maximum likelihood estimation of propensity scores and the associated inverse probability weighted estimators in the limiting case. The calibrated estimator \( \hat{\gamma}^1_{\text{CAL}} \) can be equivalently defined as a minimizer of the loss function

\[
\ell_{\text{CAL}}(\gamma) = E[T \exp[-\gamma^T f(X)]] + (1 - T)\gamma^T f(X)].
\]

In fact, setting the gradient of \( \ell_{\text{CAL}}(\gamma) \) to zero is easily shown to yield the calibration equation (4) with logistic \( \pi(X; \gamma) \). Moreover, \( \ell_{\text{CAL}}(\gamma) \) is convex in \( \gamma \) and is strictly convex and bounded from below under a certain nonseparation condition, as discussed in the Supplementary Material. Previously, the loss function \( \ell_{\text{CAL}} \) and similar ones were mainly used as a computational device (Tan, 2010, § 4.4; Vermeulen & Vansteelandt, 2015).

First, we establish an interesting relationship between the likelihood and calibration loss functions \( \ell_{\text{ML}}(\gamma) \) and \( \ell_{\text{CAL}}(\gamma) \). To allow for misspecification of model (3), we write \( \ell_{\text{ML}}(\gamma) = \kappa_{\text{ML}}(\gamma^T f) \) and \( \ell_{\text{CAL}}(\gamma) = \kappa_{\text{CAL}}(\gamma^T f) \), where for a function \( g(x) \),

\[
\kappa_{\text{ML}}(g) = E[\log(1 + \exp(g(X))] - T g(X)),
\]

\[
\kappa_{\text{CAL}}(g) = E[T \exp[-g(X)] + (1 - T)g(X)].
\]

Then \( \kappa_{\text{ML}}(g^*) \) and \( \kappa_{\text{CAL}}(g^*) \) are well-defined for the true log odds ratio \( g^*(x) = \log[\pi^*(x)/\{1 - \pi^*(x)\}] \), even when model (3) is misspecified, i.e., when \( g^*(x) \) is not of the form \( \gamma^T f(x) \). It can easily be shown that both \( \kappa_{\text{ML}}(g) \) and \( \kappa_{\text{CAL}}(g) \) are convex in \( g \). For two functions \( g(x) \) and \( g'(x) \), consider the Bregman divergences associated with \( \kappa_{\text{ML}} \) and \( \kappa_{\text{CAL}} \):

\[
D_{\text{ML}}(g, g') = \kappa_{\text{ML}}(g) - \kappa_{\text{ML}}(g') - \langle \nabla \kappa_{\text{ML}}(g') \rangle, g - g',
\]

\[
D_{\text{CAL}}(g, g') = \kappa_{\text{CAL}}(g) - \kappa_{\text{CAL}}(g') - \langle \nabla \kappa_{\text{CAL}}(g') \rangle, g - g',
\]

where \( g \) is identified as a vector \( (g_1, \ldots, g_n) \) with \( g_i = g(X_i) \), \( \nabla \kappa_{\text{CAL}}(g') \), \( g - g' \) \( = n^{-1} \sum_{i=1}^n \{(\partial g_j)/\partial g'_{j}\}{T_i \exp(-g'_j) + (1 - T_i)g'_j} \times (g_i - g'_i) \), and \( \nabla \kappa_{\text{CAL}}(g') \), \( g - g' \) is similarly defined. For two probabilities \( \rho \in (0, 1) \) and \( \rho' \in (0, 1) \), the Kullback–Liebler divergence is \( L(\rho, \rho') = \rho' \log(\rho'/\rho) + (1 - \rho') \log((1 - \rho')/(1 - \rho)) \geq 0 \). In addition, let \( K(\rho, \rho') = \rho'/\rho - 1 - \log(\rho'/\rho) \geq 0 \), which is strictly convex in \( \rho'/\rho \) with a minimum of 0 at \( \rho'/\rho = 1 \).

Proposition 1. (i) For any functions \( g(x) \) and \( g'(x) \) and the corresponding functions \( \pi(x) = [1 + \exp(-g(x))]^{-1} \) and \( \pi'(x) = [1 + \exp(-g'(x))]^{-1} \), we have that

\[
D_{\text{ML}}(g, g') = \tilde{E}[L[\pi(X), \pi'(X)]],
\]

\[
D_{\text{CAL}}(g, g') = \tilde{E}\left(\frac{T}{\pi'(X)}[K[\pi(X), \pi'(X)] + L[\pi(X), \pi'(X)]]\right).
\]
Calibrated estimation of propensity scores

(ii) As a result, we have that for any fixed value $\gamma$,

$$E[\ell_{ML}(\gamma) - \kappa_{ML}(g^*)] = E[L(\pi(X; \gamma), \pi^*(X))],$$  \hspace{1cm} (9)

$$E[\ell_{CAL}(\gamma) - \kappa_{CAL}(g^*)] = E[K(\pi(X; \gamma), \pi^*(X)) + L(\pi(X; \gamma), \pi^*(X))].$$ \hspace{1cm} (10)

To examine the implications of Proposition 1, we first briefly describe some results from the classical theory of estimation in misspecified models (White, 1982; Manski, 1988). Under standard regularity conditions, as $n \to \infty$ while $p$ is fixed, the maximum likelihood estimator $\hat{\gamma}_{ML}$ can be shown to converge in probability to a target, or limiting, value $\gamma_{ML}$, which is defined as a minimizer of the expected loss $E[\ell_{ML}(\gamma)]$ or, equivalently, the Kullback–Liebler divergence (9). Similarly, $\hat{\gamma}_{CAL}^1$ can be shown to converge in probability to a target value $\gamma_{CAL}^1$, which is defined as a minimizer of the expected loss $E[\ell_{CAL}(\gamma)]$ or, equivalently, the calibration divergence (10). If model (3) is correctly specified, then both $\hat{\gamma}_{ML}$ and $\hat{\gamma}_{CAL}^1$ coincide with $\gamma^*$ such that $\pi(\cdot; \gamma^*) = \pi^*(\cdot)$. However, if model (3) is misspecified, then $\hat{\gamma}_{ML}$ and $\hat{\gamma}_{CAL}^1$ are in general different, determined by the minimization of the corresponding divergence, (9) or (10).

Next, we demonstrate how the mean squared error of an inverse probability weighted estimator, $\hat{\mu}_{IPW}(\gamma) = \bar{E}(TY/\pi(X; \gamma))$, can be bounded from above by the likelihood and calibration divergences (9) and (10). Consider the mean squared relative error

$$\text{MSRE}(\gamma) = E[Q(\pi(X; \gamma), \pi^*(X))] = E \left[ \left( \frac{\pi^*(X)}{\pi(X; \gamma)} - 1 \right)^2 \right],$$

where $Q(\rho, \rho') = (\rho' / \rho - 1)^2$ for two probabilities $\rho \in (0, 1)$ and $\rho' \in (0, 1)$. This measure of relative errors of propensity scores can be seen to directly govern the mean squared error of the estimator $\hat{\mu}_{IPW}(\gamma)$, independently of the outcome variable $Y$.

**Proposition 2.** Suppose that $E[(Y^1 \mid X) \leq c$ and $\pi^*(X) \geq \delta$ almost surely for some constants $c > 0$ and $\delta \in (0, 1)$.

(i) For any fixed value $\gamma$,

$$E[(\hat{\mu}_{IPW}^1(\gamma) - \mu^1)^2] \leq \frac{2c}{n\delta} + c \left( 1 + \frac{2}{n\delta} \right) \text{MSRE}(\gamma).$$ \hspace{1cm} (11)

(ii) If $\pi(\cdot; \gamma) \geq a \pi^*(\cdot)$ almost surely for some constant $a \in (0, 1/2]$, then

$$\text{MSRE}(\gamma) \leq \frac{5}{3a} E[K(\pi(\cdot; \gamma), \pi^*(\cdot))] \leq \frac{5}{3a} E[\ell_{CAL}(\gamma) - \kappa_{CAL}(g^*)].$$ \hspace{1cm} (12)

The factor $5/(3a)$ in general cannot be improved up to a constant, independent of $a$.

(iii) If $\pi(\cdot; \gamma) \geq b$ almost surely for some constant $b \in (0, 1)$, then

$$\text{MSRE}(\gamma) \leq \frac{1}{2b^2} E[L(\pi(\cdot; \gamma), \pi^*(\cdot))] = \frac{1}{2b^2} E[\ell_{ML}(\gamma) - \kappa_{ML}(g^*)].$$ \hspace{1cm} (13)

The factor $1/(2b^2)$ in general cannot be improved up to a divisor of order $\log(b^{-1})$.

Setting $\gamma$ to $\gamma_{CAL}^1$ in (11) and (12) shows that if $\pi(X; \gamma_{CAL}^1) \geq a \pi^*(X)$ almost surely, then

$$E[(\hat{\mu}_{IPW}^1(\gamma_{CAL}^1) - \mu^1)^2] \leq \frac{2c}{n\delta} + \frac{5c}{3a} \left( 1 + \frac{2}{n\delta} \right) \min_{\gamma} E[\ell_{CAL}(\gamma) - \kappa_{CAL}(g^*)].$$ \hspace{1cm} (14)
The minimum calibration divergence is inflated in (14) by a leading factor $5c/(3a)$, which remains bounded from above as long as $a$ is bounded away from zero, even when some limiting propensity scores $\pi(X; \tilde{\gamma}_{\text{CAL}})$ are close to zero. In contrast, setting $\gamma$ to the limiting value $\tilde{\gamma}_{\text{ML}}$ in (11) and (13) shows that if $\pi(X; \tilde{\gamma}_{\text{ML}}) \geq b$ almost surely, then

$$E[ (\hat{\mu}_{\text{IPW}}(\tilde{\gamma}_{\text{ML}}) - \mu)^2 ] \leq \frac{2c}{n\delta} + \frac{c}{2b^2} \left( 1 + \frac{2}{n\delta} \right) \min_{\gamma} E[ \ell_{\text{ML}}(\gamma) - \kappa_{\text{ML}}(g^*) ]. \quad (15)$$

The minimum likelihood divergence is inflated in (15) by the leading factor $c/(2b^2)$, which may be arbitrarily large when some limiting propensity scores $\pi(X; \tilde{\gamma}_{\text{ML}})$ are close to zero. In general, which bound, (14) or (15), is smaller depends on both the divergences and the inflation factors, and the comparison between the bounds or the mean squared errors may vary in different data-generation settings. Nevertheless, the preceding discussion suggests that the mean squared error of $\hat{\mu}_{\text{IPW}}(\tilde{\gamma}_{\text{CAL}})$ can be more strongly controlled by the minimization of the calibration divergence than that of $\hat{\mu}_{\text{IPW}}(\tilde{\gamma}_{\text{ML}})$ by the minimization of the likelihood divergence, especially when some of the limiting propensity scores are close to zero. This result is complemented to the statistical theory in § 4.3 on how inverse probability weighted estimators with fitted propensity scores deviate from $\hat{\mu}_{\text{IPW}}(\tilde{\gamma}_{\text{CAL}})$ with limiting propensity scores. See (29) and the related discussion thereafter.

Technically, the difference between (12) and (13) can be attributed to the function $K[\pi(X; \gamma), \pi^*(X)]$ in the calibration divergence (10). On the one hand, (12) is deduced from the following lemma, proved in the Supplementary Material. The calibration divergence controls the mean squared relative error $E[ (\pi^*(X)/\pi(X; \gamma) - 1)^2 ]$ effectively through $E[ K[\pi(X; \gamma), \pi^*(X)] ]$. On the other hand, (13) is based on the inequality $L(\rho, \rho') \geq 2(\rho - \rho')^2$ (Cover & Thomas, 1991, Lemma 12.6.1). The likelihood divergence mainly controls the mean squared absolute error $E[ (\pi(X; \gamma) - \pi^*(X))^2 ]$ rather than the mean squared relative error. Lemma 1 also shows that an upper bound on $\rho'/\rho$ alone does not in general constrain the ratio $Q(\rho, \rho')/L(\rho, \rho')$.

**Lemma 1.** For a constant $a \in (0, 1/2]$, if any two probabilities $\rho \in (0, 1)$ and $\rho' \in (0, 1)$ satisfy $\rho \geq a\rho'$, then

$$Q(\rho, \rho') \leq \frac{5}{3a} K(\rho, \rho').$$

By comparison, $\sup_{\rho \geq a\rho'}{Q(\rho, \rho')/L(\rho, \rho')} = \infty$ for any constant $a > 0$.

For illustration, consider a simple setting adapted from the simulation study in § 5. Let $X \sim N(0, 1)$ and $\pi^*(X) = 1 + \exp(X)^{-1}$. For $W = f_1(X) = \exp(X/2)$, the propensity score model is $\pi(X; \gamma) = \{ 1 + \exp(\gamma_0 + \gamma_1 W) \}^{-1}$, which is misspecified. Figure 1 shows the limiting propensity scores $\pi(\cdot; \tilde{\gamma}_{\text{ML}}), \pi(\cdot; \tilde{\gamma}_{\text{CAL}})$ and $\pi(\cdot; \tilde{\gamma}_{\text{BAL}})$ (Imai & Ratkovic, 2014), conditional on $n = 40$ design points $(W_1, \ldots, W_n)$, labelled as covariate, where $W_i = \exp(X_i/2)$ with $X_i$ being the $i/401$ quantile of $N(0, 1)$. The values $\tilde{\gamma}_{\text{ML}}, \tilde{\gamma}_{\text{CAL}}$ and $\tilde{\gamma}_{\text{BAL}}$ are computed by minimizing, respectively, $\ell_{\text{ML}}(\gamma), \ell_{\text{CAL}}(\gamma)$ and $\ell_{\text{BAL}}(\gamma)$ in (2), (6) and (36) with $T_i$ replaced by $\pi^*(X_i)$. If judged by absolute errors, that is, by $|\pi^*(\cdot) - \pi(\cdot; \tilde{\gamma}_{\text{ML}})|$ etc., the three propensity scores are comparable and reasonably capture the main trend of the true propensity scores. However, substantial differences emerge when the propensity scores are compared in terms of relative errors, that is, $|\pi^*(\cdot)/\pi(\cdot; \tilde{\gamma}_{\text{ML}}) - 1|$ etc. The calibrated propensity scores are the most accurate, the maximum likelihood propensity scores are the least accurate, and the balancing propensity scores are in between, especially in the right tail of $W$ where the true propensity scores are small. If a true propensity score 0.05 is estimated by, for example, 0.005, then the relative error
is large even though the absolute error appears small. As suggested by (11), it is relative rather than absolute errors that are relevant in evaluating propensity scores used for inverse probability weighting.

4. Regularized calibrated estimation

4.1. Background

We propose a regularized calibrated estimator of $\gamma$ in propensity score model (3). There are two motivations. First, regularization is needed in the situation where $\hat{\gamma}_{\text{CAL}}$ does not exist because the calibration loss (6) is unbounded from below; see Proposition S1 in the Supplementary Material. In our numerical study, nonconvergence is found for $\hat{\gamma}_{\text{CAL}}$ in 100% of repeated simulations with $n = 200$ and $p = 50$, as shown in the Supplementary Material. Second, regularization is also needed in high-dimensional settings where the dimension of $f(X)$ is close to or greater than the sample size.

The regularized calibrated estimator, denoted by $\hat{\gamma}_{\text{RCAL}}$, is defined by minimizing the calibration loss $\ell_{\text{CAL}}(\gamma)$ with a lasso penalty (Tibshirani, 1996),

$$
\ell_{\text{RCAL}}(\gamma) = \ell_{\text{CAL}}(\gamma) + \lambda \|\gamma_{1:p}\|_1,
$$

where $\gamma_{1:p} = (\gamma_1, \ldots, \gamma_p)^T$ excluding $\gamma_0$, $\|\cdot\|_1$ denotes the $L_1$-norm such that $\|\gamma_{1:p}\|_1 = \sum_{j=1}^{p} |\gamma_j|$, and $\lambda \geq 0$ is a tuning parameter. By the Karush–Kuhn–Tucker condition for minimization of (16), the fitted propensity score, $\hat{\pi}_{\text{RCAL}}(X) = \pi(X; \hat{\gamma}_{\text{RCAL}})$, satisfies

$$
\frac{1}{n} \sum_{i=1}^{n} \frac{T_i}{\hat{\pi}_{\text{RCAL}}(X_i)} = 1,
$$

$$
\frac{1}{n} \left| \sum_{i=1}^{n} \frac{T_i f_j(X_i)}{\hat{\pi}_{\text{RCAL}}(X_i)} - \sum_{i=1}^{n} f_j(X_i) \right| \leq \lambda \quad (j = 1, \ldots, p),
$$

where equality holds in (18) for any $j$ such that the $j$th estimate $(\hat{\gamma}_{\text{RCAL}})_j$ is nonzero. The inverse probability weights, $1/\hat{\pi}_{\text{RCAL}}(X_i)$ with $T_i = 1$, still sum to the sample size $n$ by (17), but the weighted average of each covariate $f_j(X_i)$ over the treated group may differ from the overall average of $f_j(X_i)$ by no more than $\lambda$. In other words, introducing the lasso penalty to calibrated estimation yields a relaxation of the equalities (4) to box constraints (18).
When model (3) is logistic regression, a lasso-penalized maximum likelihood estimator, is obtained by minimizing

\[ \ell_{\text{RML}}(\gamma) = \ell_{\text{ML}}(\gamma) + \lambda \| \gamma \|_1, \tag{19} \]

where \( \ell_{\text{ML}}(\gamma) \) is the average negative loglikelihood in (2). Lasso-penalized maximum likelihood estimation has been studied extensively for generalized linear models (e.g., Bühlmann & van de Geer, 2011). Such estimators have also been used in fitting propensity score models to estimate treatment effects (Belloni et al., 2014; Farrell, 2015). We build on and extend previous work to handle computation and analysis of \( \hat{\gamma}_{\text{RCAL}}^1 \) and inverse probability weighted estimation.

### 4.2. Computation

We present a Fisher scoring descent algorithm for computing the estimator \( \hat{\gamma}_{\text{RCAL}}^1 \), that is, minimizing \( \ell_{\text{RCAL}}(\gamma) \) in (16) for any fixed \( \lambda \). The basic idea of the algorithm is to iteratively form a quadratic approximation to the calibration loss \( \ell_{\text{CAL}}(\gamma) \) in (6) and solve a lasso-penalized weighted least squares problem, similarly to existing algorithms for lasso-penalized maximum likelihood-based logistic regression (e.g., Friedman et al., 2010). However, a suitable quadratic approximation is obtained here only after an additional step of replacing certain sample quantities with model expectations. This idea is known as Fisher scoring and was previously used to derive the iterative reweighted least squares method for fitting generalized linear models with noncanonical links, such as probit regression (McCullagh & Nelder, 1989).

The quadratic approximation directly from a Taylor expansion of \( \ell_{\text{CAL}}(\gamma) \) about current estimates, denoted by \( \tilde{\gamma} \), is

\[
\ell_{\text{CAL},Q1}(\gamma; \tilde{\gamma}) = \ell_{\text{CAL}}(\tilde{\gamma}) + \tilde{E}\left[ -T \exp\{-f^T(X)\tilde{\gamma}\} + 1 - T \right] f^T(X)(\gamma - \tilde{\gamma}) \\
+ \frac{1}{2}(\gamma - \tilde{\gamma})^T f^T(X) [T \exp\{-f^T(X)\tilde{\gamma}\}] f(X)(\gamma - \tilde{\gamma}). \tag{20}
\]

As suggested by the quadratic term, it is tempting to recast (20) as a weighted least squares objective function in the form \((1/2) \sum_{i=1}^n w_i (T_i - f^T(X)\tilde{\gamma})^2\) with some working response \( \tilde{T}_i \), free of \( \gamma \), and weight \( w_i = T_i \exp\{-f^T(X)\tilde{\gamma}\} \). Instead, by Fisher scoring, we replace \( T_i \exp\{-f^T(X)\tilde{\gamma}\} \) with its expectation \([1 + \exp(f^T(X)\tilde{\gamma})]^{-1}\) under (3) with parameter \( \tilde{\gamma} \), and obtain

\[
\ell_{\text{CAL},Q2}(\gamma; \tilde{\gamma}) = \ell_{\text{CAL}}(\tilde{\gamma}) + \tilde{E}\left[ -T \exp\{-f^T(X)\tilde{\gamma}\} + 1 - T \right] f^T(X)(\gamma - \tilde{\gamma}) \\
+ \frac{1}{2}(\gamma - \tilde{\gamma})^T f^T(X) [1 + \exp(f^T(X)\tilde{\gamma})]^{-1} f(X)(\gamma - \tilde{\gamma}). \tag{21}
\]

which is easily shown to be a weighted least squares objective function with covariate vector \( f(X_i) \) and with working response and weight

\[
\tilde{T}_i = f^T(X_i)\tilde{\gamma} + \frac{T_i - \pi(X_i; \tilde{\gamma})}{\pi(X_i; \tilde{\gamma})(1 - \pi(X_i; \tilde{\gamma}))}, \tag{22}
\]

\[
w_i = 1 - \pi(X_i; \tilde{\gamma}), \tag{23}
\]

respectively. By comparison, in the iterative reweighted least squares algorithm for fitting logistic regression by maximum likelihood, the working response is the same as (22), but the weight is \( \pi(X_i; \tilde{\gamma})(1 - \pi(X_i; \tilde{\gamma})) \). Therefore, observations are weighted more with \( \pi(X_i; \tilde{\gamma}) \) closer to 0 for maximum likelihood estimation, but with \( \pi(X_i; \tilde{\gamma}) \) closer to 0 for calibrated estimation by (23).
To reduce computational cost, we make use of the majorization-minimization technique (Wu & Lange, 2010), similarly to existing algorithms for logistic regression. In particular, a majorizing function of (21) at current estimates \( \hat{y} \) is, by the quadratic lower bound principle (Bohning & Lindsay, 1988), the quadratic function obtained by replacing the Hessian \( \hat{E}[f^T(X)|1-\pi(X; \hat{y})f(X)] \) with \( \hat{E}[f^T(X)f(X)] \) in (21). The resulting quadratic function of \( y \), denoted by \( \ell_{\text{CAL}, Q3}(y; \hat{y}) \), can be shown to be a weighted least squares objective function with working response and weight \( \tilde{T}_i = f^T(X_i)\hat{y} + \frac{T_i}{\pi(X_i; \hat{y})} - 1 \), \( w_i = 1 \), respectively. A complication of Fisher scoring, i.e., transition from (20) to (21) is that, unlike a direct majorization of the quadratic approximation from a Taylor expansion, the function \( \ell_{\text{CAL}, Q3}(y; \hat{y}) \) may not be a majorizing function of \( \ell_{\text{CAL}}(y) \), and hence minimization of \( \ell_{\text{CAL}, Q3}(y; \hat{y}) + \lambda \|y_{1:p}\|_1 \) may not lead to a decrease of the objective function \( \ell_{\text{RCAL}}(y) = \ell_{\text{CAL}}(y) + \lambda \|y_{1:p}\|_1 \) from the current value \( \ell_{\text{RCAL}}(\hat{y}) \), as would be achieved by the majorization-minimization technique. However, the descent property, when occasionally violated, can be restored by incorporating a backtracking line search, because the direction found by minimizing \( \ell_{\text{CAL}, Q3}(y; \hat{y}) + \lambda \|y_{1:p}\|_1 \) must be a descent direction for the objective function \( \ell_{\text{RCAL}}(y) \).

**Proposition 3.** Let \( \hat{y}^{(1)} \neq \hat{y} \) be a minimizer of \( \ell_{\text{CAL}, Q2}(y; \hat{y}) + \lambda \|y_{1:p}\|_1 \) or of \( \ell_{\text{CAL}, Q3}(y; \hat{y}) + \lambda \|y_{1:p}\|_1 \), and let \( \hat{y}^{(t)} = (1 - t)\hat{y} + t\hat{y}^{(1)} \) for \( 0 < t < 1 \). Then any subgradient of \( \ell_{\text{CAL}}(\hat{y}^{(t)}) + \lambda \|y_{1:p}\|_1 \) at \( t = 0 \) is negative.

From the preceding discussions we develop the following algorithm.

**Algorithm 1.** Fisher scoring descent algorithm for minimizing (16).

(i) Set an initial value \( \gamma^{(0)} \).

(ii) Repeat the following updates for \( k = 1, 2, \ldots \) until convergence to obtain \( \hat{\gamma}_{\text{CAL}}^1 \):

(iii1) Compute \( \gamma^{(k-1/2)} = \arg \min \gamma \ell_{\text{CAL}, Q2}(y; \gamma^{(k-1)}) + \lambda \|y_{1:p}\|_1 \) or \( \gamma^{(k-1/2)} = \arg \min \gamma \ell_{\text{CAL}, Q3}(y; \gamma^{(k-1)}) + \lambda \|y_{1:p}\|_1 \).

(iii2) If \( \ell_{\text{RCAL}}(\gamma^{(k-1/2)}) < \ell_{\text{RCAL}}(\gamma^{(k-1)}) \), then set \( \gamma^{(k)} = \gamma^{(k-1/2)} \); otherwise set \( \gamma^{(k)} = (1 - t)\gamma^{(k-1/2)} + t\gamma^{(k-1/2)} \) for some \( 0 < t < 1 \) through a backtracking line search, such that \( \ell_{\text{RCAL}}(\gamma^{(k)}) < \ell_{\text{RCAL}}(\gamma^{(k-1)}) \).

Various algorithms, for example coordinate descent as in Friedman et al. (2010), can be used to solve the least squares lasso problem in step (ii2). Our implementation in the R (R Development Core Team, 2020) package RCAL (Tan, 2019b) employs the simple surrogate function \( \ell_{\text{CAL}, Q3}(y; \hat{y}) \) and then a variation of the active set algorithm of Osborne et al. (2000), which enjoys a finite termination property. We need to compute only once and save the Cholesky decomposition of the Gram matrix defined from the vectors \( \{f_j(X_1), \ldots, f_j(X_n)\} \) for the active coordinates \( y_j \) in the active set algorithm.

### 4.3. High-dimensional analysis

In this subsection we provide a high-dimensional analysis of the regularized calibrated estimator \( \hat{\gamma}^1_{\text{RCAL}} \) and the resulting estimator of \( \mu^1 \), allowing for misspecification of model (3). Our analysis of lasso-penalized M-estimators deals with the convergence of such estimators, \( \hat{\gamma}^1_{\text{RCAL}} \), to the target values, \( \hat{\gamma}^1_{\text{CAL}} \), under model misspecification, which is related to but distinct from previous results on excess prediction errors (e.g., Buhlmann & van de Geer, 2011); see the
Supplementary Material for further discussion. Moreover, our analysis of the inverse probability weighted estimator of $\mu^1$ based on $\hat{y}_{RCAL}^1$ carefully exploits the results on the calibration loss in §3.2 to obtain convergence under weaker conditions than previously realized.

As discussed in §3.2, for calibrated estimation with the loss $\ell_{CAL}(\gamma)$, the target value of $\gamma$, denoted by $\hat{y}_{CAL}^1$, is defined as a minimizer of the expected calibration loss

$$E[\ell_{CAL}(\gamma)] = E[T \exp\{-\gamma^T f(X)\} + (1 - T)\gamma^T f(X)].$$

The resulting approximation of $g^*$ is $\bar{g}_{CAL}^1 = (\hat{y}_{CAL}^1)^T f$, which is in general different from $g^*$ in the presence of model misspecification. For our theoretical analysis of $\hat{y}_{CAL}^1$, the tuning parameter in the lasso-penalized loss (16) is specified as $\lambda = A_0\lambda_0$ with a constant $A_0 > 1$ and

$$\lambda_0 = O(1)[\log((1 + p)/\epsilon)/n]^{1/2},$$

where $O(1)$ is a constant depending only on $(B_0, C_0)$ from conditions (i) and (ii) in Proposition 4, and $0 < \epsilon < 1$ is a tail probability for the error bound. For example, taking $\epsilon = 1/(1 + p)$ gives $\lambda_0 = O(1)[2 \log(1 + p)/n]^{1/2}$, a familiar rate in high-dimensional analysis.

Our first result, Proposition 4, establishes the convergence of $\hat{y}_{RCAL}^1$ to $\hat{y}_{CAL}^1$ in the $L_1$-norm $\|\hat{y}_{RCAL}^1 - \hat{y}_{CAL}^1\|_1$ and the symmetrized Bregman divergence between $\bar{g}_{RCAL}^1 = (\hat{y}_{RCAL}^1)^T f$ and $\bar{g}_{CAL}^1 = (\hat{y}_{CAL}^1)^T f$. In fact, convergence is obtained in terms of $D_{CAL}^1(g, g^*) = D_{CAL}(g, g^*) + D_{CAL}(g^*, g) + (A_0 - 1)\lambda_0\|\gamma - \gamma^*\|_1$.

See the Supplementary Material for a discussion of the technical conditions imposed and a comparison with related results in high-dimensional analysis, including those of Buhlmann & van de Geer (2011), Huang & Zhang (2012) and Negahban et al. (2012).

**Proposition 4.** Suppose that:

(i) $\bar{g}_{CAL}^1(X) \geq B_0$ almost surely for a constant $B_0 \in \mathbb{R}$, i.e., $\pi(X; \bar{y}_{CAL}^1)$ is bounded from below by $\{1 + \exp(-B_0)\}^{-1}$;

(ii) for some subset $S \subset \{0, 1, \ldots, p\}$ containing 0 and constants $\nu_0 > 0$ and $\xi_0 > 1$, if $b = (b_0, b_1, \ldots, b_p) \in \mathbb{R}^{1+p}$ satisfies $\sum_{j \notin S} |b_j| \leq \xi_0 \sum_{j \in S} |b_j|$, then $\nu_0 (\sum_{j \in S} |b_j|)^2 \leq |S| (b^T \Sigma_{CAL}^1 b)$ where $\Sigma_{CAL}^1 = E[f(X)T \exp\{-\bar{g}_{CAL}^1(X)\}]f^T(X)$;

(iii) $\max_{j=0,1,\ldots,p} |f_j(X)| \leq C_0$ for a constant $C_0 > 0$;

(iv) $|S|\lambda_0 \leq \eta_0$ for a sufficiently small constant $\eta_0 > 0$.

Then for a sufficiently large constant $A_0$ depending only on $\xi_0$, we have that with probability at least $1 - 4\epsilon$,

$$D_{CAL}^1(\bar{g}_{RCAL}^1, \bar{g}_{CAL}^1) \leq O(1)\left\{\lambda_0 \sum_{j \notin S} |\hat{y}_{CAL,j}^1| + |S|\lambda_0^2\right\},$$

where $O(1)$ depends only on $(A_0, B_0, \xi_0, \nu_0, C_0, \eta_0)$.

From Proposition 4, the following slow and fast rates can be deduced. The two rates are of distinct interest, being valid under different assumptions. Taking $S = \{0\}$ leads to a slow rate of order $\lambda_0 \sum_{j=1}^p |\hat{y}_{CAL,j}^1|$, where the corresponding compatibility assumption is explicitly satisfied under mild conditions.
Corollary 1. Suppose that conditions (i), (iii) and (iv) in Proposition 4 hold with $S = \{0\}$, and that either no linear combination of $f_1(X), \ldots, f_p(X)$ is close to being a constant or the $\psi_2$-weighted $L_2$-norms of $f_1(X), \ldots, f_p(X)$ are bounded away from above by 1, as defined in the Supplementary Material, with $\psi(T, g) = T \exp(-g) + (1 - T)g$. Then for a sufficiently large constant $A_0$, we have that with probability at least $1 - 4\epsilon$,

$$D^\dagger_{\text{CAL}}(\tilde{g}_{\text{RCAL}}^1, \tilde{g}_{\text{CAL}}^1) \leq O(1)\lambda_0 \sum_{j=1}^p |\tilde{\gamma}_{\text{CAL}, j}^1|,$$

(24)

where $O(1)$ depends only on $(A_0, B_0, C_0, \eta_0)$ and $\eta_3$ or $\eta_4$ from (S6) or (S7) in the Supplementary Material.

Taking $S = \{0\} \cup \{j: \tilde{\gamma}_{\text{CAL}, j}^1 \neq 0, j = 1, \ldots, p\}$ yields a fast rate of order $|S|\lambda_0^2$, albeit under a compatibility condition on the linear dependency between $f_1(X), \ldots, f_p(X)$, which may be violated when the number of covariates, $p$, is large.

Corollary 2. Suppose that conditions (i)–(iv) in Proposition 4 hold with $S = \{0\} \cup \{j: \tilde{\gamma}_{\text{CAL}, j}^1 \neq 0, j = 1, \ldots, p\}$. Then for a sufficiently large constant $A_0$, we have that with probability at least $1 - 4\epsilon$,

$$D^\dagger_{\text{CAL}}(\tilde{g}_{\text{RCAL}}^1, \tilde{g}_{\text{CAL}}^1) \leq O(1)|S|\lambda_0^2,$$

(25)

where $O(1)$ depends only on $(A_0, B_0, \xi_0, v_0, C_0, \eta_0)$.

We now examine implications of the preceding results and those in §3.2 on inverse probability weighted estimation. Write $\hat{\pi}_{\text{RCAL}}^1(X) = \pi(X; \tilde{\gamma}_{\text{RCAL}}^1)$, the fitted propensity score based on $\tilde{\gamma}_{\text{RCAL}}$. Consider the resulting estimator in two equivalent forms due to (17),

$$\hat{\mu}_{\text{IPW}}^1(\hat{\pi}_{\text{RCAL}}^1) = \hat{\mu}_{\text{IPW}}^1(\tilde{\gamma}_{\text{RCAL}}^1) = \tilde{E}\left\{\frac{TY}{\tilde{\gamma}_{\text{RCAL}}^1(X)}\right\}.$$

Then a high-probability bound can be obtained on the difference between $\hat{\mu}_{\text{IPW}}^1(\hat{\pi}_{\text{RCAL}}^1)$ and the limiting version $\hat{\mu}_{\text{IPW}}^1(\tilde{\pi}_{\text{CAL}}^1)$ with $\tilde{\pi}_{\text{CAL}}^1(X) = \pi(X; \tilde{\gamma}_{\text{CAL}}^1)$.

Proposition 5. (i) Suppose that the conditions in Corollary 1 hold and that $\sum_{j=1}^p |\tilde{\gamma}_{\text{CAL}, j}^1| \leq M_1$ for a constant $M_1 > 0$. Then for a sufficiently large constant $A_0$, we have that with probability at least $1 - 4\epsilon$,

$$|\hat{\mu}_{\text{IPW}}^1(\hat{\pi}_{\text{RCAL}}^1) - \hat{\mu}_{\text{IPW}}^1(\tilde{\pi}_{\text{CAL}}^1)|^2 \leq O(1)\lambda_0 \tilde{E}\left\{\frac{TY^2}{\tilde{\pi}_{\text{CAL}}^1(X)}\right\},$$

(26)

where $O(1)$ depends only on $(A_0, B_0, C_0, \eta_0, M_1)$ and $\eta_3$ or $\eta_4$ from (S6) or (S7) in the Supplementary Material.

(ii) Suppose that the conditions in Corollary 2 hold. Then for a sufficiently large constant $A_0$, we have that with probability at least $1 - 4\epsilon$,

$$|\hat{\mu}_{\text{IPW}}^1(\hat{\pi}_{\text{RCAL}}^1) - \hat{\mu}_{\text{IPW}}^1(\tilde{\pi}_{\text{CAL}}^1)|^2 \leq O(1)|S|\lambda_0^2 \tilde{E}\left\{\frac{TY^2}{\tilde{\pi}_{\text{CAL}}^1(X)}\right\},$$

(27)

where $O(1)$ depends only on $(A_0, B_0, \xi_0, v_0, C_0, \eta_0)$. 


A remarkable feature of Proposition 5 is that as $\lambda_0 \to 0$, the difference between $\hat{\mu}^1_{IPW}(\hat{\pi}^1_{RCAL})$ and $\hat{\mu}^1_{IPW}(\hat{\pi}^1_{CAL})$ can be seen to converge in probability to zero, even when the $L_1$-norm $\|\hat{\gamma}^1_{RCAL} - \hat{\gamma}^1_{CAL}\|_1$ may not converge to zero. In fact, the $L_1$-norm $\|\hat{\gamma}^1_{RCAL} - \hat{\gamma}^1_{CAL}\|_1$ is in general only bounded from above in probability under the conditions for the slow rate in Corollary 1. The situation with the fast rate in Corollary 2 is similar, but technically subtler: $\|\hat{\gamma}^1_{RCAL} - \hat{\gamma}^1_{CAL}\|_1$ is generally of order $|S|\lambda_0$, which is only required to be sufficiently small, no greater than some positive constant $\eta_0$, but need not be arbitrarily close to zero. The error bounds (26) and (27) are proved by carefully exploiting Proposition 1(i) and Lemma 1 on control of relative errors of propensity scores through the calibration loss $\ell_{CAL}$; see the proofs in the Supplementary Material.

It is interesting to discuss how results related to Proposition 5 can be obtained for inverse probability weighted estimators of $\mu^1$ based on $\hat{\gamma}_{RML}$, such as $\hat{\mu}^1_{IPW}(\hat{\pi}_{RML})$. In previous analyses (e.g., Belloni et al., 2014, § 5), such results were obtained under additional conditions ensuring that the $L_1$-norm $\|\hat{\gamma}_{RML} - \gamma_{ML}\|_1$ converges in probability to zero. Alternatively, by mimicking our proof of Proposition 5, similar results can be shown for $\hat{\mu}^1_{IPW}(\hat{\pi}_{RML})$ without requiring convergence of $\|\hat{\gamma}_{RML} - \gamma_{ML}\|_1$ to zero. Analogously to (27), it can be shown under conditions comparable to those in Corollary 2 that with high probability,

$$
|\hat{\mu}^1_{IPW}(\hat{\pi}_{RML}) - \hat{\mu}^1_{IPW}(\hat{\pi}_{ML})|^2 \leq O(1)|S|\lambda_0^2E\left\{\frac{T\bar{y}^2}{\bar{\pi}_{ML}(X)^2}\right\},
$$

(28)

where $S'$ is a counterpart of $S$. There is, however, a subtle difference between (27) and (28). Write $\hat{\pi}_{RML}^\gamma = \hat{\gamma}^\gamma_{RML}f$ and $\hat{\pi}_{ML}^\gamma = \hat{\gamma}^\gamma_{ML}f$. As discussed in the Supplementary Material, $O(1)|S|\lambda_0^2$ in (27) is obtained as $5/(3a)D_{CAL}(\hat{\pi}_{RCAL}^\gamma, \hat{\gamma}_{RCAL})$, whereas $O(1)|S|\lambda_0^2$ in (28) is obtained as $1/(2\delta^2d^2)D_{ML}(\hat{\pi}_{RML}^\gamma, \hat{\pi}_{ML}^\gamma)$, where $a > 0$ and $d' > 0$ are constants such that $\hat{\pi}_{RCAL}(X)/\hat{\pi}_{CAL}(X) \geq a$ and $\hat{\pi}_{RML}(X)/\hat{\pi}_{ML}(X) \geq d'$ with high probabilities, and $\delta > 0$ is assumed to satisfy $\hat{\pi}_{ML}(X) \geq \delta$, similar to the lower bound $\{1 + \exp(-B_0)|\}^{-1}$ in Proposition 4. This zoomed-in view shows a comparison reminiscent of that between (12) and (13). The likelihood divergence $D_{ML}(\hat{\pi}_{RML}, \hat{\pi}_{ML})$ can be inflated by a much larger factor for the error bound (28) than $D_{CAL}(\hat{\pi}_{RCAL}, \hat{\gamma}_{RCAL})$ for the error bound (27).

Finally, although Proposition 5 deals with convergence of $\hat{\mu}^1_{IPW}(\hat{\pi}_{RCAL})$ to $\hat{\mu}^1_{IPW}(\hat{\pi}_{CAL})$, which may differ from the parameter of interest $\mu^1$, we point out that Propositions 2 and 5 are complementary in providing support for the use of $\hat{\pi}_{RCAL}^\gamma$ in inverse probability weighted estimation of $\mu^1$. The argument is based on the triangle inequality,

$$
|\hat{\mu}^1_{IPW}(\hat{\pi}_{RCAL}) - \mu^1| \leq |\hat{\mu}^1_{IPW}(\hat{\pi}_{CAL}) - \mu^1| + |\hat{\mu}^1_{IPW}(\hat{\pi}_{RCAL}) - \hat{\mu}^1_{IPW}(\hat{\pi}_{CAL})|.
$$

(29)

The first and second terms in (29) represent, respectively, limiting bias and sampling variation. On the one hand, as discussed after Proposition 2, minimization of the expected calibration loss facilitates achieving a smaller first term, $|\hat{\mu}^1_{IPW}(\hat{\pi}_{CAL}) - \mu^1|$, than maximum likelihood estimation. On the other hand, by the preceding discussion, using the calibration loss with lasso penalization makes it possible to achieve a smaller second term, $|\hat{\mu}^1_{IPW}(\hat{\pi}_{RCAL}) - \hat{\mu}^1_{IPW}(\hat{\pi}_{CAL})|$, than maximum likelihood estimation with lasso penalization in high-dimensional settings.

5. SIMULATION STUDY

We conduct a simulation study extending the design of Kang & Schafer (2007) to high-dimensional, sparse settings. For $p \geq 4$, let $X = (X_1, \ldots, X_p)^T$ be multivariate normal with mean
0 and covariances \( \text{cov}(X_j, X_k) = 0 \) if \( 1 \leq j, k \leq 4 \) and \( \text{cov}(X_j, X_k) = 2^{-|j-k|} \) if \( 5 \leq j \leq p \) or \( 5 \leq k \leq p \), and let \( T \) be a binary variable such that

\[
\Pr(T = 1 \mid X) = \pi^*(X) = \left[1 + \exp\{X_1 - 0.5X_2 + 0.25X_3 + 0.1X_4\}\right]^{-1},
\]

which depends only on the four covariates \((X_1, X_2, X_3, X_4)\). Consider two versions of the logistic model (3) with the following regressors: (i) \( f_j(X) = X_j \) for \( j = 1, \ldots, p \); (ii) \( f_j(X) \) is a standardized version of \( W_j \) with sample mean 0 and sample variance 1, where \( W_1 = \exp(0.5X_1), W_2 = 10 + (1 + \exp(X_1))^{-1}X_2, W_3 = (0.04X_1X_3 + 0.6)^3, W_4 = (X_2 + X_4 + 20)^2 \) and, if \( p > 4 \), \( W_j = X_j \) for \( j = 5, \ldots, p \). Then model (3) is correctly specified in scenario (i), but misspecified in scenario (ii). For \( p = 4 \), model (3) in scenario (ii), although misspecified, appears adequate by conventional diagnostics, as shown in Kang & Schafer (2007). In addition, with an outcome variable \( Y \), this simulation with \( p = 4 \) has been widely used to study estimators for \( \mu^1 = E(Y^1) \) with observed data \( \{(T_i, Y_i^1, T_i, X_i) : i = 1, \ldots, n\} \).

We compare six estimators in the ratio form \( \hat{\mu}^1_{\text{IPW}}(\hat{\pi}) \), labelled as follows:

(I) \( \hat{\pi} \) is replaced by the true propensity score \( \pi^* \);

(II) \( \hat{\pi} = \hat{E}(T) \) obtained from model (3) with only the intercept \( f \equiv 1 \);

(III) \( \hat{\pi} = \hat{\pi}_{\text{ML}} \) obtained by maximum likelihood, i.e., by minimizing (2);

(IV) \( \hat{\pi} = \hat{\pi}_{\text{RML}} \) obtained by lasso-penalized maximum likelihood, i.e., by minimizing (19);

(V) \( \hat{\pi} = \hat{\pi}_{\text{CAL}} \) obtained by calibrated estimation, i.e., by minimizing (6);

(VI) \( \hat{\pi} = \hat{\pi}_{\text{RCAL}} \) obtained by regularized calibrated estimation, i.e., by minimizing (16).

The functions (2) and (6) are minimized using the \( \mathbb{R} \) package trust (Geyer, 2014); (19) and (16) are minimized using the \( \mathbb{R} \) package RCAL (Tan, 2019b), based on Friedman et al. (2010) and our algorithm in §4.2. The tuning parameter \( \lambda \) in (16) or (19) is determined using five-fold cross-validation based on the corresponding loss function. For \( k = 1, \ldots, 5 \) let \( T_k \) be a random subsample of size \( n/5 \) from \( \{1, 2, \ldots, n\} \). For a loss function \( \ell(\gamma) \), for example \( \ell_{\text{CAL}}(\gamma) \) in (6), denote by \( \ell(\gamma; T) \) the loss function obtained when the sample average \( \hat{E}(\cdot) \) is computed over only the subsample \( T \). The five-fold cross-validation criterion is defined as

\[
\text{CV}_5(\lambda) = (1/5) \sum_{k=1}^5 \ell(\hat{\gamma}^{(k)}; T_k),
\]

where \( \hat{\gamma}^{(k)}(\lambda) \) is a minimizer of the penalized loss \( \ell(\gamma; T_k) + \lambda \|\gamma\|_1 \) over the subsample \( T_k \) of size \( 4n/5 \). Then \( \lambda \) is selected by minimizing \( \text{CV}_5(\lambda) \) over the discrete set \( \{\lambda^*/2^j : j = 0, 1, \ldots, 10\} \), where for \( \pi_0 = \hat{E}(T) \) the value \( \lambda^* \) is computed as

\[
\lambda^* = \max_{j=1,\ldots,p} |\hat{E}\{E(T - \hat{\pi}_0)f_j(X)|\}
\]

when the likelihood loss (2) is used, or as \( \lambda^* = \max_{j=1,\ldots,p} |\hat{E}\{(T/\hat{\pi}_0 - 1)f_j(X)|\} \) when the calibration loss (6) is used. It can be shown that in either case, the penalized loss \( \ell(\gamma) + \lambda \|\gamma\|_1 \) over the original sample has a minimum at \( \gamma|_p = 0 \) for all \( \lambda \geq \lambda^* \).

The performance of an inverse probability weighted estimator \( \hat{\mu}^1_{\text{IPW}}(\hat{\pi}) \) depends not only on the closeness of \( \hat{\pi} \) to \( \pi^* \), but also on the outcome regression function \( m^*_i(X) = E(Y^1 \mid X) \) and the error \( e = Y^1 - m^*_i(X) \). Although such dependency can be exploited to achieve local efficiency and double robustness by incorporating an outcome regression model, we are interested in comparing inverse probability weighted estimators regardless of outcome regression models. Under unconfoundedness, it can be shown via conditioning on \( \{(T_i, X_i) : i = 1, \ldots, n\} \) that

\[
E(\hat{\mu}^1_{\text{IPW}}(\hat{\pi})) = E(\hat{\mu}^1_{\text{IPW}}(\hat{\pi} ; m^*_i)),
\]

\[
\text{var}(\hat{\mu}^1_{\text{IPW}}(\hat{\pi})) = \text{var}(\hat{\mu}^1_{\text{IPW}}(\hat{\pi} ; m^*_i)) + \text{var}(\hat{\mu}^1_{\text{IPW}}(\hat{\pi} ; e)),
\]

where \( \hat{\mu}^1_{\text{IPW}}(\hat{\pi} ; h) = \hat{E}(Th(X)/\hat{\pi}(X))/\hat{E}(T/\hat{\pi}(X)) \) for a function \( h(X) \) and \( \hat{\mu}^1_{\text{IPW}}(\hat{\pi} ; e) = \hat{E}(Te/\hat{\pi}(X))/\hat{E}(T/\hat{\pi}(X)) \). As a result, the mean squared error \( E[(\hat{\mu}^1_{\text{IPW}}(\hat{\pi}) - \mu)^2] \) can...
The advantage of or 800 (right) within the column for each estimator. The estimators π by the model (3) is correctly specified, with for the few observations with be decomposed as $\text{MSE}[\hat{\mu}_{\text{IPW}}(\hat{\pi}; \cdot; \cdot)] = \text{var}[\hat{\mu}_{\text{IPW}}(\hat{\pi}; \cdot; \cdot)]$, where $\text{MSE}[\hat{\mu}_{\text{IPW}}(\hat{\pi}; \cdot; \cdot)] = E[(\hat{\mu}_{\text{IPW}}(\hat{\pi}; \cdot; \cdot) - E[h(X)])^2]$. We consider several configurations for $X = (\cdot; \cdot)$: (lin1) $h(X) = X_1 + 0.5X_2 + 0.5X_3 + 0.5X_4$; (lin2) $h(X) = X_1 + 2X_2 + 2X_3 + 2X_4$; (quad1) $h(X) = \sum_{j=1}^{4} \{\max(X_j, 0)\}^2$; (quad2) $h(X) = \sum_{j=1}^{4} \{\max(-X_j, 0)\}^2$ and (exp) $h(X) = \exp(\sum_{j=1}^{4} X_j/2)$. The configuration lin1 is similar to that in Kang & Schafer (2007) up to a linear transformation.

For model (3) correctly specified or misspecified, Fig. 2 or 3 displays, respectively, Monte Carlo estimates of $\text{MSE}^{1/2}[\hat{\mu}_{\text{IPW}}(\hat{\pi}; \cdot; \cdot)]$ with five choices of $h(X)$ and $\text{var}^{1/2}[\hat{\mu}_{\text{IPW}}(\hat{\pi}; \cdot; \cdot)] = \exp(\sum_{j=1}^{4} X_j/2)$. The nonpenalized estimators $\hat{\pi}_{\text{ML}}$ and $\hat{\pi}_{\text{CAL}}$ are computed only for $p = 4, 50$ and 50. For $(p, n) = (20, 200)$ or $(50, \leq 400)$, the estimator $\hat{\pi}_{\text{CAL}}$ is obtained with nonconvergence declared by the R package trust in 30–100% of simulations, indicating that the loss function $\ell_{\text{CAL}}(\cdot; \cdot)$ may not have a finite minimum. We summarize the findings as follows.

First, for all the choices of $(n, p)$ and $h(X)$ studied, the estimator $\hat{\pi}_{\text{CAL}}^{1}$ yields similar or smaller mean squared errors than $\hat{\pi}_{\text{RML}}$, whether model (3) is correctly specified or misspecified. The advantage of $\hat{\pi}_{\text{CAL}}^{1}$ is substantial in the case of misspecified model (3).

Second, for relatively small $p$, 50 or less, the estimator $\hat{\pi}_{\text{CAL}}^{1}$, in spite of the nonconvergence issue mentioned above, consistently leads to smaller mean squared errors than $\hat{\pi}_{\text{ML}}$, whether model (3) is correctly specified or misspecified. In the case of misspecified model (3), the performance of $\hat{\pi}_{\text{ML}}$ deteriorates substantially, particularly for estimation associated with the configurations quad1 and exp for $h(X)$. A possible explanation is that $h(X)$ in these cases quickly increases as $(X_1, X_3, X_4)$ become large, which by (30) is the region where the propensity score $\pi^*(X)$ becomes small. Even a small discrepancy, especially underestimation, between $\hat{\pi}$ and $\pi^*$ for the few observations with $T = 1$ in this region can give rise to large errors for the estimates $\hat{\mu}_{\text{IPW}}(\hat{\pi}; \cdot; \cdot)$. 

Fig. 2. Root mean squared errors of $\hat{\mu}_{\text{IPW}}(\hat{\pi}; h)$ and $\hat{\mu}_{\text{IPW}}(\hat{\pi}; \pi)$ for the estimators $\hat{\pi}$ labelled I–VI when the logistic model (3) is correctly specified, with $p = 4(\Delta), 20(+)\), 50(\times), 100(\gamma)$ and with $n = 200$ (left), 400 (middle) or 800 (right) within the column for each estimator. The estimators $\hat{\pi}_{\text{ML}}$ and $\hat{\pi}_{\text{CAL}}$, i.e., III and V, are computed only for $p = 4, 20$ and 50. The results for $\hat{\pi}_{\text{CAL}}^{1}$ should be interpreted with caution for $(p, n) = (20, 200)$ and $(50, \leq 400)$. 

For model (3) correctly specified or misspecified, Fig. 2 or 3 displays, respectively, Monte Carlo estimates of $\text{MSE}^{1/2}[\hat{\mu}_{\text{IPW}}(\hat{\pi}; \cdot; \cdot)]$ with five choices of $h(X)$ and $\text{var}^{1/2}[\hat{\mu}_{\text{IPW}}(\hat{\pi}; \cdot; \cdot)] = \exp(\sum_{j=1}^{4} X_j/2)$. The nonpenalized estimators $\hat{\pi}_{\text{ML}}$ and $\hat{\pi}_{\text{CAL}}$ are computed only for $p = 4, 50$ and 50. For $(p, n) = (20, 200)$ or $(50, \leq 400)$, the estimator $\hat{\pi}_{\text{CAL}}^{1}$ is obtained with nonconvergence declared by the R package trust in 30–100% of simulations, indicating that the loss function $\ell_{\text{CAL}}(\cdot; \cdot)$ may not have a finite minimum. We summarize the findings as follows.

First, for all the choices of $(n, p)$ and $h(X)$ studied, the estimator $\hat{\pi}_{\text{CAL}}^{1}$ yields similar or smaller mean squared errors than $\hat{\pi}_{\text{RML}}$, whether model (3) is correctly specified or misspecified. The advantage of $\hat{\pi}_{\text{CAL}}^{1}$ is substantial in the case of misspecified model (3).

Second, for relatively small $p$, 50 or less, the estimator $\hat{\pi}_{\text{CAL}}^{1}$, in spite of the nonconvergence issue mentioned above, consistently leads to smaller mean squared errors than $\hat{\pi}_{\text{ML}}$, whether model (3) is correctly specified or misspecified. In the case of misspecified model (3), the performance of $\hat{\pi}_{\text{ML}}$ deteriorates substantially, particularly for estimation associated with the configurations quad1 and exp for $h(X)$. A possible explanation is that $h(X)$ in these cases quickly increases as $(X_1, X_3, X_4)$ become large, which by (30) is the region where the propensity score $\pi^*(X)$ becomes small. Even a small discrepancy, especially underestimation, between $\hat{\pi}$ and $\pi^*$ for the few observations with $T = 1$ in this region can give rise to large errors for the estimates $\hat{\mu}_{\text{IPW}}(\hat{\pi}; \cdot; \cdot)$. 

Fig. 2. Root mean squared errors of $\hat{\mu}_{\text{IPW}}(\hat{\pi}; h)$ and $\hat{\mu}_{\text{IPW}}(\hat{\pi}; \pi)$ for the estimators $\hat{\pi}$ labelled I–VI when the logistic model (3) is correctly specified, with $p = 4(\Delta), 20(+)\), 50(\times), 100(\gamma)$ and with $n = 200$ (left), 400 (middle) or 800 (right) within the column for each estimator. The estimators $\hat{\pi}_{\text{ML}}$ and $\hat{\pi}_{\text{CAL}}$, i.e., III and V, are computed only for $p = 4, 20$ and 50. The results for $\hat{\pi}_{\text{CAL}}^{1}$ should be interpreted with caution for $(p, n) = (20, 200)$ and $(50, \leq 400).$
Calibrated estimation of propensity scores

Fig. 3. Root mean squared errors of \( \hat{\mu}_{\text{IPW}}(\hat{\pi}; h) \) and \( \hat{\mu}_{\text{IPW}}(\hat{\pi}; \varepsilon) \) plotted as in Fig. 2, for the estimators \( \hat{\pi} \) labelled I–VI when the logistic model (3) is misspecified; the values are censored within the upper limit of the vertical axis (dotted line).

Third, the regularized estimator \( \hat{\pi}_{\text{RML}} \) or \( \hat{\pi}_{\text{RCAL}} \) yields smaller or slightly larger mean squared errors than the corresponding nonregularized estimator \( \hat{\pi}_{\text{ML}} \) or \( \hat{\pi}_{\text{CAL}} \) with \( p \leq 50 \), except for \( \hat{\pi}_{\text{RCAL}} \) versus \( \hat{\pi}_{\text{CAL}} \) in the configuration lin1 for \( h(X) \); see the Supplementary Material for a discussion.

In the Supplementary Material we report various additional results, including the number of samples with nonconvergence for \( \hat{\gamma}_{\text{ML}} \) and \( \hat{\gamma}_{\text{CAL}} \), the average numbers of nonzero coefficients obtained in \( \hat{\gamma}_{\text{RML}} \) and \( \hat{\gamma}_{\text{RCAL}} \), and the mean squared errors related to estimation of \( \mu^1 - \mu^0 \).

6. Application to a Medical Study

We present an empirical application of the proposed methods to a medical study described in Connors et al. (1996) on the effects of right heart catheterization. The observational study was of interest at a time when many physicians believed that the procedure led to better patient outcomes, but the benefit had not been demonstrated in any randomized clinical trials. The study of Connors et al. (1996) included \( n = 5735 \) critically ill patients admitted to five medical centres. For each patient, the data consist of the treatment status \( T \), defined as 1 if the procedure was used within 24 hours of admission and 0 otherwise, the health outcome \( Y \), defined as survival time up to 30 days, and a list of 75 covariates \( X \) specified by medical specialists in critical care. In previous analyses using propensity scores, logistic regression was employed either with main effects only (e.g., Hirano & Imbens, 2002; Vermeulen & Vansteelandt, 2015) or with interaction terms manually added (Tan, 2006) in the approach of Rosenbaum & Rubin (1984).

To capture possible dependency beyond main effects, we consider a logistic propensity score model (3), where the vector \( f(X) \) includes all main effects and two-way interactions of \( X \) except those with numbers of nonzero values less than 46, i.e., 0.8% of the sample of 5735.
The dimension of $f(X)$ is $p = 1855$, excluding the constant. All variables in $f(X)$ are standardized to have sample mean 0 and sample variance 1. We apply the methods of regularized maximum likelihood and regularized calibrated estimation, labelled as RML and RCAL, similarly to how we conducted the simulation study, with the lasso tuning parameter $\lambda$ attempted in a finer set $\{ \lambda^*/2i/4 : i = 0, 1, \ldots, 24 \}$, where $\lambda^*$ is the value leading to a zero solution $\gamma_1 = \cdots = \gamma_p = 0$.

To measure the effect of calibration in the treated sample for a function $h(X)$ using a propensity score estimate $\hat{\pi}$, we use the standardized calibration difference $CAL^1(\hat{\pi}; h) = [\hat{\mu}_h^{IPW}(\hat{\pi}; h) - \bar{E}(h(X))]/\tilde{V}^{1/2}(h(X))$, where $\bar{E}(\cdot)$ and $\tilde{V}(\cdot)$ denote the sample mean and variance, respectively, and $\hat{\mu}_h^{IPW}(\hat{\pi}; h)$ is defined as $\hat{\mu}_h^{IPW}(\hat{\pi})$ with $Y$ replaced by $h(X)$. For $f_j(X)$ standardized with sample mean 0 and sample variance 1, $CAL^1(\hat{\pi}; f_j)$ reduces to $\hat{\mu}_h^{IPW}(\hat{\pi}; f_j)$. See, for example, Austin & Stuart (2015, § 4.1.1) for a related statistic based on $\hat{\mu}_h^{IPW}(\hat{\pi}; h) - \hat{\mu}_h^{IPW}(\hat{\pi}; h)$ for balance checking. Figure 4 displays the standardized calibration differences for all the variables $f_j(X)$ and the fitted propensity scores in the treated sample, obtained from the regularized estimators $\hat{\pi}_{RML}$ and $\hat{\pi}_{RCAL}$ with the tuning parameter $\lambda$ selected by five-fold cross-validation as in § 5.

From Fig. 4 it can be seen that the maximum absolute standardized differences are reduced from 35% to about 10% based on the estimators $\hat{\pi}_{RML}$ and $\hat{\pi}_{RCAL}$. However, the latter estimator $\hat{\pi}_{RCAL}$ is obtained with a much smaller number, 32 versus 188, of nonzero estimates of coefficients $\gamma_j$. The corresponding standardized differences for these 32 nonzero coefficients precisely attain the maximum absolute value, 0.102, which is also the tuning parameter $\lambda$ used for the lasso penalty in (18). The fitted propensity scores $\hat{\pi}_{RCAL}(X_i)$ in the treated sample are consistently larger or smaller than $\hat{\pi}_{RML}(X_i)$ when close to 0 or 1, respectively. As a result, the inverse probability
weights $1/\hat{\pi}_{\text{RCAL}}^1(X_i)$ tend to be less variable than $1/\hat{\pi}_{\text{RML}}(X_i)$, which is also confirmed in the following discussion.

Figure 5 shows how the maximum absolute standardized differences are related to the numbers of nonzero estimates of $y_j$ and the relative variances of the inverse probability weights in the treated sample. As seen from Fig. 5, in the process of reducing the relative variance is defined as $\sum_{j=1}^{n_i} (\hat{w}_j - \bar{w})^2/((n_i - 1)\bar{w}^2)$, where $\bar{w} = \sum_{j=1}^{n_i} w_i/n_i$ and $n_i/n = \bar{E}(T)$. See Liu (2001, § 2.5.3) for a discussion on using the relative variance to measure the efficiency of a weighted sample. As seen from Fig. 5, in the process of reducing the standardized differences, the estimator $\hat{\pi}_{\text{RCAL}}^1$ is associated with a much smaller number of nonzero coefficients $y_j$, or greater sparsity, and with smaller relative variances of the inverse probability weights, or greater efficiency, than $\hat{\pi}_{\text{RML}}$. These results demonstrate the advantages of regularized calibrated estimation in high-dimensional settings.

Additional results are reported in the Supplementary Material, including results that parallel those in Fig. 5 for the fitted propensity score $\hat{\pi}_{\text{RCAL}}^0$ in the untreated sample, as well as the estimates of $\mu^1$, $\mu^0$ and $\mu^1 - \mu^1$ for the 30-day survival status indicating $Y \geq 30$.

7. FURTHER DISCUSSION

7.1. Dual formulation

The regularized calibrated estimator $\hat{\pi}_{\text{RCAL}}^1$ can also be derived in a dual formulation. Let $w = \{w_i > 1 : T_i = 1, i = 1, \ldots, n\}$. For fixed $\lambda > 0$, consider the following optimization problem:

$$\begin{align*}
\text{minimize} & \quad \ell_{\text{CAL}}^D(w) = \sum_{1 \leq i \leq n : T_i = 1} (w_i - 1) \log(w_i - 1) - (w_i - 1) \\
\text{subject to} & \quad \sum_{1 \leq i \leq n : T_i = 1} w_i = n, \\
& \quad \left| \sum_{1 \leq i \leq n : T_i = 1} w_i f_j(X_i) - \sum_{i=1}^{n} f_j(X_i) \right| \leq \lambda \quad (j = 1, \ldots, p).
\end{align*}$$

(31)–(33)

It can be shown directly via the Karush–Kuhn–Tucker condition that if $\hat{\pi}_{\text{RCAL}}^1$ minimizes the penalized loss $\ell_{\text{RCA1}}(\gamma)$ in (16), then the inverse probability weights $\hat{w}_i = \{\hat{\pi}_{\text{RCAL}}^1(X_i)\}^{-1} = 1 + \exp\{-(\hat{\pi}_{\text{RCAL}}^1)^T f(X_i)\}$ are a solution to (31)–(33). In the case of exact calibration ($\lambda = 0$), (31)–(33) can be obtained from Chan et al. (2016) with the particular distance measure $\ell_{\text{CAL}}^D(w)$; see also Zubizarreta (2015) for
a related method. Similarly, for the regularized likelihood estimator $\hat{\gamma}_{RML}$ minimizing $\ell_{RML}(\gamma)$ in (19), it can be shown that the fitted propensity scores $\hat{\pi}_i = \hat{\pi}_{RML}(X_i) \ (i=1, \ldots, n)$ solve the following program with $\pi = \{0 < \pi_i < 1 : i = 1, \ldots, n\}$:

$$\begin{align*}
\text{minimize} & \quad \ell_{ML}^0(\pi) = \sum_{i=1}^n (1 - \pi_i) \log(1 - \pi_i) + \pi_i \log(\pi_i) \\
\text{subject to} & \quad \sum_{i=1}^n (T_i - \pi_i) = 0, \\
& \quad \left| \sum_{i=1}^n (T_i - \pi_i) f_j(X_i) \right| \leq \lambda \quad (j = 1, \ldots, p).
\end{align*}$$

(34)

See Dudik et al. (2007) for general results relating box constraints such as (33) and (34) to the lasso penalty in a conceptually similar context. These dual formulations help to shed light on the differences between maximum likelihood and calibration estimation, which deal with, respectively, propensity scores in the probability scale and the scale of inverse probability weights.

We distinguish between two types of calibration estimators viewed in a dual formulation, which in previous work usually involved exact constraints. The first type is survey calibration (Deville & Sarndal, 1992), in which calibration weights are constructed by minimizing a measure of distance to the design weights, i.e., inverses of inclusion probabilities, subject to calibration equations. Similar ideas are used in Tan (2010, 2013) to derive improved doubly robust estimators, through adjusting the inverses of fitted propensity scores to achieve calibration constraints, possibly depending on a fitted outcome regression function. The second type of calibrated estimators, such as $\hat{\gamma}_{CAL}^1$ for $\gamma$ or $\hat{\mu}_{CAL}^1$ for $\mu^1$, are typically derived to deal with nonresponse or missing data, in a similar manner to the first type with uniform design weights. But there is a subtle difference. The survey calibration weights (Deville & Sarndal, 1992) are expected to deviate from the design weights by $O_p(n^{-1/2})$, and are mainly used to reduce the asymptotic variances of the resulting estimators of population quantities. The calibration weights of the second type can be viewed as the inverses of fitted response probabilities or propensity scores from a model implied by the choice of distance measure, and are expected to behave as $O_p(1)$ to achieve bias reduction.

7.2. Estimation of average treatment effects

So far our theory and methods have focused mainly on estimation of $\mu^1$, but they can be directly extended to estimation of $\mu^0$, and hence the average treatment effect $\mu^1 - \mu^0$. As mentioned in § 3, for estimation of $\mu^0$ with model (3), the calibrated estimator of $\gamma$, denoted by $\hat{\gamma}_{CAL}^0$, is defined as a solution to (5). By exchanging $T$ with $1 - T$ and $\gamma$ with $-\gamma$ in (6), the corresponding loss function minimized by $\hat{\gamma}_{CAL}^0$ is

$$\ell_{CAL}^0(\gamma) = \hat{E}[ (1 - T) \exp\{\gamma^T f(X) \} - T \gamma^T f(X)].$$

For fixed $\lambda \geq 0$, the regularized calibrated estimator $\hat{\gamma}_{RCAL}^0$ is defined as a minimizer of

$$\ell_{RCAL}^0(\gamma) = \ell_{CAL}(\gamma) + \lambda \|\gamma_{1:p}\|_1.$$ 

The fitted propensity score, $\hat{\pi}_{RCAL}^0(X) = \pi(X) \hat{\gamma}_{RCAL}^0$, then satisfies (17) and (18) with $T_i$ replaced by $1 - T_i$ and $\hat{\pi}_{RCAL}^0(X_i)$ replaced by $1 - \hat{\pi}_{RCAL}^0(X_i)$. The resulting estimator of $\mu^0$ is $\hat{\mu}_{IPW}^0(\hat{\pi}_{RCAL}^0) = \hat{\mu}_{IPW}(\hat{\pi}_{RCAL}^0)$, and that of $\mu^1 - \mu^0$ is $\hat{\mu}_{IPW}^1(\hat{\pi}_{RCAL}^0) - \hat{\mu}_{IPW}^0(\hat{\pi}_{RCAL}^0)$.

An interesting feature of our approach is that two different estimators of the propensity score are used when estimating $\mu^0$ and $\mu^1$. The estimators $\hat{\gamma}_{RCAL}^0$ and $\hat{\gamma}_{RCAL}^1$ may in general have different asymptotic limits when the propensity score model (3) is misspecified, even though their asymptotic limits coincide when model (3) is correctly specified. Such possible differences
7.3. Calibration or balancing

It is interesting to compare calibrated propensity scores with the covariate balancing propensity scores of Imai & Ratkovic (2014). For model (1), the covariate balancing estimator of $\gamma$, denoted by $\hat{\gamma}_{\text{BAL}}$, is defined as a solution to

$$\tilde{E} \left[ \left\{ \frac{T}{\pi(X; \gamma)} - \frac{1 - T}{1 - \pi(X; \gamma)} \right\} f(X) \right] = 0. \tag{35}$$

The same fitted propensity score $\hat{\pi}_{\text{BAL}}(X) = \pi(X; \hat{\gamma}_{\text{BAL}})$ can be used in $\hat{\mu}_{\text{IPW}}^{1}(\hat{\pi}_{\text{BAL}})$ or $\hat{\mu}_{\text{IPW}}^{1}(\hat{\pi}_{\text{BAL}})$ for estimating $\mu$ and in $\hat{\mu}_{\text{IPW}}^{0}(\hat{\pi}_{\text{BAL}})$ or $\hat{\mu}_{\text{IPW}}^{0}(\hat{\pi}_{\text{BAL}})$ for estimating $\mu^0$. Equation (35) amounts to finding a single value $\hat{\gamma}_{\text{BAL}}$ such that the left-hand sides of (4) and (5) are equal, although they may each deviate from zero. For calibrated estimation, (4) and (5) are satisfied separately by two estimators $\hat{\gamma}_{\text{BAL}}^{1}$ and $\hat{\gamma}_{\text{BAL}}^{0}$. An advantage of using the calibration equations (4) and (5) is that for $t = 0$ or 1, $\hat{\mu}_{\text{IPW}}^{0}(\hat{\pi}_{\text{CAL}})$, but not $\hat{\mu}_{\text{IPW}}^{1}(\hat{\pi}_{\text{CAL}})$, is doubly robust, i.e., remains consistent if either the propensity score model (1) or a linear outcome model is correct, $E(Y^t | X) = \alpha_t f(X)$ for a coefficient vector $\alpha_t$ (Graham et al., 2012). We also point out that with the logistic model (3), $\hat{\gamma}_{\text{BAL}}$ can be obtained by minimizing the loss function

$$\ell_{\text{BAL}}(\gamma) = \ell_{\text{CAL}}(\gamma) + \ell_{\text{CAL}}^{0}(\gamma), \tag{36}$$

which is still convex in $\gamma$. Our results developed for calibrated estimation and regularization can be adapted to $\hat{\gamma}_{\text{BAL}}$ and its regularized version. See Fig. 1 for a comparison of limiting propensity scores in a simple example with model misspecification.

7.4. Estimation of average treatment effects on the treated

There is a simple extension of our approach to estimation of the average treatment effect on the treated, $\nu^1 - \nu^0$, as defined in § 2. The parameter $\nu^1 = E(Y^1 | T = 1)$ can be directly estimated by $\tilde{E}(TY)/\tilde{E}(T)$. Two standard inverse probability weighted estimators for $\nu^0$ are

$$\nu_{\text{IPW}}^0(\hat{\pi}_{\text{ML}}) = \tilde{E} \left\{ \frac{(1 - T)\hat{\pi}_{\text{ML}}(X)Y}{1 - \hat{\pi}_{\text{ML}}(X)} \right\} / \tilde{E}(T)$$

and $\nu_{\text{IPW}}^0(\hat{\pi}_{\text{ML}})$, which is defined as $\nu_{\text{IPW}}^0(\hat{\pi}_{\text{ML}})$ with $\tilde{E}(T)$ replaced by $\tilde{E}[(1 - T)\hat{\pi}_{\text{ML}}(X)/(1 - \hat{\pi}_{\text{ML}}(X))]$, where $\hat{\pi}_{\text{ML}}(X)$ is the maximum likelihood fitted propensity score. To derive a calibrated estimator of $\gamma$, consider the set of calibration equations

$$\tilde{E} \left[ \left\{ \frac{(1 - T)\pi(X; \gamma)}{1 - \pi(X; \gamma)} - T \right\} f(X) \right] = 0. \tag{37}$$

Equation (37) is used in Imai & Ratkovic (2014) as a balancing equation in propensity score estimation for estimating $\nu^1 - \nu^0$. We point out two simple results which, though straightforward as shown below, do not seem to have been discussed before.
First, (37) is equivalent to the calibration equations (5) for $\hat{\gamma}^0_{\text{CAL}}$ when estimating $\mu^0$. This follows from the simple identity $(1 - T)\pi(X; \gamma)/(1 - \pi(X; \gamma)) - T = (1 - T)/(1 - \pi(X; \gamma)) - 1$. Therefore the same set of fitted propensity scores, for example $\hat{\pi}^0_{\text{RCAL}}(X_i)$ based on the regularized estimator $\hat{\gamma}^0_{\text{RCAL}}$, can be used for estimating $\mu^0$ by $\hat{\mu}^0_{\text{IPW}}(\hat{\pi}^0_{\text{RCAL}})$ and for estimating $\nu^0$ by $\hat{\nu}^0_{\text{IPW}}(\hat{\pi}^0_{\text{RCAL}}) = \tilde{\nu}_{\text{IPW}}(\hat{\pi}^0_{\text{RCAL}})$, due to an equation similar to (17).

Second, for the logistic model (3), the estimator $\hat{\nu}^0_{\text{HPW}}(\hat{\pi}^0_{\text{CAL}})$ is identical to the estimator of $\nu^0$ from entropy balancing (Hainmueller, 2012). In fact, $\hat{\nu}^0_{\text{HPW}}(\hat{\pi}^0_{\text{CAL}})$ can be rewritten as $\sum_{i:T_i=0} \hat{w}_i Y_i$ where, for $\hat{\gamma} = \hat{\gamma}^0_{\text{CAL}}$, $\hat{y}_{1:p} = (\hat{y}_1, \ldots, \hat{y}_p)^T$ and $f_{1:p} = (f_1, \ldots, f_p)^T$,

$$\hat{w}_i = \frac{\exp\{-\hat{y}^T_{1:p} f_{1:p}(X_i)\}}{\sum_{i':T_{i'}=0} \exp\{-\hat{y}^T_{1:p} f_{1:p}(X_{i'})\}}.$$

Equation (37) for $\hat{\gamma}^0_{\text{CAL}}$ then implies that $\sum_{i:T_i=0} \hat{w}_i = 1$ and

$$\sum_{i:T_i=0} \hat{w}_i f_j(X_i) = \frac{\tilde{E}\left\{\frac{(1 - T)\pi(X; \hat{\gamma})}{1 - \pi(X; \hat{\gamma})} f(X)\right\}}{\tilde{E}\left\{\frac{(1 - T)\pi(X; \hat{\gamma})}{1 - \pi(X; \hat{\gamma})}\right\}} = \frac{\tilde{E}\{T f(X)\}}{\tilde{E}(T)} = \frac{1}{n_1} \sum_{i:T_i=1} f(X_i)$$

for $j = 1, \ldots, p$, where $n_1/n = \tilde{E}(T)$. The weights $\{\hat{w}_i : T_i = 0, 1 \leq i \leq n\}$ and the above constraints are the same as in entropy balancing. The result can also be shown by comparing the weighting scheme in Hainmueller (2012) and the program analogous to (31)–(33), but corresponding to the calibration equations (5). From this connection, our regularized method also extends entropy balancing to allow box constraints similar to (18).

7.5. Augmented inverse probability weighting and related methods

The development in this article focuses on estimation of propensity scores to improve inverse probability weighted estimation, without invoking any outcome regression model. On the other hand, using both propensity score and outcome regression models may enhance efficiency and robustness and, particularly with high-dimensional data, enable valid confidence intervals to be constructed under suitable conditions. In the following, we briefly discuss related work in high-dimensional, parametric settings. See Robins et al. (2017) for estimation based on higher-order influence functions in nonparametric settings.

Belloni et al. (2014) and Farrell (2015) studied the augmented inverse probability weighted estimator (Robins et al., 1994), with propensity score and outcome regression models fitted by regularized maximum likelihood, and obtained valid confidence intervals when both the propensity score and the outcome regression models are correctly specified under suitable sparsistency conditions. In a 2017 Cornell University technical report, Ning, Peng and Imai also used regularized maximum likelihood estimation of propensity scores and outcome regression functions, but applied an inverse probability weighted estimator with propensity scores further adjusted through calibration equations similar to (4), depending on variable selection from the initial regularized estimation. Similar results to those of Belloni et al. (2014) and Farrell (2015) were obtained. Athey et al. (2018) combined penalized estimation of a linear outcome model and construction of balancing weights, similarly to Zubizarreta (2015), and obtained valid confidence intervals in the situation where a linear outcome model is correctly specified. In contrast with these works, Tan (2019a) developed regularized calibrated estimation for outcome regression in conjunction with that for propensity scores as studied here to achieve model-assisted inference, that is, valid confidence intervals in cases where the propensity score model is correctly specified, but the outcome regression model may be misspecified. With linear outcome models, the confidence intervals are also doubly robust.
ACKNOWLEDGEMENT

The author thanks Cun-Hui Zhang for helpful discussions, as well as the editor, an associate editor and a referee for constructive comments which have led to various improvements of the article. This research was supported in part by the Patient-Centred Outcomes Research Institute.

SUPPLEMENTARY MATERIAL

Supplementary Material available at Biometrika online includes technical proofs and additional numerical results from the simulation study and empirical application.

REFERENCES

Athey, S., Imbens, G. W. & Wager, S. (2018). Approximate residual balancing: Debiasing inference of average treatment effects in high dimensions. *J. R. Statist. Soc. B* 80, 597–623.

Austin, P. C. & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statist. Med.* 34, 3661–79.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* 81, 608–50.

Bohning, D. & Lindsay, B. G. (1988). Monotonicity of quadratic approximation algorithms. *Ann. Inst. Statist. Math.* 40, 641–63.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. New York: Springer.

Chan, K. C. G., Yam, S. C. P. & Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Statist. Soc. B* 78, 673–700.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell Jr, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M. et al. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *J. Am. Med. Assoc.* 276, 889–97.

Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley.

Devile, J. C. & Sarndal, C. E. (1992). Calibration estimators in survey sampling. *J. Am. Statist. Assoc.* 87, 376–82.

Dudik, M., Phillips, S. J. & Schapire, R. E. (2007). Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *J. Mach. Learn. Res.* 8, 1217–60.

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

Folsom, R. E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. In *Proc. Am. Statist. Assoc. Social Statist. Sect.* Alexandria, Virginia: American Statistical Association, pp. 197–202.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* 33, 1–22.

Geyer, C. J. (2014). *Trust: Trust Region Optimization*. R package version 0.1-6, available at https://CRAN.R-project.org/package=trust.

Graham, B. S., de Xavier Pinto, C. C. & Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *Rev. Econ. Stud.* 79, 1053–79.

Hainmueller, J. (2012). Entropy balancing for causal effects: Multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* 20, 25–46.

Hirano, K. & Imbens, G. W. (2002). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Serv. Out. Res. Methodol.* 2, 259–78.

Huang, J. & Zhang, C.-H. (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *J. Mach. Learn. Res.* 13, 1839–64.

Imai, K. & Ratkovic, M. (2014). Covariate balancing propensity score. *J. R. Statist. Soc. B* 76, 243–63.

Kang, J. D. Y. & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with Discussion). *Statist. Sci.* 22, 523–39.

Kim, J. K. & Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statist. Sinica* 24, 375–94.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Manski, C. F. (1988). *Analog Estimation Methods in Econometrics*. New York: Chapman & Hall.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*. New York: Chapman & Hall, 2nd ed.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. & Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* 27, 538–57.
Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, Section 9. *Statist. Sci.* 5, 465–80.

Osborne, M., Presnell, B., & Turlach, B. (2000). A new approach to variable selection in least squares problem. *IMA J. Numer. Anal.* 20, 389–404.

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Robins, J. M., Li, L., Mukherjee, R., Tchetgen Tchetgen, E., & van der Vaart, A. (2017). Minimax estimation of a functional on a structured high dimensional model. *Ann. Statist.* 45, 1951–87.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* 89, 846–66.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Statist. Assoc.* 79, 516–24.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 688–701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63, 581–90.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Am. Statist. Assoc.* 101, 1619–37.

Tan, Z. (2007). Comment: Understanding OR, PS, and DR. *Statist. Sci.* 22, 560–8.

Tan, Z. (2010). Bounded, efficient, and doubly robust estimation with inverse weighting. *Biometrika* 97, 661–82.

Tan, Z. (2013). Simple design-efficient calibration estimators for rejective and high-entropy sampling. *Biometrika* 100, 399–415.

Tan, Z. (2019a). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Ann. Statist.* to appear, arXiv: 1801.09817.

Tan, Z. (2019b). *RCAL: Regularized Calibrated Estimation*. R package version 1.0, available at https://CRAN.R-project.org/package=RCAL.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* 58, 267–88.

Vermeulen, K. & Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *J. Am. Statist. Assoc.* 110, 1024–36.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–25.

Wu, T. T. & Lange, K. (2010). The MM alternative to EM. *Statist. Sci.* 25, 492–505.

Yiu, S. & Su, L. (2018). Covariate association eliminating weights: A unified weighting framework for causal effect estimation. *Biometrika* 105, 709–22.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Am. Statist. Assoc.* 110, 910–22.

[Received on 30 May 2018. Editorial decision on 15 May 2019]