Approximate residual balancing: debiased inference of average treatment effects in high dimensions

Susan Athey, Guido W. Imbens and Stefan Wager

Stanford University, USA

[Received November 2016. Final revision December 2017]

Summary. There are many settings where researchers are interested in estimating average treatment effects and are willing to rely on the unconfoundedness assumption, which requires that the treatment assignment be as good as random conditional on pretreatment variables. The unconfoundedness assumption is often more plausible if a large number of pretreatment variables are included in the analysis, but this can worsen the performance of standard approaches to treatment effect estimation. We develop a method for debiasing penalized regression adjustments to allow sparse regression methods like the lasso to be used for $\sqrt{n}$-consistent inference of average treatment effects in high dimensional linear models. Given linearity, we do not need to assume that the treatment propensities are estimable, or that the average treatment effect is a sparse contrast of the outcome model parameters. Rather, in addition to standard assumptions used to make lasso regression on the outcome model consistent under 1-norm error, we require only overlap, i.e. that the propensity score be uniformly bounded away from 0 and 1. Procedurally, our method combines balancing weights with a regularized regression adjustment.

Keywords: Causal inference; Potential outcomes; Propensity score; Sparse estimation

1. Introduction

To identify causal effects in observational studies, practitioners may assume that treatment assignments are as good as random (or unconfounded) conditional on observed features of the units; see Rosenbaum and Rubin (1983) and Imbens and Rubin (2015) for general discussions. Motivated by this set-up, there is a large literature on how to adjust for differences in observed features between the treatment and control groups; some popular methods include regression, matching, propensity score weighting and subclassification, as well as doubly robust combinations thereof (e.g. Abadie and Imbens (2006), Heckman et al. (1998), Hirano et al. (2003), Robins et al. (1994, 1995, 2017), Rosenbaum (2002), Tan (2010), Tsiatis (2007) and Van Der Laan and Rubin (2006)).

In practice, researchers sometimes need to account for a substantial number of features to make this assumption of unconfoundedness plausible. For example, in an observational study of the effect of flu vaccines on hospitalization, we may be concerned that only controlling for differences in the age and sex distribution between controls and treated individuals may not be sufficient to eliminate biases. In contrast, controlling for detailed medical histories and personal characteristics may make unconfoundedness more plausible. But the formal asymptotic theory in the earlier literature considers only the case where the sample size increases whereas the number of features remains fixed, and so approximations based on those results may not yield...
valid inferences in settings where the number of features is large, possibly even larger than the sample size.

There has been considerable recent interest in adapting methods from the earlier literature to high dimensional settings. Belloni et al. (2014, 2017) showed that attempting to control for high dimensional confounders by using a regularized regression adjustment obtained via, for example, the lasso can result in substantial biases. Belloni et al. (2014) proposed an augmented variable selection scheme to avoid this effect, whereas Belloni et al. (2017), Chernozhukov et al. (2017), Farrell (2015) and Van der Laan and Rose (2011) built on the work of Robins et al. (1994, 1995) and discussed how a doubly robust approach to average treatment effect estimation in high dimensions can also be used to compensate for the bias of regularized regression adjustments. Despite the breadth of research on the topic, all these references rely crucially on the existence of a consistent estimator of the propensity score, i.e. the conditional probability of receiving treatment given the features, to yield $\sqrt{n}$-consistent estimates of the average treatment effect in high dimensions. (Some of these methods assume that the propensity scores can be consistently estimated by using a sparse logistic model, whereas others allow for the use of more flexible modelling strategies following, for example, McCaffrey et al. (2004), Van der Laan et al. (2007) or Westreich et al. (2010).)

In this paper, we show that, in settings where we are willing to entertain a sparse, well-specified linear model on the outcomes, efficient inference of average treatment effects in high dimensions is possible under more general assumptions than suggested by the literature discussed above. Given linearity assumptions, we show that it is not necessary to estimate treatment propensities consistently; rather, it is enough to rely on debiasing techniques building on recent developments in the high dimensional inference literature (Javanmard and Montanari, 2014, 2015; Van de Geer et al., 2014; Zhang and Zhang, 2014). In particular, in sparse linear models, we show that $\sqrt{n}$-consistent inference of average treatment effects is possible provided that we simply require overlap, i.e. that the propensity score be uniformly bounded away from 0 and 1 for all values in the support of the pretreatment variables. We do not need to assume the existence of a consistent estimator of the propensity scores, or any form of sparsity on the propensity model.

The starting point behind both our method and the doubly robust methods of Belloni et al. (2017), Chernozhukov et al. (2017), Farrell (2015), Van der Laan and Rose (2011), etc. is a recognition that high dimensional regression adjustments (such as the lasso) always shrink estimated effects, and that ignoring this shrinkage may result in prohibitively biased treatment effect estimates. The references on doubly robust estimation then proceed to show that propensity-based adjustments can be used to compensate for this bias, provided that we have a consistent propensity model that converges sufficiently fast to the truth. Conceptually, this work builds on the result of Rosenbaum and Rubin (1983), who showed that controlling for the propensity score is sufficient to remove all biases associated with observed covariates, regardless of their functional form.

If we are willing to focus on high dimensional linear models, however, it is possible to tighten the connection between the estimation strategy and the objective of estimating the average treatment effect and, in doing so, to extend the number of settings where $\sqrt{n}$-consistent inference is possible. The key insight is that, in a linear model, propensity-based methods are attempting to solve a needlessly difficult task when they seek to eliminate biases of any functional form. Rather, in linear models, it is enough to correct for linear biases. In high dimensions, this can still be challenging; however, we find that it is possible to correct approximately for such biases whenever we assume overlap.

Concretely, we study the following two-stage approximate residual balancing algorithm. First,
we fit a regularized linear model for the outcome given the features separately in the two treatment groups. In the current paper we focus on the elastic net (Zou and Hastie, 2005) and the lasso (Chen et al., 1998; Tibshirani, 1996) for this component and present formal results for the latter. In a second stage, we reweight the first-stage residuals by using weights that approximately balance all the features between the treatment and control groups. Here we follow Zubizarreta (2015) and optimize the implied balance and variance provided by the weights, rather than the fit of the propensity score. Approximate balancing on all pretreatment variables (rather than exact balance on a subset of features, as in a regularized regression, or weighting using a regularized propensity model that may not be able to capture all relevant dimensions) enables us to guarantee that the bias arising from a potential failure to adjust for a large number of weak confounders can be bounded. Formally, this second step of reweighting residuals by using the weights that were proposed by Zubizarreta (2015) is closely related to debiasing corrections studied in the high dimensional regression literature (Javanmard and Montanari, 2014, 2015; Van de Geer et al., 2014; Zhang and Zhang, 2014); we comment further on this connection in Section 3.

This approach also bears a close conceptual connection to work by Chan et al. (2016), Deville and Särndal (1992), Graham et al. (2012, 2016), Hainmueller (2012), Hellerstein and Imbens (1999), Imai and Ratkovic (2014) and Zhao (2016), who fitted propensity models to the data under a constraint that the resulting inverse propensity weights exactly balance the covariate distributions between the treatment and control groups and found that these methods outperform propensity-based methods that do not impose balance. Such an approach, however, is only possible in low dimensions; in high dimensions where there are more covariates than samples, achieving exact balance is in general impossible. One of the key findings of this paper is that, in high dimensions, it is still often possible to achieve approximate balance under reasonable assumptions and that—when combined with a lasso regression adjustment—approximate balance suffices for eliminating bias due to regularization.

In our simulations, we find that three features of the algorithm are important:

(a) the direct covariance adjustment based on the outcome data with regularization to deal with the large number of features,
(b) the weighting using the relationship between the treatment and the features and
(c) the fact that the weights are based on direct measures of imbalance rather than on estimates of the propensity score.

The finding that both weighting and regression adjustment are important is similar to conclusions drawn from the earlier literature on doubly robust estimation (e.g. Robins and Ritov (1997)), where combining both techniques was shown to extend the set of problems where efficient treatment effect estimation is possible. The finding that weights designed to achieve balance perform better than weights based on the propensity score is consistent with findings in Chan et al. (2016), Graham et al. (2012, 2016), Hainmueller (2012), Imai and Ratkovic (2014) and Zubizarreta (2015).

Our paper is structured as follows. First, in Section 2, we motivate our two-stage procedure by using a simple bound for its estimation error. Then, in Section 3, we provide a formal analysis of our procedure under high dimensional asymptotics, and we identify conditions under which approximate residual balancing is asymptotically Gaussian and enables practical inference about the average treatment effect with dimension-free rates of convergence. Finally, in Section 5, we conduct a simulation experiment and find that our method performs well in a wide variety of settings relative to other proposals in the literature. A software implementation for R is available at https://github.com/swager/balanceHD.
2. Estimating average treatment effects in high dimensions

2.1. Setting and notation

Our goal is to estimate average treatment effects in the potential outcomes framework, or Rubin causal model (Rubin, 1974; Imbens and Rubin, 2015). For each unit in a large population there is a pair of (scalar) potential outcomes \((Y_i(0), Y_i(1))\). Each unit is assigned to the treatment or not, with the treatment indicator denoted by \(W_i \in \{0, 1\}\). Each unit is also characterized by a vector of covariates or features \(X_i \in \mathbb{R}^p\), with \(p\) potentially large, possibly larger than the sample size. For a random sample of size \(n\) from this population, we observe the triple \((X_i, W_i, Y_{i}^{\text{obs}})\) for \(i = 1, \ldots, n\), where

\[
Y_{i}^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(1) & \text{if } W_i = 1, \\ Y_i(0) & \text{if } W_i = 0 \end{cases}
\]

is the realized outcome, equal to the potential outcome corresponding to the actual treatment received. The total number of treated units is equal to \(n_t\) and the number of control units equals \(n_c\). We frequently use the shorthand \(X_c\) and \(X_t\) for the feature matrices corresponding only to control or treated units respectively. We write the propensity score, i.e. the conditional probability of receiving the treatment given features, as \(e(x) = \mathbb{P}(W_i = 1|X_i = x)\) (Rosenbaum and Rubin, 1983). We focus primarily on the conditional average treatment effect for the treated sample:

\[
\tau = \frac{1}{n_t} \sum_{\{i: W_i = 1\}} \mathbb{E}[Y_i(0) - Y_i(1)|X_i].
\]

We note that the average treatment effect for the controls and the overall average effect can be handled similarly. Throughout the paper we assume unconfoundedness, i.e. that, conditional on the pretreatment variables, treatment assignment is as good as random (Rosenbaum and Rubin, 1983); we also assume a linear model for the potential outcomes in both groups.

**Assumption 1** (unconfoundedness). \(W_i \perp \perp (Y_i(0), Y_i(1)) | X_i\).

**Assumption 2** (linearity). The conditional response functions satisfy \(\mu_c(x) = \mathbb{E}[Y_i(0)|X = x] = x\beta_c\) and \(\mu_t(x) = \mathbb{E}[Y_i(1)|X = x] = x\beta_t\), for all \(x \in \mathbb{R}^p\).

Here, we shall use the linear model only for the control outcome because we focus on the average effect for the treated units, but if we were interested in the overall average effect we would need linearity in both groups. The linearity assumption is strong, but in high dimensions some strong structural assumption is in general needed for inference to be possible. Then, given linearity, we have

\[
\tau = \mu_t - \mu_c , \quad \mu_t = \bar{X}_t\beta_t , \quad \mu_c = \bar{X}_c\beta_c , \quad \bar{X}_t = \frac{1}{n_t} \sum_{i=1}^{n} \mathbb{1}\{W_i = 1\} X_i.
\]

Estimating the first term is easy: \(\hat{\mu}_t = \bar{Y}_t = \sum_{\{i: W_i = 1\}} Y_i^{\text{obs}} / n_t\) is unbiased for \(\mu_t\). In contrast, estimating \(\mu_c\) is a major challenge, especially in settings where \(p\) is large, and it is the main focus of the paper.

2.2. Baselines and background

We begin by reviewing two classical approaches to estimating \(\mu_c\), and thus also \(\tau\), in the above linear model. The first is a weighting-based approach, which seeks to reweight the control sample to make it look more like the treatment sample; the second is a regression-based approach, which
seeks to adjust for differences in features between treated and control units by fitting an accurate model to the outcomes. Though neither approach alone performs well in a high dimensional setting with a generic propensity score, we find that these two approaches can be fruitfully combined to obtain better estimators for $\tau$.

2.2.1. Weighted estimation

A first approach is to reweight the control data set by using weights $\gamma_i$ to make the weighted covariate distribution mimic the covariate distribution in the treatment population. Given the weights we estimate $\hat{\mu}_c$ as a weighted average $\hat{\mu}_c = \frac{\sum_{i:W_i=0} \gamma_i Y_i^{obs}}{\sum_{i:W_i=0} \gamma_i}$. The standard way of selecting weights $\gamma_i$ uses the propensity score: $\gamma_i = e(X_i) / \{1 - e(X_i)\} \{\sum_{i:W_i=0} e(X_j) / \{1 - e(X_j)\}\}$. To implement these methods researchers typically substitute an estimate of the propensity score into this expression. Such inverse propensity weights with non-parametric propensity score estimates have desirable asymptotic properties in settings with a small number of covariates (Hirano et al., 2003). The finite sample performance of methods based on inverse propensity weighting can be poor, however, both in settings with limited overlap in covariate distributions and in settings with many covariates. A key difficulty is that estimating the treatment effect then involves dividing by $1 - \hat{e}(X_i)$, and so small inaccuracies in $\hat{e}(X_i)$ can have large effects, especially when $e(x)$ can be close to 1; this problem is often quite severe in high dimensions.

As discussed in Section 1, if the control potential outcomes $Y_i(0)$ have a linear dependence on $X_i$, then using weights $\gamma_i$ that explicitly seek to balance the features $X_i$ is often advantageous (Deville and Särndal, 1992; Chan et al., 2016; Graham et al., 2012, 2016; Hainmueller, 2012; Hellerstein and Imbens, 1999; Imai and Ratkovic, 2014; Zhao, 2016; Zubizarreta, 2015). This is a subtle but important improvement. The motivation behind this approach is that, in a linear model, the bias for estimators based on weighted averaging depends solely on $\bar{X}_t - \sum_{i:W_i=0} \gamma_i X_i$. Therefore getting the propensity model exactly right is less important than accurately matching the moments of $\bar{X}_t$. In high dimensions, however, exact balancing weights do not in general exist. When $p \gg n_c$, there will in general be no weights $\gamma_i$ for which $\bar{X}_t - \sum_{i:W_i=0} \gamma_i X_i = 0$, and even in settings where $p < n_c$ but $p$ is large such estimators would not have good properties. Zubizarreta (2015) extended the balancing weights approach to allow for weights that achieve approximate balance instead of exact balance; however, directly using his approach does not allow for $\sqrt{n}$-consistent estimation in a regime where $p$ is much larger than $n$.

2.2.2. Regression adjustments

A second approach is to compute an estimator $\hat{\beta}_c$ for $\beta_c$ by using the $n_c$ control observations, and then to estimate $\mu_c$ as $\hat{\mu}_c = \bar{X}_t \hat{\beta}_c$. In a low dimensional regime with $p \ll n_c$, the ordinary least squares estimator for $\hat{\beta}_c$ is a natural choice and yields an accurate and unbiased estimate of $\mu_c$. In high dimensions, however, the problem is more delicate: accurate unbiased estimation of the regression adjustment is in general impossible, and methods such as the lasso, ridge regression or the elastic net may perform poorly when plugged in for $\beta_c$, in particular when $\bar{X}_t$ is far from $\bar{X}_c$. As stressed by Belloni et al. (2014), the problem with plain lasso regression adjustments is that features with a substantial difference in average values between the two treatment arms can generate large biases even if the coefficients on these features in the outcome regression are small. Thus, a regularized regression that has been tuned to optimize goodness of fit on the outcome model is not appropriate whenever bias in the treatment effect estimate due to failing to control for potential confounders is of concern. To address this problem, Belloni et al. (2014) proposed running least squares regression on the union of two sets of selected variables: one selected by a lasso regressing the outcome on the covariates, and the other
selected by a lasso logistic regression for the treatment assignment. We note that estimating \( \mu_c \) by a regression adjustment \( \hat{\mu}_c = \bar{X}_t \hat{\beta}_c \), with \( \hat{\beta}_c \) estimated by ordinary least squares on selected variables, is implicitly equivalent to using a weighted averaging estimator with weights \( \gamma \) chosen to balance the selected features (Robins et al., 2007). The approach of Belloni et al. (2014) works well in settings where both the outcome regression and the treatment regression are at least approximately sparse. However, when the propensity is not sparse, we find that the performance of such double-selection methods is often poor.

### 2.3. Approximate residual balancing

Here we propose a new method combining weighting and regression adjustments to overcome the limitations of each method. In the first step of our method, we use a regularized linear model, e.g. the lasso or the elastic net, to obtain a pilot estimate of the treatment effect. In the second step, we do ‘approximate balancing’ of the regression residuals to estimate treatment effects: we weight the residuals by using weights that achieve approximate balance of the covariate distribution between treatment and control groups. This step compensates for the potential bias of the pilot estimator that arises due to confounders that may be weakly correlated with the outcome but are important because of their correlation with the treatment assignment. We find that the regression adjustment is effective at capturing strong effects; the weighting in contrast is effective at capturing small effects. The combination leads to an effective and simple-to-implement estimator for average treatment effects with many features.

We focus on a meta-algorithm that first computes an estimate \( \hat{\beta}_c \) of \( \beta_c \) by using the full sample of control units. This estimator may take a variety of forms, but typically it will involve some form of regularization to deal with the number of features. Second we compute weights \( \gamma_i \) that balance the covariates at least approximately and apply these weights to the residuals (Cassel et al., 1976; Robins et al., 1994):

\[
\hat{\mu}_c = \bar{X}_t \hat{\beta}_c + \sum_{\{i:W_i=0\}} \gamma_i(Y_i^{obs} - X_i \hat{\beta}_c). \tag{4}
\]

In other words, we fit a model parameterized by \( \beta_c \) to capture some of the strong signals and then use direct numerical rebalancing of the control data on the features to extract left-over signal from the residuals \( Y_i^{obs} - X_i \hat{\beta}_c \). Ideally, we would hope for the first term to take care of any strong effects, whereas the rebalancing of the residuals can efficiently take care of the small spread-out effects. Our theory and experiments will verify that this is so.

A major advantage of the functional form in equation (4) is that it yields a simple and powerful theoretical guarantee, as stated below. Recall that \( X_c \) is the feature matrix for the control units. Consider the difference between \( \hat{\mu}_c \) and \( \mu_c \) for our proposed approach: \( \hat{\mu}_c - \mu_c = (\tilde{X}_t - X_c^T \gamma)(\hat{\beta}_c - \beta_c) + \gamma \varepsilon \), where \( \varepsilon \) is the intrinsic noise \( \varepsilon_i = Y_i(0) - X_i \beta_c \). With only the regression adjustment and no weighting, the difference would be \( \hat{\mu}_{c,reg} - \mu_c = (\tilde{X}_t - \tilde{X}_c)(\hat{\beta}_c - \beta_c) + 1\varepsilon/n_c \), and with only the weighting the difference would be \( \hat{\mu}_{c,weight} - \mu_c = (\tilde{X}_t - X_c^T \gamma)\hat{\beta}_c + \gamma \varepsilon \). Without any adjustment, just using the average outcome for the controls as an estimator for \( \mu_c \), the difference between the estimator for \( \mu_c \) and its actual value would be \( \hat{\mu}_{c,no-adj} - \mu_c = (\tilde{X}_t - \tilde{X}_c)\beta_c + 1\varepsilon/n_c \). The regression reduces the bias from \( (\tilde{X}_t - \tilde{X}_c)\beta_c \) to \( (\tilde{X}_t - \tilde{X}_c)(\beta_c - \beta_c) \), which will be a substantial reduction if the estimation error \( \hat{\beta}_c - \beta_c \) is small relative to \( \beta_c \). The weighting further reduces this to \( (\tilde{X}_t - X_c^T \gamma)(\hat{\beta}_c - \beta_c) \), which may be helpful if there is a substantial difference between \( \tilde{X}_t \) and \( \tilde{X}_c \). This result, which is formalized below, shows the complementary nature of the regression adjustment and the weighting. All proofs are given in the on-line appendix.
Proposition 1. The estimator \( \hat{\mu}_c - \mu_c \leq \| \hat{X}_1 - X_1^T \gamma \|_{\infty} \| \hat{\beta}_c - \beta_c \|_1 + | \sum_i (i: W_i = 0) \gamma_i \varepsilon_i |. \)

This result decomposes the error of \( \hat{\mu}_c \) into two parts. The first is a bias term reflecting the dimension \( p \) of the covariates; the second term is a variance term that does not depend on it. The upshot is that the bias term, which encodes the high dimensional nature of the problem, involves a product of two factors that can both be made reasonably small; we shall focus on regimes where the first term should be expected to scale as \( O(\sqrt{\log(p)/n}) \), whereas the second term scales as \( O(k \sqrt{\log(p)/n}) \) where \( k \) is the sparsity of the outcome model. If we are in a sufficiently sparse regime (i.e. \( k \) is sufficiently small), proposition 1 implies that our procedure will be variance dominated; and, under these conditions, we also show that it is \( \sqrt{n} \) consistent.

To exploit proposition 1, we need to make concrete choices for the weights \( \gamma \) and the parameter estimates \( \hat{\beta}_c \). First, just like Zubizarreta (2015), we choose our weights \( \gamma \) to optimize the bias and variance terms in proposition 1 directly; the functional form of \( \gamma \) is given in expression (5) below, where \( \zeta \in (0, 1) \) is a tuning parameter. We refer to them as *approximately balancing weights* since they seek to make the mean of the reweighted control sample, namely \( X_1^T \gamma \), match the treated sample mean \( \bar{X}_t \) as closely as possible. The positivity constraint on \( \gamma_i \) in expression (5) aims to prevent the method from extrapolating too aggressively, whereas the upper bound is added for technical reasons discussed in Section 3. Meanwhile, for estimating \( \hat{\beta}_c \), we simply need to use an estimator that achieves sufficiently good risk bounds under \( L_1 \)-risk. In our analysis, we focus on the lasso (Chen et al., 1998; Tibshirani, 1996); however, in experiments, we use the elastic net for additional stability (Zou and Hastie, 2005). Our complete algorithm is described in the following procedure 1 (approximately residual balancing with the elastic net).

The algorithm estimates the average treatment effect on the treated by approximately balanced residual weighting. Here, \( \zeta \in (0, 1) \), \( \alpha \in (0, 1] \) and \( \lambda > 0 \) are tuning parameters. This procedure is implemented in our R package balanceHD; we default to \( \zeta = 0.5 \) and \( \alpha = 0.9 \), and select \( \lambda \) by cross-validation using the lambda.1se rule from the glmnet package (Friedman et al., 2010). The optimization problem (5) is a quadratic programme, and so can be solved by using off-the-shelf convex optimization software; we use the interior point solver mosek by default (MOSEK, 2015).

**Step 1**: compute positive approximately balancing weights \( \gamma \) as

\[
\gamma = \arg \min_{\tilde{\gamma}} \left\{ (1 - \zeta) \| \tilde{\gamma} \|_2^2 + \zeta \| \bar{X}_t - X_1^T \gamma \|_{\infty}^2 \right\} \text{ subject to } \sum_{i: W_i = 0} \tilde{\gamma}_i = 1 \text{ and } 0 \leq \tilde{\gamma}_i \leq n^{-2/3}.
\]

**Step 2**: fit \( \hat{\beta}_c \) in the linear model by using a lasso or an elastic net:

\[
\hat{\beta}_c = \arg \min_{\beta} \left[ \sum_{i: W_i = 0} (Y_i^{\text{obs}} - X_i \beta)^2 + \lambda \left\{ (1 - \alpha) \| \beta \|_2^2 + \alpha \| \beta \|_1 \right\} \right].
\]

**Step 3**: estimate the average treatment effect \( \tau \) as

\[
\hat{\tau} = \bar{Y}_t - \left\{ \bar{X}_t \hat{\beta}_c + \sum_{i: W_i = 0} \gamma_i (Y_i^{\text{obs}} - X_i \hat{\beta}_c) \right\}.
\]

Finally, although we do use this estimator in the present paper, we note that an analogous estimator for the average treatment effect \( \mathbb{E}[Y(1) - Y(0)] \) can also be constructed:
in this paper and is available in our software package balanceHD. The conditions that are required for $\sqrt{n}$-consistent inference of $\tau_{ATE}$ using expression (8) directly mirror the conditions (sparsity, overlap etc.) listed in Section 3 for inference about the average treatment effect on the treated.

2.4. Connection to doubly robust estimation

The idea of combining weighted and regression-based approaches to treatment effect estimation has a long history in the causal inference literature. Given estimated propensity scores $\hat{e}(X_i)$, Cassel et al. (1976) and Robins et al. (1994) proposed the use of an augmented inverse propensity weighting (AIPW) estimator,

$$\hat{\mu}_c^{AIPW} = \bar{X}_i \hat{\beta}_c + \sum_{\{i: \hat{W}_i = 0\}} \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)} \frac{(Y_{i}^{obs} - X_i \hat{\beta}_c)}{\sum_{\{i: \hat{W}_i = 0\}} \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)}},$$

the difference between this estimator and ours is that Cassel et al. (1976) and Robins et al. (1994) obtained their weights $\gamma_i$ for equation (4) via propensity score modelling instead of quadratic programming. Estimators of this type have several desirable aspects: they are ‘doubly robust’ in the sense that they are consistent whenever either the propensity fit $\hat{e}(\cdot)$ or the outcome fit $\hat{\beta}_c$ is consistent, and they are asymptotically efficient in a semiparametric specification (Hahn, 1998; Hirano et al., 2003; Robins and Rotnitzky, 1995; Tsiatis, 2007). However, a practical concern with this class of methods is that they may perform less well when $1 - \hat{e}(X_i)$ is close to 0 (Hirano et al., 2003; Kang and Schafer, 2007). Several higher order refinements to the simple AIPW estimator (9) have also been considered in the literature. In particular, Kang and Schafer (2007) used the inverse propensity weights $\hat{e}(X_i) / \{1 - \hat{e}(X_i)\}$ as sample weights when estimating $\hat{\beta}_c$, whereas Scharfstein et al. (1999) and Van Der Laan and Rubin (2006) considered adding these weights as features in the outcome model; see also Robins et al. (2007) and Tan (2010) for further discussion.

Belloni et al. (2017) and Farrell (2015) studied the behaviour of AIPW in high dimensions and established conditions under which they can reach efficiency when both the propensity function and the outcome model are consistently estimable. Intriguingly, in low dimensions, doubly robust methods are not necessary for achieving semiparametric efficiency. This rate can be achieved by either non-parametric inverse propensity weighting or non-parametric regression adjustments on their own (Chen et al., 2008; Hirano et al., 2003); doubly robust methods can then be used to relax the regularity conditions that are needed for efficiency (Farrell, 2015; Robins et al., 2017). Conversely, in high dimensions, both weighting and regression adjustments are required for $\sqrt{n}$-consistency (Belloni et al., 2017; Farrell, 2015; Robins and Ritov, 1997).

Although our estimator (4) is cosmetically quite closely related to the AIPW estimator (9), the motivation behind it is quite different. A common description of the AIPW estimator is that it tries to estimate two different nuisance components, i.e. the outcome model $\hat{\mu}_c$ and the propensity model $\hat{e}$; it then achieves consistency if either of these components is itself estimated consistently, and efficiency if both components are estimated at sufficiently fast rates. In contrast, our approximate residual balancing estimator bets on linearity twice: once in fitting the outcome model via the lasso, and once in debiasing the lasso via balancing weights (5).
By relying more heavily on linearity, we can considerably extend the set of problems under which $\sqrt{n}$-consistency is possible (assuming that linearity in fact holds). As a concrete example, a simple analysis of AIPW estimation in high dimensional linear models would start by assuming that the lasso is $o_p(n^{-1/4})$ consistent in root-mean-squared error, which can be attained via the lasso by assuming a $k$-sparse true model with sparsity level $k \ll \sqrt{n}/\log(p)$; and this is, in fact, exactly the same condition as we assume in theorem 2. (A more careful analysis of AIPW estimators can trade off the accuracy of the propensity and main effect models and, instead of requiring that both the propensity and the outcome models can be estimated at $o_p(n^{-1/2})$ rates, assumes only that the product of the two rates is bounded as $o_p(n^{-1/2})$; see, for example, Farrell (2015). In high dimensions, this amounts to assuming that the outcome and propensity models are both well specified and sparse, with respective sparsity levels $k_\beta$ and $k_e$ satisfying $k_\beta k_e \ll n/\log(p)^2$. AIPW can thus be preferable to approximate residual balancing given sufficiently sparse and well-specified propensity models, with $k_e \ll \sqrt{n}/\log(p)$.) Then, in addition to this requirement on the outcome model, AIPW estimators still need to posit the existence of an $o_p(n^{-1/4})$ consistent estimator of the treatment propensities, whereas we do not need to assume anything about the treatment assignment mechanism beyond overlap. The reason for this is that the task of balancing (which is all that is needed to correct for the bias of the lasso in a linear model) is different from the task of estimating the propensity score—and is in fact often substantially easier. (These distinctions are framed in a situation where the statistician starts with a set of high dimensional covariates and needs to find a way to control for all of them at once. In this setting, linearity is a strong assumption, and so it is not surprising that making this assumption lets us considerably weaken requirements on other aspects of the problem. In other applications, however, the statistician may have started with low dimensional data, but then created a high dimensional design by listing series expansions of the original data, interactions, etc. In this setting, linearity is replaced with smoothness assumptions on the outcome model (since any smooth function can be well approximated by using a sufficiently large number of terms from an appropriately chosen series expansion). Here, variants of our procedure can be more directly compared with doubly robust methods and, in particular, the $\gamma_i$ in fact converge to the oracle inverse propensity weights $e(X_i)/\{1-e(X_i)\}$; see Hirshberg and Wager (2017) and Wang and Zubizarreta (2017) for a discussion and further results.)

2.5. Related work

Our approximately balancing weights (5) are inspired by the recent literature on balancing weights (Chan et al., 2016; Deville and Särndal, 1992; Graham et al., 2012, 2016; Hainmueller, 2012; Hellerstein and Imbens, 1999; Hirano et al., 2001; Imai and Ratkovic, 2014; Zhao, 2016). Most closely related, Zubizarreta (2015) proposed to estimate $\tau$ by using the reweighting formula as in Section 2.2.1 with weights

$$\gamma = \arg\min_{\tilde{\gamma}} \{ \| \tilde{\gamma} \|_2 \text{ subject to } \sum \tilde{\gamma}_i = 1, \tilde{\gamma}_i \geq 0, \| \bar{X}_t - X^T_{Tc} \tilde{\gamma} \|_\infty \leq t \},$$

where the tuning parameter is $t$; he called these weights stable balancing weights. These weights are of course equivalent to ours, the only difference being that Zubizarreta bounds imbalance in constraint form whereas we do so in Lagrange form. The main conceptual difference between our setting and that of Zubizarreta (2015) is that he considered problem settings where $p < n_c$ and then considered $t$ to be a practically small tuning parameter, e.g. $t = 0.1\sigma$ or $t = 0.001\sigma$. However, in high dimensions, the optimization problem (10) will not in general be feasible for small values of $t$; and in fact the bias term $\| \bar{X}_t - X^T_{Tc} \tilde{\gamma} \|_\infty$ becomes the dominant source of error in estimating $\tau$. We call our weights $\gamma$ ‘approximately’ balancing to remind the reader of this
fact. In settings where it is only possible to achieve approximate balance, weighting alone as considered in Zubizarreta (2015) will not yield a $\sqrt{n}$-consistent estimate of the average treatment effect, and it is necessary to use a regularized regression adjustment as in estimator (4).

Similar estimators have been considered by Graham et al. (2012, 2016) and Hainmueller (2012) in a setting where exact balancing is possible, with slightly different objection functions. For example, Hainmueller (2012) used $\sum \gamma_i \log(\gamma_i)$ instead of $\sum \gamma_i^2$, leading to

$$\gamma = \arg\min_{\tilde{\gamma}} \left\{ \sum_{i: W_i = 0} \tilde{\gamma}_i \log(\tilde{\gamma}_i) \text{ subject to } \sum \tilde{\gamma}_i = 1, \tilde{\gamma}_i \geq 0, \|\tilde{X}_t - X_T\|_\infty = 0 \right\}. \quad (11)$$

This estimator has attractive conceptual connections to logistic regression and maximum entropy estimation, and in a low dimensional setting where $W|X$ admits a well-specified logistic model the methods of Graham et al. (2012, 2016) and Hainmueller (2012) are doubly robust (Zhao and Percival, 2017); see also Hirano et al. (2001), Imbens et al. (1998) and Newey and Smith (2004). In terms of our immediate concerns, however, the variance of $\hat{\tau}$ depends on $\gamma$ through $\|\gamma\|^2$ and not $\sum \gamma_i \log(\gamma_i)$, so our approximately balancing weights are more directly induced by our statistical objective than those defined in result (11).

Finally, in this paper, we have emphasized an asymptotic analysis point of view, where we evaluate estimators via their large sample accuracy. From this perspective, our estimator—which combines weighting with a regression adjustment as in expression (4)—appears largely to dominate pure weighting estimators; in particular, in high dimensions, we achieve $\sqrt{n}$-consistency whereas pure weighting estimators do not. In contrast, stressing practical concerns, Rubin (2008) strongly argued that ‘designed-based’ inference leads to more credible conclusions in applications by better approximating randomized experiments. In our context, design-based inference amounts to using a pure weighting estimator of the form $\sum \gamma_i Y_i$ where the $\gamma_i$ are chosen without looking at the $Y_i$. The methods that were considered by Chan et al. (2016), Graham et al. (2012), Hainmueller (2012), Zubizarreta (2015), etc. all fit within this design-based paradigm, whereas ours does not.

3. Asymptotics of approximate residual balancing

3.1. Approximate residual balancing as debiased linear estimation

As we have already emphasized, approximate residual balancing is a method that enables us to make inferences about average treatment effects without needing to estimate treatment propensities as nuisance parameters; rather, we build on recent developments on inference in high dimensional linear models (Cai and Guo, 2017; Javanmard and Montanari, 2014, 2015; Ning and Liu, 2017; Van de Geer et al., 2014; Zhang and Zhang, 2014). Our main goal is to understand the asymptotics of our estimates for $\mu_c = \tilde{X}_t \beta_c$. In the interest of generality, however, we begin by considering a broader problem, namely that of estimating generic contrasts $\xi \beta_c$ in high dimensional linear models. This detour via linear theory will help to highlight the statistical phenomena that make approximate residual balancing work, and to explain why—unlike the methods of Belloni et al. (2017), Chernozhukov et al. (2017) or Farrell (2015)—our method does not require consistent estimability of the treatment propensity function $e(x)$.

The problem of estimating sparse linear contrasts $\xi \beta_c$ in high dimensional regression problems has received considerable attention, including notable recent contributions by Javanmard and Montanari (2014, 2015), Van de Geer et al. (2014) and Zhang and Zhang (2014), which, however, exclusively considered the setting where $\xi$ is a sparse vector, and, in particular, focused on the case where $\xi$ is the $j$th basis vector $e_j$, i.e. the target estimand is the $j$th co-ordinate of $\beta_c$. But, in our setting, the contrast vector $\tilde{X}_t$ defining our estimand $\mu_c = \tilde{X}_t \beta_c$ is random and thus generically
dense; moreover, we are interested in applications where $m_t = \mathbb{E}[\tilde{X}_t]$ itself may also be dense.

Thus, a direct application of these methods is not appropriate in our problem. (As a concrete example, theorem 6 of Javanmard and Montanari (2014) shows that their debiased estimator $\hat{\beta}_c^{(\text{debiased})}$ satisfies $\sqrt{n}(\hat{\beta}_c^{(\text{debiased})} - \beta_c) = Z + \Delta$, where $Z$ is a Gaussian random variable with desirable properties and $\|\Delta\|_{\infty} = o(1)$. If we simply consider sparse contrasts of $\hat{\beta}_c$, then this error term $\Delta$ is negligible; however, in our setting, we would have a prohibitively large error term $\tilde{X}_t\Delta$ that may grow polynomially in $p$.) An extension of this line of work to the problem of estimating dense, generic contrasts $\theta = \xi\beta_c$ turns out to be closely related to our approximate residual balancing method for treatment effect estimation. To make this connection explicit, define the estimator

$$\hat{\theta} = \xi\hat{\beta}_c + \sum_{\{i: W_i = 0\}} \gamma_i(y_i^{\text{obs}} - X_i\hat{\beta}_c),$$

where

$$\gamma = \arg \min_{\gamma} \|\gamma\|_2^2 \text{ subject to } \|\xi - X^T\gamma\|_{\infty} \leq K\left\{\frac{\log(p)}{n_c}\right\}, \max_i |\gamma_i| \leq n_c^{-2/3},$$

$\hat{\beta}_c$ is a properly tuned sparse linear estimator and $K$ is a tuning parameter that is discussed below. If we set $\xi \leftarrow \tilde{X}_t$, then this estimator is just our treatment effect estimator from procedure 1. (Here, we phrased the imbalance constraint in constraint form rather than in Lagrange form; the reason for this is that, although there is a 1:1 mapping between these two settings, we found the former easier to work with formally whereas the latter appears to yield more consistent numerical performance. We also dropped the constraints $\Sigma\gamma_i = 1$ and $\gamma_i \geq 0$ for now but we shall revisit them in Section 3.3.) Conversely, in the classical parameter estimation setting with $\xi \leftarrow e_j$, the above procedure is algorithmically equivalent to that proposed by Javanmard and Montanari (2014, 2015). Thus, the estimator (12) can be thought of as an adaptation of the method of Javanmard and Montanari (2014, 2015) that debiases $\hat{\beta}_c$ specifically along the direction of interest $\xi$.

We begin our analysis in Section 3.2 by considering a general version of estimator (12) under fairly strong ‘transformed independence design’ generative assumptions on $X_c$. Although these assumptions may be too strong to be palatable in practical data analysis, this result lets us make a crisp conceptual link between approximate residual balancing and the debiased lasso. In particular, we find that (theorem 1) $\hat{\theta}$ from estimator (12) is $\sqrt{n}$ consistent for $\theta$ provided that $\xi^T\Sigma_c^{-1}\xi = O(1)$, where $\Sigma_c$ is the covariance of $X_c$. Interestingly, if $\Sigma_c = I_{p\times p}$, then in general $\xi^T\Sigma_c^{-1}\xi = \|\xi\|_2^2 = O(1)$ if and only if $\xi$ is very sparse, and so the classical debiased lasso theory that was reviewed above is essentially sharp despite only considering the sparse $\xi$ case (see also Cai and Guo (2017)). In contrast, whenever $\Sigma_c$ has latent correlation structure, it is possible to have $\xi^T\Sigma_c^{-1}\xi = O(1)$ even when $\xi$ is dense and $\|\xi\|_2 \gg 1$, provided that $\xi$ is aligned with the large latent components of $\Sigma_c$. We also note that, in the application to treatment effect estimation, $\tilde{X}_t^T\Sigma_c^{-1}\tilde{X}_t$ will in general be much larger than 1; however, in corollary 1 we show how to overcome this issue.

To our knowledge, this was the first result for $\sqrt{n}$-consistent inference about dense contrasts of $\beta_c$ at the time that we first circulated our manuscript. We note, however, simultaneous and independent work by Zhu and Bradic (2016), who developed a promising method for testing hypotheses of the form $\xi\beta_c = 0$ for potentially dense vectors $\xi$; their approach uses an orthogonal moments construction that relies on regressing $\xi X_t$ against a $(p - 1)$-dimensional design that captures the components of $X_t$ orthogonal to $\xi$.
Finally, in Section 3.3, we revisit the specific problem of high dimensional treatment effect estimation via approximate residual balancing under substantially weaker assumptions on the design matrix $X_c$: rather than assuming a generative ‘transformed independence design’ model, we simply require overlap and standard regularity conditions. The cost of relaxing our assumptions on $X_c$ is that we now obtain slightly looser performance guarantees; however, our asymptotic error rates are still in line with those that we could obtain from doubly robust methods. We also discuss practical, heteroscedasticity robust confidence intervals for $\hat{\beta}_c$ for any $c \in \mathbb{R}^n$, and the approximation error does not improve as $n_c$ and $p$ both grow large. Thus, our first task is to identify a class of problems for which programme (13) has a solution.

3.2. Debiasing dense contrasts

As we begin our analysis of $\hat{\beta}_c$ defined in expression (12), it is first important to note that the optimization programme (13) is not always feasible. For example suppose that $p = 2n_c$, that $X_c = (I_{n_c \times n_c} I_{n_c \times n_c})$ and that $\xi$ consists of $n$ times ‘1’ followed by $n$ times ‘−1’; then $\|\xi - X_c^T \gamma\|_\infty \geq 1$ for any $\gamma \in \mathbb{R}^n$, and the approximation error does not improve as $n_c$ and $p$ both grow large.

The following lemma establishes such a result for random designs, in the case of vectors $\xi$ for which $\xi^T \Sigma_c^{-1} \xi$ is bounded; here $\Sigma_c = \text{var}(X_i|W_i = 0)$ denotes the population variance of control features. We also rely on the following regularity condition, which will be needed for an application of the Hanson–Wright concentration bound for quadratic forms following Rudelson and Vershynin (2013).

**Assumption 3** (transformed independence design). Suppose that we have a sequence of random design problems with $X_c = Q \Sigma_c^{1/2}$, where $\mathbb{E}[Q_{ij}] = 0$, $\text{var}(Q_{ij}) = 1$, for all indices $i$ and $j$, and the individual entries $Q_{ij}$ are all independent. (To simplify our exposition, this assumption implicitly rules out the use of an intercept. Our analysis would go through verbatim, however, if we added an intercept $X_1 = 1$ to the design.) Moreover suppose that the $Q$-matrix is sub-Gaussian for some $\zeta > 0$, $\mathbb{E}[\exp\{\zeta^2 |Q_{ij} - \mathbb{E}[Q_{ij}]|\}] \leq \exp(\zeta^2 t^2 / 2)$ for any $t > 0$, and that $(\Sigma_c)_{ij} \leq S$ for all $j = 1, \ldots, p$.

**Lemma 1.** Suppose that we have a sequence of problems for which assumption 3 holds and, moreover, $\|\xi^T \Sigma_c^{-1} \xi \| \leq V$ for some constant $V > 0$. Then, there is a universal constant $C > 0$ such that, setting $K = C \zeta^2 \sqrt{(VS)}$, the optimization problem (13) is feasible with probability tending to 1; and, in particular, the constraints are satisfied by $\gamma_c^* = (1/n_c) \xi^T \Sigma_c^{-1} X_i$.

Lemma 1 is the key to our analysis of approximate residual balancing. Because, with high probability, the weights $\gamma^*$ from lemma 1 provide one feasible solution to the constraint in problem (13), we conclude that, again with high probability, the actual weights that we use for approximate residual balancing must satisfy $\|\gamma\|_2^2 \leq \|\gamma^*\|_2^2 \approx n_c^{-1} \xi^T \Sigma_c^{-1} \xi$. To turn this insight into a formal result, we need assumptions on both the sparsity of the signal and the covariance matrix $\Sigma_c$.

**Assumption 4** (sparsity). We have a sequence of problems indexed by $n$, $p$ and $k$ such that the parameter vector $\beta_c$ is $k$ sparse, i.e. $\|\beta_c\|_0 \leq k$, and that $k \log(p) / \sqrt{n} \rightarrow 0$. (In recent literature, there has been some interest in methods that require only approximate, rather than exact, $k$-sparsity. We emphasize that our results also hold with approximate rather than exact sparsity, as we use only our sparsity assumption to obtain bounds on $\|\hat{\beta}_c - \beta_c\|_1$ that can be used in conjunction with proposition 1. For simplicity of exposition, however, we restrict our present discussion to the case of exact sparsity.)
This sparsity requirement is quite strong. However, many analyses that seek to establish asymptotic normality in high dimensions rely on such an assumption. For example, Javanmard and Montanari (2014), Van de Geer et al. (2014) and Zhang and Zhang (2014) all made this assumption when seeking to provide confidence intervals for individual components of $\beta_c$; Belloni et al. (2014) used a similar assumption where they allowed for additional non-zero components, but they assumed that, beyond the largest $k$ components with $k$ satisfying the same sparsity condition, the remaining non-zero elements of $\beta_c$ are sufficiently small that they can be ignored, in what they referred to as approximate sparsity. (There are, of course, some exceptions to this assumption. In recent work, Javanmard and Montanari (2015) showed that inference of $\beta_c$ is possible even when $k \ll n \log(p)$ in a setting where $X$ is a random Gaussian matrix with either a known or extremely sparse population precision matrix; Wager et al. (2016) showed that lasso regression adjustments allow for efficient average treatment effect estimation in randomized trials even when $k \ll n \log(p)$, whereas the method of Zhu and Bradic (2016) for estimating dense contrasts $\xi \beta_c$ does not rely on sparsity of $\beta_c$ and instead places assumptions on the joint distribution of $\xi X_i$ and the individual regressors. The point in common between these results is that they let us weaken the sparsity requirements at the expense of strengthening our assumptions about the $X$-distribution.)

Next, our analysis builds on well-known bounds on the estimation error of the lasso (Bickel et al., 2009; Hastie et al., 2015) that require $X_c$ to satisfy a form of the restricted eigenvalue condition. Below, we make a restricted eigenvalue assumption on $\Sigma_c^{1/2}$; then, we shall use results from Rudelson and Zhou (2013) to verify that this also implies a restricted eigenvalue condition on $X_c$.

**Assumption 5** (well-conditioned covariance). Given the sparsity level $k$ that was specified above, the covariance matrix $\Sigma_c^{1/2}$ of the control features satisfies the $\{k, 2\omega, 10\}$-restricted eigenvalue defined as follows, for some $\omega > 0$. For $1 \leq k \leq p$ and $L \geq 1$, define the set $C_k(L)$ as

$$C_k(L) = \left\{ \beta \in \mathbb{R}^p : \|\beta\|_1 \leq L \sum_{j=1}^k |\beta_{i,j}| \text{ for some } 1 \leq i_1 < \ldots < i_j \leq p \right\}.$$  

Then, $\Sigma_c^{1/2}$ satisfies the $\{k, \omega, L\}$-restricted eigenvalue condition if $\beta^T \Sigma_c \beta \geq \omega \|\beta\|_2^2$ for all $\beta \in C_k(L)$.

**Theorem 1.** Under the conditions of lemma 1, suppose that the control outcomes $Y_i(0)$ are drawn from a sparse, linear model as in assumptions 1, 2, 3 and 4, that $\Sigma_c^{1/2}$ satisfies the restricted eigenvalue property (assumption 5), and that we have a minimum estimand size $\|\xi\|_\infty \geq \kappa > 0$. Suppose, moreover, that we have homoscedastic noise: $\text{var}\{\varepsilon_i(0)X_i\} = \sigma^2$ for all $i = 1, \ldots, n$, and also that the response noise $\varepsilon_i(0) := Y_i(0) - \mathbb{E}[Y_i(0)|X_i]$ is uniformly sub-Gaussian with parameter $\sigma^2 > 0$. Finally, suppose that we estimate $\theta$ by using estimator (12), with the optimization parameter $K$ selected as in lemma 1 and the lasso penalty parameter set to $\lambda_n = 5 \xi \sqrt{\log(p)/n_c}$. Then, $\hat{\theta}$ is asymptotically Gaussian,

$$\begin{align*}
(\hat{\theta} - \theta)/\|\gamma\|_2 &\Rightarrow \mathcal{N}(0, \sigma^2), \\
nc \|\gamma\|^2/\xi^T \Sigma_c^{-1} \xi &\leq 1 + o_p(1).
\end{align*}$$

The minimum estimand size assumption is needed to rule out pathological superefficient behaviour. As a concrete example, suppose that $X_i \sim \mathcal{N}(0, I_{pxp})$, and that $\xi_j = 1/\sqrt{p}$ for $j = 1, \ldots, p$ with $p \gg n_c$ Then, with high probability, the optimization problem (13) will yield $\gamma = 0$. This leaves us with a simple lasso estimator $\hat{\theta} = \xi \hat{\beta}_c$, whose risk scales as $\mathbb{E}[\|\hat{\theta} - \theta\|^2] = \mathcal{O}\{k^2 \log(p)/(pn_c)\} \ll 1/n_c$. The problem with this superefficient estimator is that it is not necessarily asymptotically Gaussian.
The statement of theorem 1 highlights a connection between our debiased estimator (12), and the ordinary least squares (OLS) estimator. Under classical large sample asymptotics with \( n \gg p \), it is well known that the OLS estimator \( \hat{\theta}^{\text{OLS}} = \xi^T (X_0^T X_0)^{-1} X_0^T Y \), satisfies
\[
\sqrt{n_c (\hat{\theta}^{\text{OLS}} - \theta)} / \sqrt{\rho} \Rightarrow \mathcal{N}(0, \sigma^2),
\]
\[
\sqrt{n_c (\hat{\theta}^{\text{OLS}} - \theta - \sum_{i: W_i = 0} \gamma_i \varepsilon_i(0))} \to_d 0,
\]
where \( \gamma_i^* \) is as defined in lemma 1. By comparing this characterization with our result in theorem 1, it becomes apparent that our debiased estimator \( \hat{\theta} \) has been able to recover the large sample qualitative behaviour of \( \hat{\theta}^{\text{OLS}} \), despite being in a high dimensional \( p \gg n \) regime. The connection between debiasing and OLS ought not appear too surprising. After all, under classical assumptions, \( \hat{\theta}^{\text{OLS}} \) is known to be the minimum variance unbiased linear estimator for \( \theta \), whereas the weights \( \gamma \) in equation (13) were explicitly chosen to minimize the variance of \( \hat{\theta} \) subject to the estimator being nearly unbiased.

A downside of the above result is that our main goal is to estimate \( \mu_c = \bar{X}_t \beta_c \), and this contrast defining vector \( \bar{X}_t \) fails to satisfy the bound on \( \bar{X}_t^T \Sigma^{-1} \bar{X}_t \) as assumed in theorem 1. In fact, because \( \bar{X}_t \) is random, this quantity will in general be of the order of \( p/n \). In the result below, we show how to circumvent this problem under the weaker assumption that \( m_t^T \Sigma_c^{-1} m_t \) is bounded; at a high level, the proof shows that the stochasticity \( \bar{X}_t \) does not invalidate our previous result. We note that, because \( \bar{Y}_t \) is uncorrelated with \( \hat{\mu}_c \) conditionally on \( \bar{X}_t \), the following result also immediately implies a central limit theorem for \( \hat{\mu}_c = \bar{Y}_t - \hat{\mu}_c \) where \( \bar{Y}_t \) is the average of the treated outcomes.

**Corollary 1.** Under the conditions of theorem 1, suppose that we want to estimate \( \mu_c = \bar{X}_t \beta_c \) by replacing \( \xi \) with \( \bar{X}_t \) in expression (12), and let \( m_t = \mathbb{E}[X|W = 1] \). Suppose, moreover, that we replace all the assumptions made about \( \xi \) in theorem 1 with the following assumptions: throughout our sequence of problems, the vector \( m_t \) satisfies \( m_t^T \Sigma_c^{-1} m_t \leq V \) and \( \|m_t\|_\infty \geq \kappa \). Suppose, finally, that \( (X_0 - m_1)_j | W_i = 1 \) is sub-Gaussian with parameter \( \nu^2 > 0 \), and that the overall odds of receiving treatment \( \mathbb{P}(W = 1) / \mathbb{P}(W = 0) \) tend to a limit \( \rho \) bounded away from 0 and \( \infty \). Then, setting the tuning parameter in problem (13) as \( K = C \xi^2 / (VS + \nu \sqrt{2.1 \rho}) \), we obtain
\[
(\hat{\mu}_c - \mu_c) / \|\gamma\|_2 \Rightarrow \mathcal{N}(0, \sigma^2),
\]
\[
n_c \|\gamma\|_2^2 / m_t^T \Sigma_c^{-1} m_t \leq 1 + o_p(1).
\]

The asymptotic variance bound \( m_t^T \Sigma_c^{-1} m_t \) is exactly the Mahalanobis distance between the mean treated and control subjects with respect to the covariance of the control sample. Thus, our result shows that we can achieve asymptotic inference about \( \tau \) with a \( 1/\sqrt{n} \) rate of convergence, irrespective of the dimension of the features, subject only to a requirement on the Mahalanobis distance between the treated and control classes, and comparable sparsity assumptions on the \( Y \)-model as used by the rest of the high dimensional inference literature, including Belloni et al. (2014, 2017), Chernozhukov et al. (2017) and Farrell (2015). However, unlike this literature, we make no assumptions on the propensity model beyond overlap and do not require it to be estimated consistently. In other words, by relying more heavily on linearity of the outcome function, we can considerably relax the assumptions that are required to obtain \( \sqrt{n} \)-consistent treatment effect estimation.

### 3.3. A robust analysis with overlap

Our discussion so far, leading up to corollary 1, gives a characterization of when and why we
should expect approximate residual balancing to work. However, from a practical perspective, the assumptions that are used in our derivation—in particular the transformed independence design assumption—were stronger than those which we may feel comfortable making in applications.

In this section, we propose an alternative analysis of approximate residual balancing based on overlap. Informally, overlap requires that each unit has a positive probability of receiving each of the treatment and control conditions, and thus that the treatment and control populations cannot be too dissimilar. Without overlap, estimation of average treatment effects relies fundamentally on extrapolation beyond the support of the features and thus makes estimation inherently sensitive to functional form assumptions; and, for this reason, overlap has become a common assumption in the literature on causal inference from observational studies (Crump et al., 2009; Imbens and Rubin, 2015). For estimation of the average effect for the treated we in fact need only the propensity score to be bounded from above by $1 - \eta$, but for estimation of the overall average effect we would require both the lower and the upper bound on the propensity score. If we are willing to assume overlap, we can relax the transposed independence design assumption into much more routine regularity conditions on the design, as in assumption 7.

**Assumption 6 (overlap).** There is a constant $0 < \eta$ such that $\eta \leq e(x) \leq 1 - \eta$ for all $x \in \mathbb{R}^p$.

**Assumption 7 (design).** Our design $X$ satisfies the following two conditions. First, the design is sub-Gaussian, i.e. there is a constant $\nu > 0$ such that the distribution of $X_j$ conditional on $W = w$ is sub-Gaussian with parameter $\nu^2$ after centering. Second, we assume that $X_c$ satisfies the $\{k, \omega, 4\}$-restricted eigenvalue condition as defined in assumption 5, with probability tending to 1.

Following lemma 1, our analysis again proceeds by guessing a feasible solution to our optimization problem, and then using it to bound the variance of our estimator. Here, however, we use inverse propensity weights as our guess:

$\gamma \propto e(X_i) / \{1 - e(X_i)\}$. Our proof hinges on showing that the actual weights that we obtain from the optimization problem are at least as good as these inverse propensity weights, and thus our method will be at most as variable as the method that uses AIPW (9) with these oracle propensity weights.

**Theorem 2.** Suppose that we have $n$ independent and identically distributed training examples satisfying assumptions 1, 2, 4, 6 and 7, and that the treatment odds $P(W = 1) / P(W = 0)$ converge to $\rho$ with $0 < \rho < \infty$. Suppose, moreover, that we have homoscedastic noise: $\text{var}\{e_i(w)\} = \sigma^2$ for all $i = 1, \ldots, n$, and also that the response noise $e_i(w) := Y_i(w) - E[Y_i(w) | X_i]$ is uniformly sub-Gaussian with parameter $\nu^2 > 0$. Finally, suppose that we use estimator (4) with weights (5), except that we replace the Lagrange form penalty on the imbalance with a hard constraint $\| \bar{X}_t - X^T \bar{\gamma} \|_\infty \leq K \sqrt{\{\log(p)/n_c\}}$, with $K = \nu \sqrt{2 \{\rho + (\eta^{-1} - 1)^2\}}$. Moreover, we fit the outcome model by using a lasso with penalty parameter set to $\lambda_n = 5\nu \sqrt{2 \{\log(p)/n_c\}}$. Then,

\[ \frac{\hat{\mu}_c - \mu_c}{\|\gamma\|_2} \Rightarrow \mathcal{N}(0, \sigma^2), \]

\[ \frac{\hat{\tau} - \tau}{\sqrt{(n_i^{-1} + \|\gamma\|_2^2)}} \Rightarrow \mathcal{N}(0, \sigma^2), \]

where $\tau$ is the expected treatment effect on the treated (2). Moreover,

\[ \limsup_{n \to \infty} n_c \|\gamma\|_2^2 \leq \rho^{-2} E \left[ \left\{ \frac{e(X_i)}{1 - e(X_i)} \right\}^2 \left| W_i = 0 \right. \right]. \]
The rate of convergence that is guaranteed by expression (19) is the same as what we would obtain if we actually knew the true propensities and could use them for weighting (Robins et al., 1994, 1995). Here, we achieve this rate although we have no guarantees that the true propensities \( e(X_i) \) are consistently estimable. Finally, we note that, when the assumptions to corollary 1 hold, the bound (17) is stronger than expression (19); however, there are designs where the bounds match (Wang and Zubizarreta, 2017).

Finally, in applications, it is often of interest to have confidence intervals for \( \mu_c \) and \( \tau \) rather than just point estimates; below, we propose such a construction. Much like the sandwich variance estimates for OLS regression, our proposed confidence intervals are heteroscedasticity robust even though the underlying point estimates were motivated by using an argument written in terms of a homoscedastic sampling distribution.

**Corollary 2.** Under the conditions of theorems 1 or 2, suppose instead that we have heteroscedastic noise \( \nu_{\text{min}}^2 \leq \text{var}\{\varepsilon_i(W_i)|X_i, W_i\} \leq \nu^2 \) for all \( i = 1, \ldots, n \). Then, the following results hold:

\[
(\hat{\mu}_c - \mu_c) / \sqrt{\hat{V}_c} \Rightarrow \mathcal{N}(0, 1),
\]

\[
\hat{V}_c = \sum_{\{i: W_i = 0\}} \gamma_i^2 (Y_i - X_i \hat{\beta}_c)^2.
\]

To provide inference about \( \tau \), we also need error bounds for \( \hat{\mu}_t \). Under sparsity assumptions that are comparable with those made for \( \beta_c \) in theorem 2, we can verify that

\[
(\hat{\mu}_t - \mu_t) / \sqrt{\hat{V}_t} \Rightarrow (0, 1),
\]

\[
\hat{V}_t = \frac{1}{n^2} \sum_{\{i: W_i = 1\}} (Y_i - X_i \hat{\beta}_t)^2,
\]

where \( \hat{\beta}_t \) is obtained by using the lasso with \( \lambda_n = 5 \nu \sqrt{\log(p)/n_c} \). Moreover, \( \hat{\mu}_c \) and \( \hat{\mu}_t \) are independent conditionally on \( X \) and \( W \), thus implying that \( (\hat{\tau} - \tau) / (\hat{V}_c + \hat{V}_t)^{1/2} \Rightarrow \mathcal{N}(0, 1) \). This last expression is what we use for building confidence intervals for \( \tau \).

4. **Application: the efficacy of welfare-to-work programmes**

Starting in 1986, California implemented the ‘Greater avenues to independence’ (GAIN) programme, with an aim to reduce dependence on welfare and to promote work among disadvantaged households. The GAIN programme provided its participants with a mix of educational resources such as English as a second language courses and vocational training, and job search assistance. This programme is described in detail by Hotz et al. (2006). To evaluate the effect of GAIN, the Manpower Development Research Corporation conducted a randomized study between 1988 and 1993, where a random subset of GAIN registrants were eligible to receive GAIN benefits immediately, whereas others were embargoed from the programme until 1993 (after which point they were allowed to participate in the programme). All experimental subjects were followed for a 3-year post-randomization period.

The randomization for the GAIN evaluation was conducted separately by county; following Hotz et al. (2006), we consider data from Alameda, Los Angeles, Riverside and San Diego counties. As discussed in detail in Hotz et al. (2006), the experimental conditions differed noticeably across counties, both in terms of the fraction of registrants who were eligible for GAIN, i.e. the treatment propensity, and in terms of the subjects participating in the experiment. For example, the GAIN programmes in Riverside and San Diego counties sought to register all welfare cases...
in GAIN, whereas the programmes in Alameda and Los Angeles counties focused on long-term welfare recipients.

The fact that the randomization of the GAIN evaluation was done at the county level rather than at the state level presents us with a natural opportunity to test our method, as follows. We seek to estimate the average treatment effect of GAIN on the treated; however, we hide the county information from our procedure and instead try to compensate for sampling bias by controlling for a large amount of covariates. We used spline expansions of age and prior income, indicators for race, family status, etc., for a total of $p = 93$ covariates. Meanwhile, we can check our performance against a gold standard estimate of the average treatment effect that is stratified by county and thus guaranteed to be unbiased. (More formally, in our experiments, we set the gold standard using the county-stratified oracle estimator on bootstrap samples of the full $n = 19170$ sample. We use bootstrap samples to correct for the correlation of estimators $\hat{\tau}$ obtained by using the full data set and subsamples of it. We also note that, given this set-up, the quantity that we are using as our gold standard is not an estimate of $\tau$, i.e. the conditional average treatment effect on the treated sample, and should rather be thought of as an estimate of $E[\tau]$, i.e. the average treatment effect on the treated population. Since we are in a setting with a fairly weak signal, this should not make a noticeable difference in practice.)

We compare the behaviour of various methods for estimating the average treatment effect on the treated by using randomly drawn subsamples of the original data. In addition to approximate residual balancing, we consider AIPW (9) and double selection following Belloni et al. (2014) as our baselines. We also show the behaviour of an ‘oracle’ procedure that gets to observe the hidden county information and then simply estimates treatment effects for each county separately, and the ‘naive’ difference-in-means estimator that ignores the features $X$. In very small samples, the oracle procedure is not always well defined because some samples may result in counties where either everyone or no one is treated.

Fig. 1 compares the performance of the various methods. We see that approximate residual balancing and double selection both do well in terms of mean-squared error. Moreover, confidence intervals built via approximate residual balancing achieve effectively nominal coverage; double selection also achieves reasonable coverage and improves with $n$. In contrast, AIPW does not perform well here. The problem appears to be that estimating treatment propensities is quite difficult, and a cross-validated logistic elastic net often learns an effectively constant propensity model.

5. Simulation experiments

5.1. Methods under comparison

In addition to approximate residual balancing as described in procedure 1, the methods that we use as baselines are as follows: naive difference-in-means estimation $\hat{\tau} = \bar{Y}_t - \bar{Y}_c$ that ignores the covariate information $X$; the elastic net (Zou and Hastie, 2005) or, equivalently, procedure 1 with trivial weights $\gamma_i = 1/n_c$; approximate balancing or, equivalently, procedure 1 with trivial parameter estimates $\hat{\beta}_c = 0$ (Zubizarreta, 2015); inverse propensity weighting, as discussed in Section 2.2.1, with propensity estimates $\tilde{e}(X_i)$ obtained by elastic net logistic regression, with the propensity scores trimmed at 0.05 and 0.95; AIPW, which pairs elastic net regression adjustments with the above inverse propensity weights (9); the weighted elastic net, motivated by Kang and Schafer (2007), that uses inverse propensity weights as sample weights for the elastic net regression; targeted maximum likelihood estimation (TMLE), which fine tunes the elastic net regression estimates along the direction specified by the inverse propensity weights (Van Der
Laan and Rubin, 2006); OLS after model selection where, in the spirit of Belloni et al. (2014), we run lasso linear regression for $Y|X, W = 0$ and lasso logistic regression for $W|X$, and then compute the OLS estimate for $\tau$ on the union of the support of the three lasso problems.

Unless otherwise specified, all outcome and propensity models were fitted by using a (linear or logistic) elastic net. Whenever there is a ‘$\lambda$’ regularization parameter to be selected, we use cross-validation with the lambda.1se rule from the glmnet package (Friedman et al., 2010). Belloni et al. (2014) recommended selecting $\lambda$ by using more sophisticated methods, such as
Fig. 2. Illustrating simulation designs: (a) low dimensional version of the many-clusters simulation setting \( X \), control; \( \circ \), treated \( X \)-observations); (b) schematic diagram of a misspecified simulation setting, along the first covariate \( X_i \), ‘treatment effect’ curve (not to scale along the \( y \)-axis); ---, density of control; ----, density of treated)
Table 1. Root-mean-squared error $\sqrt{\mathbb{E}[(\hat{\tau} - \tau)^2]}$ in the two-cluster setting†

| Method                       | Results for the following experiments: |                      |                      |                      |                      |                      |
|------------------------------|----------------------------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
|                              |                                        | Beta model dense     | Beta model harmonic  | Beta model moderately sparse | Beta model very sparse |
|                              |                                        | Propensity model dense |                      |                      |                      |                      |
|                              |                                        |                      |                      |                      |                      |                      |
| Naive                        | 6.625                                  | 7.119                | 3.557                | 3.924                | 1.257                | 1.256                | 0.711                | 0.722                |
| Elastic net                  | 4.328                                  | 1.058                | 2.190                | 0.665                | 0.716                | 0.350                | 0.237                | 0.204                |
| Approximate balance          | 3.960                                  | 1.179                | 2.130                | 0.686                | 0.789                | 0.362                | 0.464                | 0.316                |
| Approximate residual balance | 3.832                                  | 0.423                | 1.854                | 0.320                | 0.495                | 0.213                | 0.185                | 0.165                |
| Inverse propensity weight    | 5.341                                  | 3.094                | 2.866                | 1.707                | 1.026                | 0.596                | 0.586                | 0.398                |
| AIPW                         | 4.082                                  | 0.618                | 2.031                | 0.415                | 0.607                | 0.242                | 0.209                | 0.166                |
| Weighted elastic net         | 4.086                                  | 0.562                | 1.984                | 0.385                | 0.575                | 0.232                | 0.207                | 0.171                |
| TMLE elastic net             | 3.811                                  | 0.591                | 1.843                | 0.399                | 0.495                | 0.239                | 0.192                | 0.165                |
| Double selection + OLS       | 6.625                                  | 0.620                | 3.540                | 0.430                | 0.525                | 0.233                | 0.254                | 0.165                |

†We used $n = 500$ and $p = 2000$ and scaled the signal such that $\|\beta\|_2 = 2$. All numbers are averaged over 400 simulation replications.

The square-root lasso (Belloni et al., 2011). However, in our simulations, our implementation of Belloni et al. (2014) still attains excellent performance in the regimes that the method is designed to work in. Similarly, our confidence intervals for $\tau$ are built by using a cross-validated choice of $\lambda$ instead of the fixed choice that is assumed by corollary 3. Our implementation of approximate residual balancing, as well as all the discussed baselines, is available in the R package balanceHD.

5.2. Simulation designs
We consider five different simulation settings. Our first setting is a two-cluster lay-out, with data drawn as $Y_i = (C_i + Z_i)\beta + W_i + \varepsilon_i$. Here, $W_i \sim \text{Bernoulli}(0.5)$, $Z_i \sim \mathcal{N}(0, I_{p \times p})$, $\varepsilon_i \sim \mathcal{N}(0, 1)$ and $C_i \in \mathbb{R}^p$ is a cluster centre that is one of $C_i \in \{0, \delta\}$, such that $\mathbb{P}(C_i = 0|W_i = 0) = 0.8$ and $\mathbb{P}(C_i = 0|W_i = 1) = 0.2$. We consider two settings for the between-cluster vector $\delta$: a ‘dense’ setting where $\delta = (4/\sqrt{n})1$, and a ‘sparse’ setting where $\delta = (40/\sqrt{n})1(\{j = 1 \text{ mod } 10\})$. Our second many-cluster lay-out is closely related to the first, except now we have 20 cluster centres $C_i \in \{c_1, \ldots, c_{20}\}$, where all the cluster centres are independently generated as $c_k \sim \mathcal{N}(0, I_{p \times p})$. To generate the data, we first draw $C_i$ uniformly at random from one of the 20 cluster centres and then set $W_i = 1$ with probability $\eta$ for the first 10 clusters and $W_i = 1$ with probability $1 - \eta$ for the last 10 clusters; we tried both $\eta = 0.1$ and $\eta = 0.25$. We illustrate this simulation concept in Fig. 2(a). In both cases, we chose $\beta$ as one of dense, harmonic,

$$\beta \propto (1, 1/\sqrt{2}, \ldots, 1/\sqrt{p}),$$

moderately sparse,

$$\beta \propto (1/10, 1/11, \ldots, 1/(p + 9)),$$


Table 2. Root-mean-squared error $\sqrt{\mathbb{E}[(\hat{\tau} - \tau)^2]}$ in the many-cluster setting†

| Method                      | Beta model dense | Beta model harmonic | Beta model moderately sparse | Beta model very sparse |
|-----------------------------|------------------|---------------------|------------------------------|------------------------|
|                             | Overlap (\(\eta\) 0.1) | Overlap (\(\eta\) 0.25) | Overlap (\(\eta\) 0.1) | Overlap (\(\eta\) 0.25) | Overlap (\(\eta\) 0.1) | Overlap (\(\eta\) 0.25) |
| Naive                       | 4.921            | 3.283               | 5.100                        | 3.231                   | 4.776                     | 3.270                        | 5.078                     | 3.348                        |
| Elastic net                 | 2.527            | 1.353               | 1.618                        | 0.869                   | 0.727                     | 0.385                        | 0.168                     | 0.108                        |
| Approximate balance         | 2.575            | 1.324               | 2.505                        | 1.379                   | 2.567                     | 1.248                        | 2.434                     | 1.396                        |
| Approximate residual balance| 2.123            | 1.172               | 1.528                        | 0.866                   | 0.653                     | 0.383                        | 0.158                     | 0.105                        |
| Inverse propensity weight   | 2.626            | 1.983               | 2.625                        | 1.943                   | 2.586                     | 1.896                        | 2.568                     | 2.000                        |
| AIPW                        | 2.102            | 1.233               | 1.515                        | 0.852                   | 0.656                     | 0.376                        | 0.154                     | 0.103                        |
| Weighted elastic net        | 2.061            | 1.176               | 1.576                        | 0.862                   | 0.727                     | 0.388                        | 0.194                     | 0.108                        |
| TMLE elastic net            | 2.095            | 1.208               | 1.500                        | 0.847                   | 0.656                     | 0.375                        | 0.163                     | 0.103                        |
| Double selection + OLS      | 2.726            | 1.526               | 3.364                        | 1.840                   | 2.201                     | 1.092                        | 0.211                     | 0.114                        |

†We used \(n = 800\) and \(p = 4000\) and scaled the signal such that \(\|\beta\|_2 = 3\). All numbers are averaged over 400 simulation replications.

\[
\beta \propto (10, \ldots, 10, 1, \ldots, 1, 0, \ldots, 0),
\]

or very sparse,

\[
\beta \propto (1, \ldots, 1, 0, \ldots, 0).
\]

The signal strength was scaled such that \(\|\beta\|_2\) is a two-cluster layout and \(\|\beta\|_2 = 3\) in the many-cluster layout.

Our next two simulations are built by using more traditional structural models. We first consider a sparse two-stage setting closely inspired by an experiment of Belloni et al. (2014). Here \(X_i \sim \mathcal{N}(0, \Sigma)\) with \(\Sigma_{ij} = \rho|\!\!-j|\), and \(\theta_i = X_i \beta_Y + \epsilon_{i1}\). Then, \(W_i \sim \text{Bernoulli}[1/\{1 + \exp(\theta_i)\}]\), and finally \(Y_i = X_i \beta_Y + 0.5W_i + \epsilon_{i2}\) where \(\epsilon_{i1}\) and \(\epsilon_{i2}\) are independent standard Gaussian. Following Belloni et al. (2014), we set the structure model as \((\beta_Y)_j \propto 1/j^2\) for \(j = 1, \ldots, p\); for the propensity model, we consider both a ‘very sparse’ propensity model \((\beta_W)_j \propto 1/j\), and also a ‘dense’ propensity model \((\beta_W)_j \propto 1/\sqrt{j}\). A potential criticism of this simulation design is that the signal is perhaps unusually sparse (in the fifth column of Table 3, adjusting for differences in the two most important covariates removes 93% of the bias that is associated with all the covariates); moreover, we note that all the important coefficients of both \(\beta_Y\) and \(\beta_W\) are close to each other in terms of their indices; thus, the effect of using a correlated design may be mitigated. Thus, we also ran a moderately sparse two-stage simulation, just like that above one, except that we now used choices for \(\beta_Y\) as above, the only difference being that we shifted the indices of the \(\beta_s\), multiplying them by 23 mod \(p\) (for example, the harmonic set-up now has \((\beta)_j \propto 1/[10 + \{23(j - 1) \bmod p\}]\)). Here, we drew the treatment assignments from a well-specified logistic model, \(W_i \sim \text{Bernoulli}[1/\{1 + \exp(-\Sigma_{100}^{10}|X_{ij}|/40)\}]\).

To test the robustness of all the methods considered, we also ran a misspecified simulation. Here, we first drew \(X_i \sim \mathcal{N}(0, I_{p \times p})\), and defined latent parameters \(\theta_i = \log[1 + \exp(-2 - \)
Table 3. Root-mean-squared error $\sqrt{\mathbb{E}[(\hat{\tau} - \tau)^2]}$ in the sparse two-stage setting†

| Method                          | Results for the following experiments: |
|---------------------------------|----------------------------------------|
|                                 | Propensity model sparse                | Propensity model dense                |
|                                 | First-stage signal strength $\|\beta_W\|_2 = 1$ | First-stage signal strength $\|\beta_W\|_2 = 4$ | First-stage signal strength $\|\beta_W\|_2 = 1$ | First-stage signal strength $\|\beta_W\|_2 = 4$ |
|                                 | Structure signal strength $\|\beta_Y\|_2$ | Structure signal strength $\|\beta_Y\|_2$ | Structure signal strength $\|\beta_Y\|_2$ | Structure signal strength $\|\beta_Y\|_2$ |
| Naive                           | 0.963                                  | 1.701                                  | 6.804                                  | 0.535                                  |
| Elastic net                     | 0.277                                  | 0.648                                  | 0.619                                  | 0.202                                  |
| Approximate balance             | 0.195                                  | 0.585                                  | 2.313                                  | 0.198                                  |
| Approximate residual balance    | 0.109                                  | 0.287                                  | 0.326                                  | 0.107                                  |
| Inverse propensity weight       | 0.484                                  | 0.932                                  | 3.722                                  | 0.301                                  |
| AIPW                            | 0.164                                  | 0.374                                  | 0.384                                  | 0.130                                  |
| Weighted elastic net            | 0.174                                  | 0.686                                  | 0.708                                  | 0.132                                  |
| TMLE elastic net                | 0.161                                  | 0.234                                  | 0.227                                  | 0.122                                  |
| Double selection + OLS          | 0.081                                  | 0.115                                  | 0.123                                  | 0.092                                  |

†We used $n = 1000$, $p = 2000$ and $\rho = 0.5$. All numbers are averaged over 400 simulation replications.

2($X_1$)$\{}/0.915. We then drew $W_i \sim \text{Bernoulli}\{1 - \exp(-\theta_i)\}$, and finally $Y_i = (X_i)_1 + \ldots + (X_i)_{10} + \theta_i(2W_i - 1)/2 + \varepsilon_i \sim \mathcal{N}(0, 1)$. We varied $n$ and $p$. This simulation setting, loosely inspired by the classic programme evaluation data set of LaLonde (1986), is illustrated in Fig. 2(b); note that the average treatment effect on the treated is much greater than the overall average treatment effect here.

5.3. Results
In the first two experiments, for which we report results in Tables 1 and 2, the outcome model $Y|X$ is reasonably sparse, whereas the propensity model has overlap but is not in general sparse. In relative terms, this appears to hurt the double-selection method most. Meanwhile, in Table 3, we find that the method of Belloni et al. (2014) has excellent performance—as expected—when both the propensity and the outcome models are sparse. However, if we make the problem somewhat more difficult (Table 4), its performance decays substantially, and double selection lags both the approximate residual balancing and the propensity-based methods in its performance.

Generally, we find that the balancing performs substantially better than propensity score weighting, with or without direct covariate adjustment. We also find that combining direct covariate adjustment with weighting does better than weighting on its own, irrespectively of whether the weighting is based on balance or on the propensity score. In these experiments, the weighted elastic net and TMLE also somewhat improve over AIPW.

Encouragingly, approximate residual balancing also does a good job in the misspecified setting from Table 5. It appears that our stipulation that the approximately balancing weights (5) must
be non-negative (i.e. $\gamma_i \geq 0$) helps to prevent our method from extrapolating too aggressively. Conversely, least squares with model selection does not perform well although both the outcome and the propensity models are sparse; apparently, it is more sensitive to the misspecification here. Perhaps the reason that AIPW and TMLE do not do as well here is that there are very strong linear effects.

Table 4. Root-mean-squared error $\sqrt{E[(\hat{\tau} - \tau)^2]}$ in the moderately sparse two-stage setting†

| Method                  | Results for the following experiments: |
|-------------------------|----------------------------------------|
|                         | Beta model dense | Beta model harmonic | Beta model moderately sparse | Beta model very sparse |
|                         | Auto-covariance (ρ) 0.5 | Auto-covariance (ρ) 0.9 | Auto-covariance (ρ) 0.5 | Auto-covariance (ρ) 0.9 | Auto-covariance (ρ) 0.5 | Auto-covariance (ρ) 0.9 |
| Naive                   | 1.236          | 2.659          | 1.088          | 1.938          | 0.951          | 1.096          | 0.814          | 0.814          |
| Elastic net             | 1.075          | 1.235          | 0.631          | 0.597          | 0.225          | 0.132          | 0.098          | 0.096          |
| Approximate balance     | 1.153          | 1.125          | 1.034          | 0.994          | 0.921          | 0.717          | 0.827          | 0.569          |
| Approximate residual balance | 0.994          | 1.146          | 0.614          | 0.554          | 0.219          | 0.131          | 0.109          | 0.100          |
| Inverse propensity weight AIPW | 1.236          | 2.645          | 1.084          | 1.932          | 0.950          | 1.091          | 0.814          | 0.813          |
| Elastic net             | 1.082          | 1.231          | 0.629          | 0.597          | 0.224          | 0.132          | 0.098          | 0.096          |
| Weighted elastic net    | 1.089          | 1.234          | 0.624          | 0.597          | 0.225          | 0.132          | 0.099          | 0.096          |
| TMLE elastic net        | 1.065          | 1.233          | 0.624          | 0.597          | 0.225          | 0.132          | 0.999          | 0.096          |
| Double selection + OLS  | 1.312          | 2.659          | 1.064          | 1.938          | 0.629          | 0.204          | 0.092          | 0.097          |

†We used $n = 600$, $p = 2000$ and scaled the signal such that $\|\beta\|_2 = 1$. All numbers are averaged over 400 simulation replications.

Table 5. Root-mean-squared error $\sqrt{E[(\hat{\tau} - \tau)^2]}$ in the misspecified setting†

| Method                  | Results for the following experiments: |
|-------------------------|----------------------------------------|
|                         | n = 400                               | n = 1000            |
|                         | p = 100      | p = 200      | p = 400      | p = 800      | p = 1600     | p = 100      | p = 200      | p = 400      | p = 800      | p = 1600     |
| Naive                   | 1.734       | 1.738       | 1.734       | 1.736       | 1.747       | 1.724       | 1.679       | 1.706       | 1.698       | 1.720       |
| Elastic net             | 0.446       | 0.468       | 0.492       | 0.517       | 0.540       | 0.376       | 0.380       | 0.389       | 0.401       | 0.413       |
| Approximate balance     | 0.523       | 0.582       | 0.609       | 0.656       | 0.700       | 0.297       | 0.327       | 0.379       | 0.395       | 0.464       |
| Approximate residual balance | 0.249       | 0.276       | 0.270       | 0.295       | 0.310       | 0.168       | 0.175       | 0.176       | 0.179       | 0.194       |
| Inverse propensity weight AIPW | 1.060       | 1.081       | 1.111       | 1.154       | 1.189       | 0.831       | 0.831       | 0.874       | 0.875       | 0.940       |
| Weighted elastic net    | 0.340       | 0.359       | 0.377       | 0.406       | 0.425       | 0.249       | 0.254       | 0.261       | 0.266       | 0.285       |
| TMLE elastic net        | 0.347       | 0.365       | 0.381       | 0.407       | 0.428       | 0.273       | 0.275       | 0.282       | 0.286       | 0.301       |
| Double selection + OLS  | 0.285       | 0.292       | 0.301       | 0.320       | 0.339       | 0.250       | 0.250       | 0.246       | 0.244       | 0.246       |

†All numbers are averaged over 400 simulation replications.
We evaluate coverage of confidence intervals in the ‘many-cluster’ setting for various choices of $\beta$, $n$ and $p$; results are given in Table 6. Coverage is generally better with more overlap ($\eta = 0.25$) rather than less ($\eta = 0.1$), and with sparser choices of $\beta$. Moreover, coverage rates appear to improve as $n$ increases, suggesting that we are in a regime where the asymptotics from corollary 3 are beginning to apply.

**Table 6.** Coverage for approximate residual balancing confidence intervals as constructed in corollary 3, with data generated as in the many-cluster setting†

| $n$   | $p$  | Results for $\beta_j \propto 1 \{ j \leq 10 \}$ | Results for $\beta_j \propto 1/j$ | Results for $\beta_j \propto 1/j^2$ |
|-------|------|---------------------------------------------|---------------------------------|----------------------------------|
|       |      | $\eta = 0.25$                             | $\eta = 0.25$                   | $\eta = 0.25$                   |
| 400   | 800  | 0.95                                       | 0.97                           | 0.87                            |
| 400   | 1600 | 0.92                                       | 0.94                           | 0.88                            |
| 400   | 3200 | 0.90                                       | 0.94                           | 0.86                            |
| 800   | 800  | 0.94                                       | 0.97                           | 0.92                            |
| 800   | 1600 | 0.95                                       | 0.95                           | 0.91                            |
| 800   | 3200 | 0.94                                       | 0.96                           | 0.94                            |
| 1600  | 800  | 0.96                                       | 0.96                           | 0.97                            |
| 1600  | 1600 | 0.95                                       | 0.98                           | 0.94                            |
| 1600  | 3200 | 0.95                                       | 0.96                           | 0.95                            |

†We scaled the signal such that $\|\beta\|_2 = 3$. The target coverage is 0.95. All numbers are averaged over 1000 simulation replications.

6. Discussion

In this paper, we introduced approximate residual balancing as a method for unconfounded average treatment effect estimation in high dimensional linear models. Under standard assumptions from the high dimensional inference literature, our method enables $\sqrt{n}$-consistent inference of the average treatment effect without any structural assumptions on the treatment assignment mechanism beyond overlap.

Widely used doubly robust methods, pioneered by Robins et al. (1994) and studied further by several researchers (e.g. Belloni et al. (2017), Farrell (2015), Kang and Schafer (2007), Scharfstein et al. (1999), Robins et al. (2007), Tan (2010) and Van Der Laan and Rubin (2006)) approach this problem by trying to estimate two nuisance components, the outcome model and the propensity model. These methods then achieve consistency if either nuisance component is itself consistently estimated, and they achieve semiparametric efficiency if both components are estimated sufficiently fast. In contrast, our method ‘bets’ on linearity twice: both in fitting the lasso and in attempting to balance away its bias. In well-specified linear models, this bet allows us to extend the class of problems considerably for $\sqrt{n}$-consistent inference of average treatment effects is possible; thus, if a practitioner believes linearity to be a reasonable assumption in a given problem, our estimator may be a promising choice.

We end by mentioning two important questions that are left open by this paper. First, it would be important to develop a better understanding of how to choose the tuning parameter $\zeta$ in expression (5) that trades off bias and variance in our balancing. Results from theorems 1 and 2 provide some guidance on choosing $\zeta$ (via a constraint form characterization); however, in our experiments, we achieved good performance by simply setting $\zeta = \frac{1}{2}$ everywhere. The difficulty in choosing $\zeta$ is that we are trying to trade off an observable quantity (sampling variance) against an
unobservable quantity (residual bias) and so cannot rely on simple methods like cross-validation that require unbiased estimates of the loss criterion that we are trying to minimize. It would be of considerable interest either to devise a data-adaptive choice for $\zeta$, or to understand why a fixed choice $\zeta = \frac{1}{2}$ appears to achieve systematically good performance.

It would also be interesting to extend our approach to generalized linear models, where there is a non-linear link function $\psi$ for which $E[Y_i(c)|X_i=x] = \psi(x/\beta_c)$. In causal inference applications, this setting frequently arises when the outcomes $Y_{i}^{\text{obs}}$ are binary, and we are willing to work with a logistic regression model. In this case, the first-order error component from using a pilot estimator $\hat{\beta}_c$ for estimating $\mu_c$ with a plug-in estimator $n_c^{-1}\sum_{i=1}^{n_c} \psi(X_i/\hat{\beta}_c)$ would be of the form $n_c^{-1}\sum_{i=1}^{n_c} \gamma_i (Y_{i}^{\text{obs}} - \psi(X_i/\hat{\beta}_c))$. An analogue to proposition 1 then suggests using an estimator

$$
\hat{\mu}_c = \frac{1}{n_c} \sum_{\{W_i=1\}} \psi(X_i/\hat{\beta}_c) + \sum_{\{W_i=0\}} \gamma_i \{Y_{i}^{\text{obs}} - \psi(X_i/\hat{\beta}_c)\},
$$

$$
\gamma = \arg \min \left\{ \zeta \left( \frac{1}{n_t} \sum_{\{W_i=1\}} \psi'(X_i/\hat{\beta}_c)X_i - \sum_{\{W_i=0\}} \tilde{\gamma}_i \psi'(X_i/\hat{\beta}_c) \right)^2 + (1 - \zeta) \sum_{\{W_i=0\}} \tilde{\gamma}_i^2 \psi'(X_i/\hat{\beta}_c) \text{ subject to } \sum_{\{W_i=0\}} \tilde{\gamma}_i = 1, \ 0 \leq \tilde{\gamma}_i \leq n_c^{-2/3} \right\}.
$$

However, for brevity, we leave a study of this estimator to further work.

**Acknowledgements**

We are grateful for detailed comments from Jelena Bradic, Edgar Dobriban, Bryan Graham, Chris Hansen, Nishanth Mundru, Jamie Robins and José Zubizarreta, and for discussions with seminar participants at the Atlantic Causal Inference Conference, Boston University, the Columbia Causal Inference Conference, Columbia University, Cowles Foundation, the Econometric Society Winter Meeting, the Evidence in Governance and Politics standards meeting, the European Meeting of Statisticians, the International Conference on Machine Learning, the Institute for Operations Research and the Management Sciences, Stanford University, the University of North Carolina at Chapel Hill, University of Southern California and the World Statistics Congress.

**References**

Abadie, A. and Imbens, G. W. (2006) Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74, 235–267.

Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2017) Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85, 233–298.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014) Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.*, 81, 608–650.

Belloni, A., Chernozhukov, V. and Wang, L. (2011) Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98, 791–806.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37, 1705–1732.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W. and Candès, E. J. (2015) SLOPE: adaptive variable selection via convex optimization. *Ann. Appl. Statist.*, 9, 1103–1140.

Cai, T. T. and Guo, Z. (2017) Confidence intervals for high-dimensional linear regression: minimax rates and adaptivity. *Ann. Statist.*, 45, 615–646.

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1976) Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615–620.
Chan, K. C. G., Yam, S. C. P. and 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.

Chen, S. S., Donoho, D. L. and Saunders, M. A. (1998) Atomic decomposition by basis pursuit. *SIAM J. Scient. Comput.*, 20, 33–61.

Chen, X., Hong, H. and Tarozzi, A. (2008) Semiparametric efficiency in GMM models with auxiliary data. *Ann. Statist.*, 36, 808–843.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2017) Double/debiased machine learning for treatment and structural parameters. *Econometr. J.*, to be published.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2009) Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96, 187–199.

Deville, J.-C. and Särndal, C.-E. (1992) Calibration estimators in survey sampling. *J. Am. Statist. Ass.*, 87, 376–383.

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

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

Graham, B., Pinto, C. and Egel, D. (2012) Inverse probability tilting for moment condition models with missing data. *Rev. Econ. Stud.*, 79, 1053–1079.

Graham, B., Pinto, C. and Egel, D. (2016) Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST). *J. Bus. Econ. Statist.*, 34, 288–301.

Hahn, J. (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66, 315–331.

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

Hastie, T., Tibshirani, R. and Wainwright, M. (2015) *Statistical Learning with Sparsity: the Lasso and Generalizations*. Boca Raton: CRC.

Heckman, J. J., Ichimura, H. and Todd, P. (1998) Matching as an econometric evaluation estimator. *Rev. Econ. Stud.*, 65, 261–294.

Hellerstein, J. and Imbens, G. (1999) Imposing moment restrictions by weighting. *Rev. Econ. Statist.*, 81, 1–14.

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

Hirano, K., Imbens, G., Ridder, G. and Rubin, D. (2001) Combining panels with attrition and refreshment samples. *Econometrica*, 69, 1645–1659.

Hirshberg, D. A. and Wager, S. (2017) Balancing out regression error: efficient treatment effect estimation without smooth propensities. *Preprint arXiv:1712.00038*. Columbia University, New York.

Hotz, V. J., Imbens, G. W. and Klerman, J. A. (2006) Evaluating the differential effects of alternative welfare-to-work training components: a reanalysis of the California GAIN program. *J. Lab. Econ.*, 24, 521–566.

Imai, K. and Ratkovic, M. (2014) Covariate balancing propensity score. *J. R. Statist. Soc. B*, 76, 243–263.

Imbens, G. W. and Rubin, D. B. (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press.

Imbens, G., Spady, R. and Johnson, P. (1998) Information theoretic approaches to inference in moment condition models. *Econometrica*, 66, 333–357.

Javanmard, A. and Montanari, A. (2014) Confidence intervals and hypothesis testing for high-dimensional models. *J. Mach. Learn. Res.*, 15, 2869–2909.

Javanmard, A. and Montanari, A. (2015) De-biasing the lasso: optimal sample size for Gaussian designs. *Preprint arXiv:1508.02757*. Marshall School of Business, University of Southern California, Los Angeles.

Kang, J. and Schafer, J. (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, 22, 523–529.

LaLonde, R. J. (1986) Evaluating the econometric evaluations of training programs with experimental data. *Am. Econ. Rev.*, 76, 604–620.

McCaffrey, D. F., Ridgeway, G. and Morral, A. R. (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Meth.*, 9, 403–425.

MOSEK (2015) *MOSEK Rmosek Package*. MOSEK, Copenhagen. (Available from [http://docs.mosek.com/8.0/rmosek.pdf](http://docs.mosek.com/8.0/rmosek.pdf).)

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

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

Robins, J., Li, L., Mukherjee, R., Tchetgen Tchetgen, E. and van der Vaart, A. (2017) Minimax estimation of a functional on a structured high dimensional model. *Ann. Statist.*, 45, 1951–1987.

Robins, J. M. and Rotnitzky, Y. (1997) Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statist. Med.*, 16, 285–319.

Robins, J. and Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Statist. Ass.*, 90, 122–129.
Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, 89, 846–866.

Robins, J., Rotnitzky, A. and Zhao, L. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Ass.*, 90, 106–121.

Robins, J., Sued, M., Lei-Gomez, Q. and Rotnitzky, A. (2007) Comment: Performance of double-robust estimators when inverse probability weights are highly variable. *Statist. Sci.*, 22, 544–559.

Rosenbaum, P. R. (2002) *Observational Studies*. New York: Springer.

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

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

Rubin, D. B. (2008) For objective causal inference, design trumps analysis. *Ann. Appl. Statist.*, 2, 808–840.

Rudelson, M. and Vershynin, R. (2013) Hanson-Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.*, 18, 1–9.

Rudelson, M. and Zhou, S. (2013) Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory*, 59, 3434–3447.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Am. Statist. Ass.*, 94, 1096–1120.

Su, W. and Candes, E. (2016) SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, 44, 1038–1068.

Tan, Z. (2010) Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97, 661–682.

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

Tsiatis, A. (2007) *Semiparametric Theory and Missing Data*. New York: Springer Science and Business Media.

Van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42, 1160–1202.

Van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007) Super learner. *Statist. Appl. Genet., Molec. Biol.*, 6, article 25.

Van der Laan, M. J. and Rose, S. (2011) *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer Science and Business Media.

Van der Laan, M. J. and Rubin, D. (2006) Targeted maximum likelihood learning. *Int. J. Biostatist.*, 2, no. 1.

Wager, S., Du, W., Taylor, J. and Tibshirani, R. J. (2016) High-dimensional regression adjustments in randomized experiments. *Proc. Natl. Acad. Sci. USA*, 113, 12673–12678.

Wang, Y. and Zubizarreta, J. R. (2017) Approximate balancing weights: characterizations from a shrinkage estimation perspective. *Preprint arXiv:1705.00998*, Columbia University, New York.

Westreich, D., Lessler, J. and Funk, M. J. (2010) Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J. Clin. Epidemiol.*, 63, 826–833.

Zhang, C.-H. and Zhang, S. S. (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc. B*, 76, 217–242.

Zhao, Q. (2016) Covariate balancing propensity score by tailored loss functions. *Preprint arXiv:1601.05890*, University of Pennsylvania, Philadelphia.

Zhao, Q. and Percival, D. (2017) Entropy balancing is doubly robust. *J. Causal Inf.*, 5, no. 1.

Zhu Y. and Bradic, J. (2016) Linear hypothesis testing in dense high-dimensional linear models. *J. Am. Statist. Ass.*, to be published.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67, 301–320.

Zubizarreta, J. R. (2015) Stable weights that balance covariates for estimation with incomplete outcome data. *J. Am. Statist. Ass.*, 110, 910–922.

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Proofs’.