LIKELIHOOD RATIO INFERENCE FOR MISSING DATA MODELS

KARUN ADUSUMILLI AND TAISUKE OTSU

Abstract. Missing or incomplete outcome data is a ubiquitous problem in biomedical and social sciences. Under the missing at random setup, inverse probability weighting is widely applied to estimate and make inference on the population objects of interest, but it is known that its performance can be poor in practical sample sizes. Recently, to overcome this problem, several alternative weighting methods have been proposed that directly balance the distributional characteristics of covariates. These existing balancing methods are useful for obtaining point estimates of the population objects. The purpose of this paper is to develop a new weighting scheme, based on Empirical Likelihood, that would be useful for conducting interval estimation or hypothesis testing. In particular, we propose re-weighting the covariate balancing weights so that the resulting objective function admits an asymptotic chi-square calibration. Our re-weighting method is naturally extended to inference on treatment effects, data combination models, and high-dimensional covariates. Simulation and empirical examples illustrate usefulness of the proposed method.

1. Introduction

Missing or incomplete outcome data is a ubiquitous problem in biomedical and social sciences (see, e.g., Little and Rubin, 2002, and Imbens and Rubin, 2015). Under the assumption of missing at random or selection on observable covariates, inverse probability weighting is widely applied to estimate and make inference on the population objects of interest, such as population means and average treatment effects (Horvitz and Thompson, 1952; Rosenbaum and Rubin, 1983; Rosenbaum, 1987; Robins, Rotnitzky and Zhao, 1994; Hirano, Imbens and Ridder, 2003, among many others). Although inverse probability or propensity score weighting balances pretreatment variables in the population, its performance in finite samples leaves much to be desired. Recently several alternative weighting methods that directly balance distributional characteristics of covariates have been proposed to overcome this problem (Hainmueller, 2012; Zubizarreta, 2015; Graham, Pinto and Egel, 2012, 2016; Athey, Imbens and Wager, 2016; Chan, Yam and Zhang, 2016). These methods, which are based on balancing weights, have been employed to obtain point estimates for the population objects of interest, and the above authors have reported desirable performances of the empirical balancing approach in various contexts.

The purpose of this paper is to develop a new weighting scheme that would be useful for conducting statistical inference (i.e., interval estimation and hypothesis testing) as opposed to point estimation. In particular, we propose to *re-weight* the covariate balancing weights so that the resulting objective function admits an asymptotic chi-square calibration. The basic idea is to capture the empirical likelihood increments when going from the baseline likelihood based on

This research was partly supported by the ERC Consolidator Grant (SNP 615882) (Otsu).

empirical balancing for point estimation, to the one obtained by adding the identifying moment conditions for objects of interest. Since our likelihood ratio statistic is asymptotically pivotal, the resulting confidence set circumvents estimation of the asymptotic variance, which typically involves several nonparametric components (e.g., conditional means and variances and the propensity score). Also the confidence set is range preserving and transformation respecting, and its shape is determined by the data.

Our re-weighting method for constructing asymptotically pivotal statistics can be naturally extended for providing inference on treatment effects, data combination models, and high-dimensional covariates. For treatment effects, we re-weight the empirical balancing weights of Hainmueller (2012) and Chan, Yam and Zhang (2016) to yield an asymptotically pivotal likelihood ratio statistic. Our approach is general enough to cover average and quantile treatment effects, among other quantities. For the data combination models, we consider the setup of Chen, Hong and Tarozzi (2008) and re-weight the balancing weights that approximate the odds ratio of the propensity scores. For high-dimensional covariates, we re-weight the approximate balancing weights of Zubizarreta (2015) and Wang and Zubizarreta (2017). Simulation and empirical examples illustrate usefulness of the proposed method.

This paper also contributes to the literature on empirical likelihood methods (see, Owen, 2001, for a survey). Qin and Zhang (2007) introduced the empirical likelihood approach to missing response problems with parametric propensity scores. Qin, Zhang and Leung (2009) proposed a unified empirical likelihood approach to missing data problems. We refer Qin (2017) for a comprehensive survey on the empirical likelihood methods, particularly in the context of missing and biased sampling problems.

This paper is organized as follows. Section 2 introduces the basic setup and empirical balancing approach. In Section 3, we develop the empirical likelihood ratio statistic. Section 4 discusses extensions to inferences on treatment effects, data combination models, high-dimensional covariates by approximate balancing, and overidentified models.

2. Setup and empirical balancing

Consider a sequence of random variables $\{Y_{1i}, X_i : \ i = 1, \ldots, N\}$. Here X_i is a d-dimensional vector of covariates that is observable for all $i = 1, \ldots, N$. On the other hand, we observe scalar Y_{1i} only for a limited selection of individuals. In particular, we observe $Y_i = Y_{1i}D_i$ for all $i = 1, \ldots, N$, where D_i is the selection indicator (taking the value of one if Y_{1i} is observable, and zero otherwise). We wish to conduct statistical inference on the p-dimensional vector of parameters, β, which is identified by the moment condition

$$E[\psi(Y_1, X, \beta)] = 0. \tag{1}$$

We focus on the just-identified case (i.e., ψ and β have the same dimension) and discuss an extension to the over-identified case in Section 4.4. Let $p(x) = P(D = 1|X = x)$ denote the propensity score. We impose the following assumptions.

Assumption. *(i)* $\{Y_{1i}, X_i, D_i : i = 1, \ldots, N\}$ *is an independent and identically distributed random sequence; (ii)* Y_1 *is independent of* D *given* X*; (iii) There exists* $\kappa > 0$ *such that* $\kappa \leq p(x) \leq 1$ *for all* $x \in \mathbb{R}^d$*.*

This setup encompasses a number of examples including missing at random, attrition in panel data, missing regressors, among others. In Section 4.1, we extend this setup to analyze treatment effects under unconfoundedness.

Under the above assumptions, we are interested in testing the parameter hypothesis $\mathrm{H}_0 : \beta = \beta_0$ against $\mathrm{H}_1 : \beta \neq \beta_0$ for some given value β_0. For the just-identified case, it follows from the results in Graham (2011) that this testing problem is equivalent to testing the 'identifying moment'

$$E\left[\frac{D}{p(X)}\psi(Y_1, X, \beta_0)\right] = 0, \tag{2}$$

subject to the 'auxiliary moment'

$$E\left[\frac{D}{p(X)} - 1 \middle| X\right] = 0. \tag{3}$$

This setup has been widely studied and many estimation and inference methods on β have been proposed (see, e.g., Graham, 2011, and references therein). One of the recent promising proposals is the empirical balancing (EB) approach (Hainmueller, 2012; Zubizarreta, 2015; Chan, Yam and Zhang, 2016). The basic idea of EB methods is to find and utilize weights to approximate the inverse of the unknown propensity score so that the parameters β can be estimated by the method of moments from (2) with the EB weights.

More precisely, let $\{q_j(x) : j = 1, \ldots, K\}$ denote a series of approximating or basis functions, such as splines, power series, and Fourier series. As K increases, information contained in the auxiliary moment (3) can be eventually captured by the growing vector of unconditional moments

$$E\left[\left(\frac{D}{p(X)} - 1\right) q^K(X)\right] = 0, \tag{4}$$

where $q^K(x) = (q_1(x), \ldots, q_K(x))'$. Suppose the observations are ordered in such a way that the first n observations correspond to the respondents (i.e., $D_i = 1$ for $i = 1, \ldots, n$ and 0 for $i = n+1, \ldots, N$). Then the EB weights $(\gamma_1, \ldots, \gamma_n)$ for the respondents can be obtained as solutions of

$$\min_{t_1, \ldots, t_n} \sum_{i=1}^{n} \log t_i \quad \text{s.t.} \quad \frac{1}{N}\sum_{i=1}^{n} t_i q^K(X_i) = \frac{1}{N}\sum_{i=1}^{N} q^K(X_i). \tag{5}$$

The EB weights provide a weighting scheme that exactly satisfies the unconditional moments in (4) and may be considered as an approximation to the reciprocal of the propensity score. Based on the EB weights, a point estimator of β can be obtained by solving the estimating equation $\sum_{i=1}^{n} \gamma_i \psi(Y_{1i}, X_i, \hat{\beta}) = 0$. Chan, Yam and Zhang (2015) extended this idea to estimation of the average treatment effects (see Section 4.1) and established global semiparametric efficiency of their estimator. An interesting feature of this estimation strategy is that it does not involve explicit preliminary modeling or estimation of the propensity score function.

The construction of the EB weights above is closely related to empirical likelihood (Owen, 1988). Chan, Yam and Zhang (2016) extended this construction to a general divergence family, which is analogous to the generalized empirical likelihood (Newey and Smith, 2004). However, the EB weights in (5) are designed to impose the auxiliary moment conditions in (4) but not to test the identifying moment (2) of interest. Also since Y_1 can be missing, we cannot directly construct empirical likelihood from the identifying moment itself. In this paper, we propose a new inference method to *re-weight* the EB weights in (5) so that the likelihood ratio statistic will converge to the chi-square limiting distribution.

3. Likelihood ratio statistic

We now present our likelihood ratio statistic. The basic idea is to capture the likelihood increments between the baseline likelihood in (5) to the one obtained by adding the identifying moment condition (2). To this end, we consider the following maximization problem

$$\ell(\beta_0) = \max_{w_1,\ldots,w_N} \sum_{i=1}^N \log(Nw_i),$$

$$\text{s.t.} \quad \sum_{i=1}^N w_i = 1, \quad \sum_{i=1}^N w_i(\gamma_i D_i - 1)q^K(X_i) = 0, \quad \sum_{i=1}^N w_i \gamma_i D_i \psi(Y_{1i}, X_i, \beta_0) = 0. \tag{6}$$

Here $(\gamma_1, \ldots, \gamma_n)$ are the EB weights obtained in (5). For $(\gamma_{n+1}, \ldots, \gamma_N)$ (for non-respondents), we set zeros, but they can be any values. Note that without the last condition in (6) on the identifying moment condition, the above maximization problem is solved by uniform weights $w_i = N^{-1}$ for all $i = 1, \ldots, N$ (because of (5)). Therefore, the above maximum $\ell(\beta_0)$ indeed corresponds to the likelihood increment by adding the last condition in (6). Letting $g_i^K = (\gamma_i D_i \psi(Y_{1i}, X_i, \beta_0)', \ (\gamma_i D_i - 1)q^K(X_i)')$, the dual form of $\ell(\beta_0)$ is written as

$$\ell(\beta_0) = -\min_{\lambda} \sum_{i=1}^N \log(1 + \lambda' g_i^K). \tag{7}$$

In practice, we employ this dual form to implement our inference method. The asymptotic distribution of the likelihood ratio statistic $\ell(\beta_0)$ under $\mathrm{H}_0 : \beta = \beta_0$ is obtained as follows.

Theorem. *Suppose Assumptions (i)-(iii) hold true. Furthermore, assume (iv) the support of X is a Cartesian product of compact intervals, and $E[Y_1^2] < \infty$; (v) $p(x)$ is s-times continuously differentiable with $s > 13d$, and $E[\psi(Y_1, X, \beta_0)|X = x]$ is s_1-times continuously differentiable with $s_1 > \frac{3d}{2}$; and (vi) $K = O(N^a)$ with $\left(\frac{s}{d} - 2\right)^{-1} < a < \frac{1}{11}$. Then under $\mathrm{H}_0 : \beta = \beta_0$,*

$$-2\{\ell(\beta_0) - N \log N\} \xrightarrow{d} \chi_p^2, \quad \text{as } N \to \infty.$$

Our assumptions are analogous to those in Chan, Yam and Zhang (2016). The theorem says that our likelihood ratio statistic is asymptotically pivotal and converges to the chi-square distribution under the null hypothesis. Based on this result, the $100(1-\alpha)\%$ asymptotic confidence set for β can be given by $\{\beta : -2\{\ell(\beta_0) - N \log N\} \leq \chi_{p,\alpha}^2\}$, where $\chi_{p,\alpha}^2$ is the $(1-\alpha)$-th quantile of the χ_p^2 distribution. Furthermore, it is straightforward to extend this theorem for testing the null $\mathrm{H}_0 : r(\beta) = \theta_0$ for a possibly nonlinear function $r : \mathbb{R}^p \to \mathbb{R}^{p_1}$ with $p_1 \leq p$. In this case, the

likelihood ratio statistic is obtained by $-2\{\max_{\beta:r(\beta)=\theta_0} \ell(\beta) - N \log N\}$, which can be shown to converge to the $\chi^2_{p_1}$ distribution.

Although we focus on the likelihood ratio statistic in (6), it is relatively straightforward to extend our result to statistics defined by a general divergence family (say, $D(N, w)$ instead of $\log(Nw)$) as in Chan, Yam and Zhang (2016). For example, if D is the power divergence family of Cressie and Read (1984), then the resulting test statistics are asymptotically equivalent to ours.

The proof of this theorem (see, (11) and (12) below) indicates that our likelihood ratio statistic has the same local power function as the Wald or t-test based on the globally semiparametric efficient estimator as derived by Graham (2011). However, in contrast to the Wald test, we circumvent estimation of the asymptotic variance V, which typically involves estimating the conditional mean $E[\psi(Y_1, X, \beta_0)|X]$, the conditional variance $Var(\psi(Y_1, X, \beta_0)|X)$, and the propensity score $p(X)$.

4. Extensions

4.1. **Treatment effects.** It is straightforward to extend our likelihood ratio construction to conduct inference on various measures of treatment effects under unconfoundedness. Let Y_1 and Y_0 be potential outcomes associated with a binary treatment variable D. The observed outcome is $Y = DY_1 + (1 - D)Y_0$. Also let X be a vector of covariates. and $p(X) = P(D = 1|X)$ be the propensity score. Our setup covers inference on parameters β identified as a solution to the moment condition:

$$E[\psi_1(Y_1, X, \beta) - \psi_0(Y_0, X, \beta)] = 0,$$

where ψ_1 and ψ_0 have the same dimension as β. This setup accommodates many popular inferential problems as special cases. For example, if we set $\psi_1(Y_1, X, \beta) = Y_1 - \beta$ and $\psi_0(Y_0, X, \beta) = Y_0$, then β is the average treatment effect. Again, we consider the testing problem $\mathrm{H}_0 : \beta = \beta_0$ against $\mathrm{H}_1 : \beta \neq \beta_0$, which is equivalent to testing the identifying moment

$$E\left[\frac{D}{p(X)}\psi_1(Y, X, \beta_0) - \frac{1 - D}{1 - p(X)}\psi_0(Y, X, \beta_0)\right] = 0, \tag{8}$$

subject to the auxiliary moments

$$E\left[\frac{D}{p(X)} - 1\middle| X\right] = 0 \quad \text{and} \quad E\left[\frac{1 - D}{1 - p(X)} - 1\middle| X\right] = 0.$$

The second auxiliary moment is needed to provide the balance constraints for the control sample.

We shall order the observations in such a way that the first n terms correspond to treated observations (i.e., $D_i = 1$ for $i = 1, \ldots, n$ and 0 for $i = n + 1, \ldots, N$). Following Hainmueller (2012) and Chan, Yam and Zhang (2016), we can obtain the EB weights $(\bar{\gamma}_1, \ldots, \bar{\gamma}_N)$ as the solution of

$$\min_{t_1,\ldots,t_N} \sum_{i=1}^{N} \log t_i,$$

$$\text{s.t.} \quad \frac{1}{N}\sum_{i=1}^{N} t_i D_i q^K(X_i) = \frac{1}{N}\sum_{i=1}^{N} q^K(X_i), \quad \frac{1}{N}\sum_{i=1}^{N} t_i (1 - D_i) q^K(X_i) = \frac{1}{N}\sum_{i=1}^{N} q^K(X_i).$$

Here the weights $(\bar{\gamma}_1, \ldots, \bar{\gamma}_n)$ asymptotically approximate $p(X)^{-1}$, while the the remainder of the weights $(\bar{\gamma}_{n+1}, \ldots, \bar{\gamma}_N)$ asymptotically approximate $(1 - p(X))^{-1}$. It is implicitly assumed that the vector of functions $q^K(\cdot)$ includes the constant function. Hainmueller (2012) and Chan, Yam and Zhang (2016) employed these EB weights to obtain an estimator that attains the semi-parametric efficiency bound. Here, we instead re-weight the EB weights to obtain a likelihood ratio statistic for the identifying moment in (8).

In particular, our likelihood ratio statistic can be constructed as

$$\bar{\ell}(\beta_0) = \max_{w_1,\ldots,w_N} \sum_{i=1}^{N} \log(N w_i),$$

$$\text{s.t.} \quad \sum_{i=1}^{N} w_i = 1, \quad \sum_{i=1}^{N} w_i(\bar{\gamma}_i D_i - 1) q^K(X_i) = 0, \quad \sum_{i=1}^{N} w_i(\bar{\gamma}_i(1 - D_i) - 1) q^K(X_i) = 0,$$
$$\sum_{i=1}^{N} w_i \bar{\gamma}_i \{D_i \psi_1(Y_i, X_i, \beta_0) - (1 - D_i)\psi_0(Y_i, X_i, \beta_0)\} = 0.$$

The dual form of $\bar{\ell}(\beta_0)$ is obtained in the same manner as $\ell(\beta_0)$. Also, under analogous conditions to the ones in Theorem, it can be shown that $-2\{\bar{\ell}(\beta_0) - N \log N\} \xrightarrow{d} \chi^2_p$ under H_0, where p is the dimension of β_0. Similar comments to the ones for Theorem apply. Again, our likelihood ratio statistic is asymptotically pivotal, and is free from variance estimation. If we are interested in some p_1-dimensional function $r(\beta)$ (e.g., quantile treatment effects), the likelihood ratio statistic can be modified as $-2\{\max_{\beta: r(\beta) = \theta_0} \bar{\ell}(\beta) - N \log N\} \xrightarrow{d} \chi^2_{p_1}$.

4.2. **Data combination models.** Data combination models are another interesting class of missing data models. Our inferential results can be extended to cover these as well. Let $Z = (W_1, W_0, X')'$ denote a vector of random variables from a study population. We are interested in conducting inference for the p-dimensional vector of parameters, β, which is just-identified by the moment condition

$$E_s[\psi(Z, \beta)] = 0,$$

where $E_s[\cdot]$ denotes the expectation under the study sample. However we do not observe the entire vector Z. Rather, we only observe N_s measurements of $(W_1, X')'$ from the study sample, but we have access to N_a measurements of $(W_0, X')'$ drawn from an auxiliary sample. Thus the variables X are common to the both samples.

We shall assume that the conditional distribution of W_0 given X is the same in the both samples (however the marginal distributions of X may differ). Also, we assume that the support of X in the auxiliary sample is at least as large as the study sample. Under these conditions, Chen, Hong and Tarozzi (2008) showed that the parameter β is identified as long as $\psi(\cdot)$ is separable in W_1 and W_0 in the sense that

$$\psi(W_1, W_0, X, \beta) = \psi_s(W_1, X, \beta) - \psi_a(W_0, X, \beta),$$

for some $\psi_s(\cdot)$ and $\psi_a(\cdot)$.

Following Graham, Pinto and Egel (2016), we employ a multinomial sampling framework by assuming that a unit is drawn at random from the distribution of the study sample with probability π, and from that of the auxiliary sample with probability $1-\pi$. Let D denote a binary random variable that takes value 1 when the observation is in the study sample and 0 when it is in the auxiliary sample. Under this framework we can treat the 'merged' realization $(D_i, X_i, D_i W_{1i}, (1-D_i)W_{0i})$ as a random draw from a synthetic 'merged' population (Graham, Pinto and Egel, 2016). Let $P(\cdot)$ and $E[\cdot]$ denote the probability and expectation, respectively, in this merged population. The propensity score in this context is then defined as $p(x) = P(D=1|X=x)$. The support condition above assures existence of some $\kappa > 0$ such that $\kappa \leq p(x) \leq 1$ for all $x \in \mathbb{R}^d$. Importantly, we do not place any functional form assumptions on the propensity score, apart from some smoothness assumptions.[1] Finally, let $W = DW_1 + (1-D)W_0$ denote the observed 'outcome' variable.

The above framework covers many important statistical problems including estimation of the average treatment effect on the treated (ATT), two-sample instrumental variables (Angrist and Krueger, 1992), counterfactual distributions (Dinardo, Fortin and Lemieux, 1996), poverty maps (Tarozzi, 2007), semiparametric differences-in-differences (Abadie, 2005), and models with mismeasured regressors in the presence of validation samples (Carroll and Wand, 1991). We refer to Graham, Egel and Pinto (2016) for further discussion.

As before, we consider the testing problem $\mathrm{H}_0 : \beta = \beta_0$ against $\mathrm{H}_1 : \beta \neq \beta_0$, which is equivalent to test the identifying moment

$$E\left[D\psi_s(W, X, \beta_0) - (1-D)\frac{p(X)}{1-p(X)}\psi_a(W, X, \beta_0)\right] = 0,$$

subject to the auxiliary moment

$$E\left[D - (1-D)\frac{p(X)}{1-p(X)}\middle| X\right] = 0,$$

(see, Graham, 2011, and Zhao, 2018). Let $N = N_a + N_s$. We shall order the observations such that the first N_a terms correspond to the auxiliary sample (i.e., $D_i = 0$ for $i = 1, \ldots, N_a$ and 1 for $i = N_a + 1, \ldots, N$). The EB weights $(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_{N_a})$ for data combination models are obtained as the solution of

$$\min_{t_1,\ldots,t_{N_a}} \sum_{i=1}^{N_a} \log t_i \quad \text{s.t.} \quad \frac{1}{N}\sum_{i=1}^{N} D_i q^K(X_i) = \frac{1}{N}\sum_{i=1}^{N} t_i(1-D_i)q^K(X_i).$$

Here the weights $(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_{N_a})$ asymptotically approximate $p(X)/(1-p(X))$. For $i = N_a + 1, \ldots, N$, we set $\tilde{\gamma}_i = 1$. In this case, our likelihood ratio statistic is obtained as

$$\tilde{\ell}(\beta_0) = \max_{w_1,\ldots,w_N} \sum_{i=1}^{N} \log(N w_i),$$

[1] Note that unlike the 'standard' missing data and treatment effect models considered in the previous sections, knowledge of the propensity score is in fact asymptotically useful in this context. Still, our test procedure is first order (semi-parametrically) efficient under the regime of unknown form of the propensity score.

$$\text{s.t.} \qquad \sum_{i=1}^{N} w_i = 1, \quad \sum_{i=1}^{N} w_i \{D_i - \tilde{\gamma}_i(1-D_i)\} q^K(X_i) = 0,$$
$$\sum_{i=1}^{N} w_i \tilde{\gamma}_i \{D_i \psi_s(W_i, X_i, \beta_0) - (1-D_i)\psi_a(W_i, X_i, \beta_0)\} = 0.$$

The dual form of $\tilde{\ell}(\beta_0)$ is obtained in the same manner as $\ell(\beta_0)$. Also under analogous conditions to the ones in Theorem, it can be shown that $-2\{\tilde{\ell}(\beta_0) - N \log N\} \xrightarrow{d} \chi^2_p$ under H_0, where p is the dimension of β_0.

4.3. **Approximate balancing weights.** Consider again the setup in Section 2. When the dimension of the auxiliary moments in (4) is large (e.g., due to high-dimensional covariates), exact balancing of the auxiliary moments is typically infeasible. For such cases, Zubizarreta (2015) proposed approximate balancing weights as a solution to the following optimization problem:

$$\min_{t_1,\ldots,t_n} \sum_{i=1}^{n} \log t_i \quad \text{s.t.} \quad \left| \frac{1}{N} \sum_{i=1}^{n} t_i q^K(X_i) - \frac{1}{N} \sum_{i=1}^{N} q^K(X_i) \right| \leq \delta_N,$$

where δ_N is a positive tuning constant chosen by the practitioner for approximate balancing. Wang and Zubizarreta (2017) showed that the dual of the above maximization is given by

$$\min_{\omega} \frac{1}{N} \sum_{i=1}^{N} \{D_i \rho(\omega' q^K(X_i)) - N^{-1} \omega' q^K(X_i)\} + \delta_N \left\|\omega\right\|_1,$$

where $\rho(x) = \exp(1+x)$. Letting $\hat{\omega}$ be the minimizer of this problem, the approximate balancing weights are obtained as $\rho'(\hat{\omega}' q^K(X_i))$ for $i = 1, \ldots, N$. Thus the weights are obtained as solutions to a constrained L_1 minimization problem.

Define

$$\omega^* = \arg\min_{\omega} E[D_i \rho(\omega' q^K(X_i)) - N^{-1} \omega' q^K(X_i)] + \delta_N \left\|\omega\right\|_1.$$

Suppose that the sparsity assumption $\left\|\omega^*\right\|_0 \leq s_N$ holds for some s_N slowly growing to ∞. In this context, sparsity essentially implies that it suffices to balance only a finite (or slowly growing) subset of basis functions from $q^K(\cdot)$ in order to obtain good balancing weights. Using the approximate balancing weights directly for inference is however not ideal since these suffer from substantial bias due to the shrinkage implicit in the L_1 minimization for ω. To alleviate this bias, we propose adjusting these weights using a post-LASSO type correction.

Let $\hat{S} = \{j : \hat{\omega}_j \neq 0\}$ denote the set of indices corresponding to the non-zero elements of $\hat{\omega}$. We then extract the subset of basis functions $q^K(\cdot)$ corresponding to $\hat{S}$, which is denoted by $q^K_{\hat{S}}(\cdot)$. The post-LASSO approximate balancing weights $(\gamma^A_1, \ldots, \gamma^A_n)$ are obtained as the solution of

$$\min_{t_1,\ldots,t_n} \sum_{i=1}^{n} \log t_i \quad \text{s.t.} \quad \frac{1}{N} \sum_{i=1}^{n} t_i q^K_{\hat{S}}(X_i) = \frac{1}{N} \sum_{i=1}^{N} q^K_{\hat{S}}(X_i).$$

The likelihood ratio statistic is then constructed analogously to Section 2.

$$\ell^A(\beta_0) = \max_{w_1,\ldots,w_N} \sum_{i=1}^{N} \log(N w_i),$$

$$\text{s.t.} \quad \sum_{i=1}^{N} w_i = 1, \quad \sum_{i=1}^{N} w_i(\gamma_i^A D_i - 1) q_{\hat{S}}^K(X_i) = 0, \quad \sum_{i=1}^{N} w_i \gamma_i^A D_i \psi(Z_i, \beta_0) = 0.$$

The dual form of $\ell^A(\beta_0)$ is obtained in the same manner as $\ell(\beta_0)$. Also under analogous conditions to the ones in Theorem, it can be shown that $-2\{\ell^A(\beta_0) - N \log N\} \xrightarrow{d} \chi_p^2$ under H_0, where p is the dimension of β_0. Again, our likelihood ratio statistic is asymptotically pivotal and is free from variance estimation.

4.4. **Over-identified models.** Thus far we have considered inference under just-identification. In some applications however, the parameters β could be over-identified (e.g., moment conditions with side information, and instrumental variable models with more instruments than regressors). While our testing procedure still controls size in such contexts, it is no longer first-order efficient. In this section we show how it can be modified to recover efficiency.

Consider the missing data setup in Section 2. Suppose now that the dimension p_1 of the moment function $\psi(\cdot)$ is greater than p, the dimension of β. Then we can construct a likelihood ratio test by considering the discrepancy in the log-likelihoods evaluated at the estimated and hypothesized values of β. In particular, based on the likelihood ratio statistic in (6), the likelihood ratio test statistic for testing $\mathrm{H}_0 : \beta = \beta_0$ against $\mathrm{H}_1 : \beta \neq \beta_0$ is given by

$$\ell^R(\beta_0) = -2 \left\{ \min_{\beta} \ell(\beta) - \ell(\beta_0) \right\}.$$

Under analogous conditions to the Theorem in Section 3, it can be shown that $\ell^R(\beta_0) \xrightarrow{d} \chi_p^2$ under H_0. Note that the degree of freedom of the limiting distribution is p, the dimension of β. On the other hand, the statistic $\ell(\beta_0)$ converges to the chi-square distribution with degree of freedom p_1, the dimension of ψ.

5. Simulation

In this section, we study the finite sample performance of the proposed likelihood ratio test.

We consider three different data generating processes (DGPs). The first DGP (DGP1) is taken from Abadie and Imbens (2016, Supplementary material), adapted for the case of missing data. We generate a two dimensional vector (X_1, X_2) of covariates by drawing both variables from a uniform $[-1/2, 1/2]$ distribution independently of each other. The 'true' outcome variable is generated as $Y_1 = 5 + 2X_1 + 4X_2 + U$, where U is a standard normal random variable. The propensity score is given by the logistic function

$$p(X) \equiv P(D = 1|X) = \frac{\exp(X_1 + tX_2)}{1 + \exp(X_1 + tX_2)}, \tag{9}$$

and the treatments are generated as $W \sim \text{Bernoulli}(p(X))$. For DGP1, we choose $t = 2$. Finally, the observed outcome variables are generated as $Y = DY_1$.

The second DGP (DGP2) differs from DGP1 only in the choice of t: in particular, we set $t = 4$. The effect of increasing t is to reduce the amount of overlap in the propensity score (i.e., we weaken Assumption (iii)).

TABLE 1. Rejection frequencies under the null for inference on β_a

		$N = 100$		$N = 200$		$N = 500$		$N = 5000$	
		$K = 3$	$K = 5$	$K = 3$	$K = 5$	$K = 3$	$K = 5$	$K = 3$	$K = 5$
DGP1	LR	0.067	0.077	0.056	0.062	0.057	0.063	0.048	0.054
	Wald	0.064	0.074	0.057	0.062	0.058	0.063	0.047	0.054
DGP2	LR	0.092	0.119	0.074	0.082	0.054	0.066	0.059	0.056
	Wald	0.098	0.118	0.076	0.085	0.056	0.066	0.058	0.056
DGP3	LR	0.105	0.147	0.083	0.102	0.079	0.065	0.126	0.057
	Wald	0.143	0.226	0.102	0.122	0.091	0.071	0.129	0.058

For the third DGP (DGP3), we generate a two dimensional vector (X_1, X_2) of covariates by drawing both variables (independently of each other) as $X_a \sim 2V - 1$ for $a = 1, 2$, with $V \sim$ beta$(2, 4)$. The unobserved outcome variable is generated as $Y_1 = 5+X_1^2+X_2^2+U$. The remainder of DGP3 follows the same construction as DGP1, i.e., we obtain the propensity scores by setting $t = 2$ in (9). Compared to DGP1, the distribution of Y is more asymmetric due to the use of an asymmetric beta distribution for the covariates. Additionally, the underlying model for Y_1 is more non-linear.

We first consider inference on the average outcome $\beta_a = E[Y_{1i}]$. Table 1 reports the performance of the likelihood ratio (LR) procedure for inference on β_a under the null, along with inference based on the Wald statistic using the variance estimate proposed by Chan, Yam and Zhang (2016). The nominal significance level is 0.05. For all DGPs, we report results with $K = 3$, corresponding to $q^K(X) = (1, X)$, and $K = 5$, corresponding to $q^K(X) = (1, X, X^2)$. All the simulation results are based on 2,500 Monte Carlo repetitions.

From Table 1, we can observe that when the DGP is linear and the overlap is good (e.g., DGP1 and DGP2), both the LR and Wald procedures behave very similarly. However, when there is non-linearity in the covariates and the underlying distribution is asymmetric, the LR procedure provides more accurate inference, as seen in the simulation results for DGP3.

Next we consider inference on the median outcome $\beta_m = \text{median}(Y_{1i})$. Here the 'identifying' moment condition for β_m is given by $E[1(Y_i < \beta_m) - 0.5] = 0$. Table 2 reports the performance of the LR procedure for inference on β_m under the null for all the DGPs. Again the nominal significance level is 0.05. The LR procedure provides excellent size control for all DGPs, with the proviso that one employs a proper choice of K for DGP3. Note also that the Wald statistic is difficult to obtain here due to the complicated nature of the variance estimate for quantile estimators; indeed we are not aware of any variance estimate that has been proposed for this context.

6. Real data example

We illustrate our inferential procedure by applying it on data taken from the influential study of Card and Krueger (1994). These authors were interested in studying the effect of the raise,

TABLE 2. Rejection frequencies under the null for inference on β_m

	$N = 100$		$N = 200$		$N = 500$		$N = 5000$	
	$K = 3$	$K = 5$	$K = 3$	$K = 5$	$K = 3$	$K = 5$	$K = 3$	$K = 5$
DGP1	0.056	0.067	0.058	0.052	0.048	0.047	0.056	0.053
DGP2	0.072	0.098	0.054	0.076	0.058	0.055	0.091	0.063
DGP3	0.128	0.206	0.097	0.101	0.069	0.080	0.088	0.049

in 1993, of New Jersey's state minimum wage on employment. To this end, they collected data on employment in fast food restaurants in New Jersey and neighboring Pennsylvania, following the minimum wage hike. The restaurants in Pennsylvania, which did not witness a change in the minimum wage, form the control group. While the original study was based on a differences-in-differences design, later authors including Rosenbaum (2002) and Imbens and Rubin (2015) re-analyzed the data as if it arose from an unconfoundedness assumption, i.e., conditional on covariates, the probability of being treated (i.e., being from New Jersey as opposed to Pennsylvania) does not depend on the potential outcomes. Subsequently, our results in this section are based on the latter assumption.

The data consist of 273 restaurants from New Jersey (treated units), and 67 from Pennsylvania (control units). The covariate data consist of the following pre-treatment variables: number of employed in each restaurant prior to minimum wage hike (empft), starting wages (wage_st), average duration for the first raise (inctime), and indicators for the identity of the chain: (burger king, kfc, roys, wendys). The outcome (Y) is the number of employed in each restaurant after the increase in minimum wage (part time employees are weighted by 0.5). Our parameter of interest, β_0, is the average treatment effect on employment levels due to the minimum wage hike.

To provide inference on β_0, we consider two empirical balancing schemes: one where we only balance a single covariate, empft, i.e., $q^K(X) = (1, \text{empft})$, corresponding to $K = 2$; and the other where we balance all the covariates X, i.e., $q^K(X) = X$, corresponding to $K = 7$. The first scheme in particular is based on the analysis of Imbens and Rubin (2015) who found that empft was the only variable selected by their iterative balance checking algorithm for inclusion in the propensity score. Table 3 presents 90 and 95% confidence regions for β_0 based on our inferential procedure, along with the Wald confidence regions. We also report the estimates, $\hat{\beta}$, of β_0 under both $K = 2$ and 7. Both values are very close to the estimate of $\hat{\beta}_m = 0.84$ obtained by Imbens and Rubin (2015) using matching.

TABLE 3. Confidence regions for β_0 using Likelihood Ratio and Wald procedures

	$K=2$		$K=7$	
Estimate	$\hat{\beta}=0.840$		$\hat{\beta}=0.873$	
	90% CI	95% CI	90% CI	95% CI
LR	$[-0.782,\ 2.382]$	$[-1.110,\ 2.682]$	$[-0.608,\ 2.262]$	$[-0.909,\ 2.527]$
Wald	$[-0.766,\ 2.445]$	$[-1.073,\ 2.753]$	$[-0.590,\ 2.335]$	$[-0.870,\ 2.615]$

APPENDIX A. PROOF OF THEOREM

Following the arguments of Owen (1988), we can establish a quadratic expansion of the dual form in (7) as

$$-2\{\ell(\beta_0) - N\log N\} = \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g_i^K\right)\left(\frac{1}{N}\sum_{i=1}^{N} g_i^K g_i^{K\prime}\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g_i^K\right) + o_p(1). \tag{10}$$

Let $h_i^K = (\gamma_i D_i - 1)q^K(X_i)$. Note that $\sum_{i=1}^N h_i^K = 0$ by (5). Using the definition $g_i^K = \left(\gamma_i D_i \psi(Y_{1i}, X_i, \beta_0)', (h_i^K)'\right)'$, the first term in the right hand side of (10) can be written as

$$\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\gamma_i D_i\psi(Y_{1i}, X_i, \beta_0)\right)\hat{V}^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\gamma_i D_i\psi(Y_{1i}, X_i, \beta_0)\right),$$

where $\hat{V} = \hat{V}_0 - \Psi' Q(Q'Q)^{-1}Q'\Psi$ with $Q = N^{-1/2}(h_1^K, \ldots, h_N^K)'$ and

$$\begin{aligned}\Psi &= N^{-1/2}\left(\gamma_1 D_1\psi(Y_{1i}, X_i, \beta_0), \ldots, \gamma_N D_N\psi(Y_{1N}, X_N, \beta_0)\right)',\\ \hat{V}_0 &= \frac{1}{N}\sum_{i=1}^{N}\gamma_i^2 D_i\psi(Y_{1i}, X_i, \beta_0)\psi(Y_{1i}, X_i, \beta_0)'.\end{aligned}$$

Thus, it is sufficient to show that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\gamma_i D_i\psi(Y_{1i}, X_i, \beta_0) \xrightarrow{d} N(0, V), \tag{11}$$

$$\hat{V} \xrightarrow{p} V = E\left[\frac{\Sigma_0(X)}{p(X)} + \Gamma(X)\Gamma(X)'\right], \tag{12}$$

where $\Sigma_0(x) = Var(\psi(Y_1, X, \beta_0)|X = x)$ and $\Gamma(x) = E[\psi(Y_1, X, \beta_0)|X = x]$.

Equation (11) follows by analogous arguments as in the proof of Chan, Yam and Zhang (2016, Theorem 1).

It now remains to prove (12). Let $\tilde{V}$ denote the variance estimate when γ_i is replaced with $p(X_i)^{-1}$, i.e., $\tilde{V} \equiv \tilde{V}_0 - \tilde{\Psi}' Q(Q'Q)^{-1}Q'\tilde{\Psi}$ with

$$\begin{aligned}\tilde{\Psi} &= N^{-1/2}\left(\frac{D_1}{p(X_1)}\psi(Y_{1i}, X_i, \beta_0), \ldots, \frac{D_N}{p(X_N)}\psi(Y_{1N}, X_N, \beta_0)\right)',\\ \tilde{V}_0 &= \frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{p(X_i)^2}\psi(Y_{1i}, X_i, \beta_0)\psi(Y_{1i}, X_i, \beta_0)'.\end{aligned}$$

Now by analogous arguments as in Chan, Yam and Zhang (2016, Lemma 1), we have

$$\max_{1\leq i\leq N}\left|\gamma_i - \frac{1}{p(X_i)}\right| = O_p(N^{-a}),$$

for some $a > 0$. Consequently, it follows $\hat{V} - \tilde{V} = o_p(1)$. Hence it suffices for (12) to show that

$$\tilde{V} \xrightarrow{p} V. \tag{13}$$

As means of proving (13), consider estimation of the parameter τ under the moment restrictions $E\left[\frac{D}{p(X)}\psi(Y_1, X, \beta_0) - \tau\right] = 0$ and $E\left[\frac{D}{p(X)} - 1\middle| X\right] = 0$. Clearly the true value of τ under H_0 is 0. Furthermore, V is the semiparametric efficiency bound for the estimation of τ under the previous moment restrictions (see, e.g., Graham, 2011). Additionally, for any given K consider the following (unconditional) moment function

$$m_i^K(\tau) = \begin{pmatrix} \frac{D_i}{p(X_i)}\psi(Y_{1i}, X_i, \beta_0) - \tau \\ \left(\frac{D_i}{p(X_i)} - 1\right) q^K(X_i) \end{pmatrix}.$$

Let V_K denote the variance of the efficient Generalized Method of Moments (GMM) estimator for τ using the moment condition $E[m_i^K(\tau)] = 0$. Now by employing the usual arguments for sieve estimators, it is possible to show that $V_K - V \to 0$ as $K \to \infty$. At the same time, $\tilde{V}$ is the sample counterpart of V_K, hence under the regularity conditions in the theorem, it can be shown $\tilde{V} - V_K = o_p(1)$. Combining these results, the conclusion follows.

References

[1] Abadie, A. and G. W. Imbens (2016) Matching on the estimated propensity score, *Econometrica*, 84, 781-807.

[2] Athey, S., Imbens, G. W. and S. Wager (2016) Approximate residual balancing: de-biased inference of average treatment effects in high dimensions, Working paper.

[3] David, C. and A. B. Krueger (1994) Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania, *American Economic Review*, 84(4), 772-793.

[4] Chan, K. C. G., Yam, S. C. P. and Z. Zhang (2016) Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting, *Journal of the Royal Statistical Society*, B 78, 673-700.

[5] Chen, X., Hong, H. and A. Tarozzi (2008) Semiparametric efficiency in GMM models with auxiliary data, *Annals of Statistics*, 36, 808-843.

[6] Cressie, N. and R. C. Read (1984) Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society*, B 46, 440-464.

[7] Graham, B. S. (2011) Efficiency bounds for missing data models with semiparametric restrictions, *Econometrica*, 79, 437-452.

[8] Graham, B., Pinto, C. and D. Egel (2012) Inverse probability tilting for moment condition models with missing data, *Review of Economic Studies*, 79, 1053-1079.

[9] Graham, B. S., Pinto, C. and D. Egel (2016) Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST), *Journal of Business and Economic Statistics*, 34, 288-301.

[10] Hainmueller, J. (2012) Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies, *Political Analysis*, 20, 25-46.

[11] Hirano, K., Imbens, G. W. and G. Ridder (2003) Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica*, 71, 1161-1189.

[12] Horvitz, D. G. and D. J. Thompson (1952) A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, 47, 663-685.

[13] Imbens, G. W. and D. B. Rubin (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

[14] Little, R. J. A. and D. B. Rubin (2002) *Statistical Analysis with Missing Data*, Wiley.

[15] Newey, W. K. and R. J. Smith (2004) Higher order properties of GMM and generalized empirical likelihood estimators, *Econometrica*, 72, 219-255.

[16] Owen, A. B. (1988) Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, 75, 237-249.

[17] Owen, A. B. (2001) *Empirical Likelihood*, Chapman & Hall/CRC.

[18] Qin, J. (2017) *Biased Sampling, Over-identified Parameter Problems and Beyond*, Springer.

[19] Qin, J. and B. Zhang (2007) Empirical-likelihood-based inference in missing response problems and its application in observational studies, *Journal of the Royal Statistical Society*, B 69, 101-122.

[20] Qin, J., Zhang, B. and D. H. Y. Leung (2009) Empirical likelihood in missing data problems, *Journal of the American Statistical Association*, 104, 1492-1503.

[21] Robins, J. M., Rotnitzky, A. and L. P. Zhao (1994) Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, 89, 846-886.

[22] Rosenbaum, P. R. (1987) Model-based direct adjustment, *Journal of the American Statistical Association*, 82, 387-394.

[23] Rosenbaum, P. R. (2002) *Observational studies*, Springer, New York.

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

[25] Wang, Y. and J. R. Zubizaretta (2017) Approximate balancing weights: characterizations from a shrinkage estimation perspective, Working paper.

[26] Zubizaretta, J. R. (2015) Stable weights that balance covariates for estimation with incomplete outcome data, *Journal of the American Statistical Association*, 110, 910-922.

Department of Economics, University of Pennsylvania, 133 South 36th Street, Philadelphia, PA 19104, USA.

E-mail address: akarun@sas.upenn.edu

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

E-mail address: t.otsu@lse.ac.uk