

# AUGMENTED BALANCING ESTIMATORS OF THE AVERAGE TREATMENT EFFECT ON THE TREATED IN CROSS-SECTIONAL AND PANEL DESIGNS

APOORVA LAL

**ABSTRACT.** Recent developments in the use of machine learning methods for causal inference typically target the average treatment effect (ATE) and frequently rely on estimating a propensity score using nonparametric regression learners and inverting it to plug into the doubly-robust IPW score. In observational studies, however, the ATE is frequently difficult to target because of the failure of overlap, which is compounded by the inversion step; researchers often target the average treatment effect on the treated (ATT) in such cases. We propose a unified framework for augmented balancing estimators for the ATT in a wide variety of research designs used by in the social sciences, including cross-sectional, two-period difference in differences, and longitudinal data settings. These estimators combine flexible nonparametric outcome models for the response surface with with balancing weights that directly targets in-sample covariate balance using an empirical risk minimization (ERM) procedure, and provide a software implementation in the R package `aba1`. In simulation studies, we find that balancing weights outperform a wide variety of commonly used estimators, including AIPW estimators that involve inverting a propensity score. We conclude with empirical applications.

# 1. Introduction

For many social scientific questions, researchers face the challenging task of disentangling the causal effects from confounding factors that may influence both the treatment assignment and the outcome of interest. Researchers are increasingly careful about stating their target estimand and the assumptions that allows them to identify it using data. For example, in cross sectional settings, it is well known that nonparametric identification requires both unconfoundedness and overlap. While flexible covariate adjustment with many controls and flexible functional form has been popularized as ‘double machine learning’ (Chernozhukov et al., 2018), overlap is discussed less often, and is particularly challenging in observational research where many units may have negligible probability of treatment. As a result, the Average Treatment effect on the Treated (ATT) is a commonly targeted estimand in a wide variety of observational settings where the propensity score is may be zero for a substantial share of units.

In this paper, we propose a unified framework for augmented balancing estimators for the ATT in a wide variety of research designs used by in the social sciences. The framework begins with the key ingredient of the modern semiparametric (‘double machine learning’) approach to causal inference, which is the use of multiple flexible models to fit both the outcome (henceforth *outcome model*) and treatment assignment (henceforth *treatment model*), which has the appealing ‘double-robustness’ property of yielding consistent estimates if one of the two is correctly specified<sup>1</sup>. It departs from the standard Augmented Inverse-Propensity Weighting (AIPW) approach (Robins, Rotnitzky, and Zhao, 1994; Hahn, 1998) by replacing the inverse propensity score, which greatly magnifies bias under misspecification (Kang and Schafer, 2007), with *balancing scores* that directly optimize for in-sample balance and have been show to possess superior finite sample properties (Hainmueller, 2012; Zubizarreta, 2015; Zhao and Percival, 2016; Hirshberg and Wager, 2021). With this recipe in hand, we propose augmented balancing estimators for the ATT in cross-sectional, two-period difference in differences, and panel data settings, and provide an accompanying R package `abal`.

We contribute to a growing literature on the use of machine learning methods for causal inference, and a closely related literature on the use of balancing (‘synthetic control’) weights for panel data. The former strand of the literature has largely focussed on estimating causal effects in cross-sectional settings under unconfoundedness (Robinson, 1988; Robins, Rotnitzky, and Zhao, 1994; Hahn, 1998; Chernozhukov et al., 2018), and has only recently been extended to related research designs with panel data (Abadie, 2005; Sant’Anna and Zhao, 2020; Ben-Michael, Feller, and Rothstein, 2021). The latter literature focuses on specific problems that arise in the panel data setting (Abadie and Gardeazabal, 2003; Abadie, Diamond, and Hainmueller, 2010; Doudchenko and Imbens, 2016; Athey et al., 2021). By nesting these problems in a common framework and providing an exposition of the core

---

<sup>1</sup>Parameters that characterize the outcome and treatment model are nuisance parameters. This property is generalized into a family of ‘Neyman Orthogonal’ scores by Chernozhukov et al. (2018) and Chernozhukov et al. (2022) that are robust to local perturbations in the nuisance parameters around their true values.

identification assumptions and their implications, and producing a performant software implementation that implements a wide variety of these methods through a common interface, we hope to make this powerful set of methods more accessible to applied researchers, as well as to spur future methodological research in the use of combining multiple modelling strategies to conduct robust causal inference.

## 2. Methodology

We observe IID draws of units from a large population with measurements  $(\mathbf{Y}_{it}, \mathbf{W}_{it}, \mathbf{X}_i) \in \mathbb{R} \times \{0, 1\} \times \mathcal{X}$  where  $\mathbf{Y}_{it}$  is a  $T$ -vector of outcomes,  $\mathbf{W}_{it}$  is a treatment dummy, and  $\mathbf{X}_i$  is a  $d$ -vector of baseline covariates taking values in the covariate space  $\mathcal{X} \subseteq \mathbb{R}^d$ .  $i \in [N]$  indexes units and  $t \in [T]$  indexes time periods, with  $T$  being the total number of time periods. When  $T = 1$ , the data is said to be *cross-sectional*, and we drop the second subscript; while for  $T > 1$ , it is said to be *panel*. We partition the units into Treated  $\mathcal{T} := \{i : W_i = 1\}$  and Control  $\mathcal{C} := \{i : W_i = 0\}$ , with  $n_t = |\mathcal{T}|$  and  $n_c = |\mathcal{C}|$  corresponding with the number of treatment and control units respectively. We define the fraction of treated units in the sample as  $\rho = n_t/n$ .

Some additional pieces of notation will be used repeatedly henceforth. Two key *nuisance functions* are the propensity score and outcome model. The **propensity score** is the conditional probability of receiving the treatment  $\pi(\mathbf{x}) := \mathbb{E}[W|\mathbf{X}] = \Pr(W = 1|\mathbf{X} = \mathbf{x})$ . Also, define, for each treatment level  $w$ , a corresponding regression function  $\mu^{(w)}(\mathbf{x}) = \mathbb{E}[Y^{(w)}|\mathbf{X} = \mathbf{x}]$ , which we also refer to as the **outcome model**. Estimators for these functions,  $\hat{\pi}(\cdot), \hat{\mu}(\cdot)$  are to be fit using flexible nonparametric/machine-learning estimators. A superscript on a nuisance function  $\hat{\mu}^{(w)}$  means that the model is fit on the sub-sample with treatment level  $w$  only.

In the current paper, the estimand is the Average Treatment Effect on the Treated (ATT), which is the average difference between potential outcomes for the treated group during treated periods (which the only period in the cross sectional setting, second period in the two-period setting, and an arbitrary share of periods in the panel setting).

$$\tau^{\text{ATT}} = \mathbb{E}\left[Y_{iT}^{(1)}|W_{iT} = 1\right] - \mathbb{E}\left[Y_{iT}^{(0)}|W_{iT} = 1\right] \quad (2.1)$$

The ATT is a more modest but realistic goal in observational work wherein one takes assignment as given and seeks to estimate the treatment effect (average difference between treated and control potential outcomes) for treated units. It is often an alternative to targeting the average treatment effect (ATE) in the presence of units with very small propensity scores  $\pi(\mathbf{x}) \approx 0$ , which makes the computation of  $\mathbb{E}[Y^{(1)}]$  infeasible, or at best extremely imprecise thanks to the presence of  $1/\pi(\mathbf{x})$  in the semiparametrically efficient variance (Hahn, 1998). The treated potential outcome  $Y_{it}^{(1)}$  is observed for treated units, so the learning problem amounts to constructing an estimator for the counterfactual potential

outcome for treated units  $\mathbb{E} \left[ Y_{iT}^{(0)} | W_{iT} = 1 \right]$  in the absence of treatment, which we refer to as  $\hat{\xi}$  henceforth.

Different estimators for the control potential outcome mean  $\hat{\xi} := \hat{\mathbb{E}} \left( Y_{iT}^{(0)} | W_{iT} = 1 \right)$  can be constructed under different identification assumptions depending on the data at hand, which we turn to next.

**2.1. Cross-sectional Setting.** In this section, we drop the time subscript because  $T = 1$ . Since we only have access to one period, and as such only observe  $Y^1 \forall i \in \mathcal{T}$  and  $Y^0 \forall i \in \mathcal{C}$ , imputing counterfactual means necessitates versions of well-known **selection-on-observables** assumptions (see Imbens (2004) for a review).

**Assumption 1 (No Interference / Stable Unit Treatment Value Assumption (SUTVA)).** This assumption asserts that a unit’s realized outcome is generated as

$$Y_i = W_i Y^{(1)} + (1 - W_i) Y^{(0)} \quad (2.2)$$

This imposes that each unit  $i$  has two potential outcomes  $Y^{(1)}, Y^{(0)}$  corresponding with unit  $i$ ’s treatment states, and not any other units. This rules out unit  $i$ ’s outcomes being affected by unit  $j$ ’s assignment  $\forall i \neq j \in [N]$ .

The corresponding individual level treatment effect is  $\tau_i := Y^{(1)} - Y^{(0)}$ , which is unidentifiable because only one of the two potential outcomes are revealed for any given unit  $i$  due to the Fundamental Problem of Causal Inference. We are interested in estimating the ATT, which is the average of  $\tau_i$ s in the treated group  $\mathcal{T}$ .

**Assumption 2 (Unconfoundedness for Controls).**

$$Y^{(0)} \perp\!\!\!\perp W | \mathbf{X}_i \quad (2.3)$$

We assume that the control potential outcome  $Y^{(0)}$  is independent of the treatment conditional on baseline covariates. This can be weakened to **mean independence for control**  $\mathbb{E} [Y^{(0)} | W, \mathbf{X}] = \mathbb{E} [Y^{(0)} | \mathbf{X}]$ , however it is difficult to construct examples where mean independence holds but unconfoundedness doesn’t.

**Assumption 3 (Weak Overlap).**

$$\Pr(W = 1 | \mathbf{X}) < 1 \text{ a.s.} \quad (2.4)$$

This imposes that there are no covariate profiles  $\mathbf{X}$  wherein the treatment is deterministic, since that would mean that we cannot find comparable control units and must therefore rely on extrapolation.

Under assumptions 1, 2, and 3, the counterfactual mean under control for treated units  $\widehat{\mathbb{E}}[Y^0|W = 1]$  is nonparametrically identified and can be constructed using observed data (Heckman, Ichimura, and Todd, 1998). Estimators for  $\widehat{\mathbb{E}}[Y^0|W = 1]$  take one of three forms: Regression (R), Weighting (W), or Hybrid approaches (H).

**2.1.1. Outcome Modelling.** The Regression approach involves fitting an outcome model  $\mu^{(0)}(\mathbf{X}_i)$  on control outcomes, and projecting it on all units. This gives us the following regression estimator for the average control potential outcome for the treated

$$\widehat{\xi}^{\text{OM}} := \frac{1}{n_t} \sum_{i \in \mathcal{T}} \widehat{\mu}^{(0)}(\mathbf{X}_i) \quad (2.5)$$

where the imputed potential outcomes  $\widehat{\mu}^{(0)}$  are averaged over treated units  $\mathcal{T}$  alone, with the control units used solely to fit  $\widehat{\mu}^{(0)}$ . Consistency hinges on the model  $\widehat{\mu}^{(0)}$  being correctly specified, and as such flexible regression estimators are preferred. Outcome modelling also be viewed as a re-weighting estimator (Kline, 2011; Chattopadhyay and Zubizarreta, 2021; Bruns-Smith et al., 2023). For the ATT, the weight for each control unit  $i$  is  $\gamma_i^{\text{Reg}} = (\mathbf{X}_i - \bar{\mathbf{X}}_c)'(\widehat{\Sigma}_0)^{-1}(\bar{\mathbf{X}}_t - \bar{\mathbf{X}}_c)$  where  $\bar{\mathbf{X}}_t, \bar{\mathbf{X}}_c$  are covariate means in the treatment and control groups respectively, and  $\widehat{\Sigma}_0$  is the scaled covariance matrix in the control group. These weights exactly balance the elements of  $\mathbf{X}_i$  across  $\mathcal{T}$  and  $\mathcal{C}$  if  $\frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})} = \mathbf{X}'_i \boldsymbol{\gamma}$ .

**2.1.2. Reweighting.** An important implication of the unconfoundedness assumption 2 is that treatment and control units are *balanced* on observable covariates. In a seminal paper, Rosenbaum and Rubin (1983) show that a scalar  $\gamma$  is said to be a *balancing score* if it satisfies conditional independence between treatment and covariates; in other words,  $Y^{(0)} \perp\!\!\!\perp W | \mathbf{X}$  is equivalent to  $Y^{(0)} \perp\!\!\!\perp W | \gamma$ , which performs dimension reduction<sup>2</sup>. We seek weights  $\gamma$  that satisfy the following conditions

$$\text{Population Balance : } \mathbb{E}[W\mathbf{X}] = \mathbb{E}[\gamma(1 - W)\mathbf{X}] \quad (2.6)$$

$$\text{Sample Balance : } \frac{1}{n_t} \sum_{i \in \mathcal{T}} \mathbf{X}_i = \frac{1}{n_c} \sum_{i \in \mathcal{C}} \gamma_i \mathbf{X}_i \quad (2.7)$$

where the latter sample balance condition 2.1 is the feasible condition that can be verified using observed data. This allows us to construct estimators for the average control potential outcome for the treated

$$\xi^{\text{wt}} = \sum_{i \in \mathcal{C}} \gamma_i Y_i \quad (2.8)$$

<sup>2</sup>this is a modified version of the statement in RR1983, which is interested in the Average Treatment Effect (ATE) and therefore focuses on full unconfoundedness  $Y^{(1)}, Y^{(0)} \perp\!\!\!\perp W | \mathbf{X}$

Methods for constructing  $\gamma$  include *modelling*, such as by fitting a model for the propensity score  $\pi(\mathbf{x})$ , or by solving the above balance condition directly. Matching estimators (Heckman, Ichimura, and Todd, 1998; Smith and Todd, 2005; Abadie and Imbens, 2006) are a special case of the above formulation where the weights  $\gamma_i$  are chosen separately for each treated observation  $i$  and are constrained to be nonzero for the  $k$  best matches.

To estimate these weights, we may first want to characterize the error of a general weighting estimator for the control potential outcome  $\mathbb{E}[Y^{(0)}]$ , which takes the form 2.8. The error of this estimator can be decomposed in the vein of Ben-Michael et al. (2021) as

$$\widehat{\xi}^{\text{wt}} - \xi = \underbrace{\frac{1}{n} \sum_i (1 - W_i) \widehat{\gamma}_i \mu^{(0)}(\mathbf{x}_i) - \frac{1}{n} \sum_i W_i \mu^{(0)}(\mathbf{x}_i)}_{\text{Bias from Imbalance}} - \underbrace{\frac{1}{n} \sum_{i=1}^n (1 - W_i) \gamma_i \varepsilon_i}_{\text{Noise}} + \underbrace{\frac{1}{n} \sum_{i=1}^n W_i \widehat{\mu}^{(0)}(\mathbf{x}_i) - \xi}_{\text{Sampling}} \quad (2.9)$$

where the second two terms are a weighted average of sampling noise and sampling variation, which means that the design-conditional bias is the first term. The decomposition in 2.9 also clearly emphasizes that a weighting based estimator of the missing counterfactual is intrinsically tied to the unknown conditional mean function  $\mu^{(0)}(\mathbf{x}_i)$ : the balancing approach implicitly requires one to take a stand on the *model class*  $\mathcal{M}$  of the conditional mean  $\mu^{(0)}$  we are trying to model.

**2.1.2.1. Modelling the Propensity Score.** The Rosenbaum and Rubin (1983) balancing result shows that for the ATT, the asymptotically balancing weight is

$$\gamma_i = \frac{\mathbb{E}[W = 1 | \mathbf{X}_i]}{\mathbb{E}[W = 0 | \mathbf{X}_i]} = \frac{\mathbb{E}[W = 1 | \mathbf{X}_i]}{1 - \mathbb{E}[W = 1 | \mathbf{X}_i]}$$

The most direct and widely used implementation of this approach is via Inverse Propensity Weighting, which uses the plug-in approach to fit a (typically parametric) model for the propensity score  $\pi(\mathbf{X}) := \Pr(W = 1 | \mathbf{X})$  and plug in these predicted probabilities to construct the sample analogue of the above population quantity. Imai and Ratkovic (2014) show that using logistic  $\gamma_i^{\text{IPW}} = \widehat{\pi}(\mathbf{X}_i) / [1 - \widehat{\pi}(\mathbf{X}_i)]$  regression to estimate the propensity score has a balancing interpretation, where weights are calculated to balance a particular function of covariates. We extend the result in 2.1.

**Proposition 2.1 (Balancing interpretation of Logistic Propensity Score).**

Parametric propensity score modelling involves estimating  $\Pr(W_i = 1 | \mathbf{X}_i) = F(\mathbf{X}_i^\top \boldsymbol{\beta})$  as a linear predictor passed through a link function  $F$  where  $F(\cdot) : \mathbb{R} \rightarrow [0, 1]$ , where the most popular choice is  $F(\cdot) = \Lambda(\cdot) := \frac{\exp(\cdot)}{1 + \exp(\cdot)}$  using logit regression. This gives rise the following log-likelihood

$$\begin{aligned} \log \mathcal{L} &= \sum_{i=1}^n \{W_i \log \Lambda(\mathbf{X}_i^\top \boldsymbol{\beta}) + (1 - W_i) \log(1 - \Lambda(\mathbf{X}_i^\top \boldsymbol{\beta}))\} \quad \text{Differentiate for Likelihood Eq} \\ \frac{\partial \log \mathcal{L}}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^n \left\{ W_i \frac{\Lambda'(\mathbf{X}_i^\top \boldsymbol{\beta})}{\Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})} - (1 - W_i) \frac{\Lambda'(\mathbf{X}_i^\top \boldsymbol{\beta})}{1 - \Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})} \right\} = 0 \\ &= \sum_{i=1}^n \left\{ \frac{W_i}{\Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})} - \frac{1 - W_i}{1 - \Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})} \right\} \Lambda'(\mathbf{X}_i^\top \boldsymbol{\beta}) = 0 \end{aligned}$$

where  $\Lambda'(\cdot) =: \lambda(\cdot)$  is the density  $d\Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})/d(\mathbf{X}_i^\top \boldsymbol{\beta})$ . Imai and Ratkovic (2014) therefore conclude that estimating the propensity score by maximum likelihood balances a *particular* function of the covariates: the first derivative of the link function  $f(\mathbf{X}_i^\top \boldsymbol{\beta})$ . For logistic regression, we can claim something stronger: using properties of the logistic distribution, we can plug in  $f(\mathbf{X}_i^\top \boldsymbol{\beta}) = \Lambda'(\mathbf{X}_i^\top \boldsymbol{\beta}) = \Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})(1 - \Lambda(\mathbf{X}_i^\top \boldsymbol{\beta}))$  allows us to see that that the function being balanced across the two groups is the estimated propensity score itself

$$\sum_{i=1}^n \left\{ \frac{W_i}{\Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})} - \frac{1 - W_i}{1 - \Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})} \right\} \Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})(1 - \Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})) = 0$$

This implies that instead of balancing covariates moments (which is often the informally stated goal of reweighting), logistic regression relies entirely on the dimension reduction of covariates via the link function and balances on  $\Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})$  (and its complement  $1 - \Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})$ ). From Rosenbaum and Rubin (1983), we know that conditioning on the true propensity score is equivalent to conditioning on covariates under selection on observables, so when  $\pi(\mathbf{X})$  is correctly specified as  $\Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})$ , balancing on the logistic propensity score is sufficient. However, if the propensity score is misspecified (i.e. the true propensity score cannot be approximated by  $\Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})$ ), balancing on  $\Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})(1 - \Lambda(\mathbf{X}_i^\top \boldsymbol{\beta}))$  provides no guarantees about balance along covariates, and indeed may worsen balance since  $\Lambda(\mathbf{X}_i^\top \boldsymbol{\beta})$  is now simply a low-dimensional summary statistic of covariates, with especially high weights placed on covariates that predict treatment assignment. Therefore, solving for weights that directly balance covariates are preferable over the logistic propensity score under (likely) misspecification.

**2.1.2.2. Balancing.** In 2.1, we saw that modelling the propensity score without explicitly prioritising balance may yield poor properties under misspecification. This is related to the empirical observation that by inverting  $1 - \hat{\pi}(\mathbf{X}_i)$ , one risks magnifying estimation errors in  $\hat{\pi}$ , and therefore incurring considerable bias. An alternative set of methods seek to solve for weights  $\gamma$  by solving the *sample balance* condition 2.7 directly.

Weights that solve the sample balance condition are known as *covariate-balancing propensity scores* (CBPS) following Imai and Ratkovic (2014). These may be specified to mimic propensity scores and be characterised by a finite dimensional parameter vector  $1/\pi_\beta(\mathbf{X}_i)$ ,

as in the Imai and Ratkovic work. Alternatively, they can be characterized as solutions to a mathematical program

**Defn 2.1 (Sample Balance Program).**

$$\min_{\gamma} = \sum_{i \in \mathcal{C}} h(\gamma_i) \quad \text{s.t.} \quad (2.10)$$

$$\left| \frac{1}{n_c} \sum_{i \in \mathcal{C}} \gamma_i \phi_k(\mathbf{X}_i) - \frac{1}{n_t} \sum_{i \in \mathcal{T}} \phi_k(\mathbf{X}_i) \right| \leq \delta_k \quad k = [K] \quad (2.11)$$

$$\sum_{i \in \mathcal{C}} \gamma_i = 1 \quad (2.12)$$

$$\sum_{i \in \mathcal{C}} \gamma_i \geq 0 \quad (2.13)$$

where  $h(\gamma_i)$  is a convex loss function of the weights, and condition 2.11 asserts that covariate basis functions  $\phi_k(\cdot)$ ,  $k = \{1, \dots, K\}$  (including but not limited to moments of covariates  $\mathbf{X}_i$ ) are balanced across treatment and control groups within tolerance<sup>3</sup>  $\delta_k$ , which is chosen to be exactly zero when  $\mathbf{X}_i$  is low-dimensional but is infeasible in high dimensions. Condition 2.12 ensures that weights sum to 1 in the treated group. Condition 2.13 is an *optional* non-negativity constraint that forces the weights to be on a  $|\mathcal{C}|$  dimensional simplex and avoids extrapolation.

Intuitively, we want to solve for a set of weights that depart minimally from uniform weights of  $1/n_c$  while guaranteeing us sample balance, which motivates the term ‘minimal weights’ (Wang and Zubizarreta, 2019). This framework nests entropy balancing (Hainmueller, 2012) with  $h(x) = x \log x$ , quadratic balancing (Zubizarreta, 2015), with  $h(x) = (x - 1/n_c)^2$ . The constrained optimization problem can be posed as a Generalized Empirical Likelihood (GEL) problem (Imbens2002-mr)<sup>4</sup>. Solving the constrained optimization problem targets a  $n_c$  vector, which is computationally challenging because of the saddle-point structure of the GEL problem. However, since  $h$  is convex, the corresponding dual for the balancing problem is unconstrained and can be solved much more easily. Wang and Zubizarreta (2019)[Thm 1] show that the dual unconstrained problem takes on the following form

<sup>3</sup>which can be substantively motivated or chosen to be a scalar if the  $\mathbf{X}_i$  are scaled to be on the unit interval or normalized beforehand, which is advised for numerical stability

<sup>4</sup>where choices of  $h(\cdot)$  correspond with different choices of  $\lambda$  in the Cressie-Read divergence

$$l_{\lambda}(p, q) := \frac{1}{\lambda(1 + \lambda)} \sum_{i=1}^n p_i \left[ \left( \frac{p_i}{q_i} \right)^{\lambda} - 1 \right]$$

. with base weights  $q_i$  set to uniform. Entropy balancing corresponds with the Exponential tilting  $\lambda = 0$ , which minimizes the Kullback-Liebler distance between uniform and EL weights. Quadratic balancing corresponds with the Euclidian likelihood estimator ( $\lambda = -2$ ).



$$\begin{aligned} \boldsymbol{\lambda}^* &= \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} -\frac{1}{n_c} \sum_{i \in \mathcal{C}} (\rho(\boldsymbol{\phi}(\mathbf{X}_i)^\top \boldsymbol{\lambda})) + \frac{1}{n_t} \sum_{i \in \mathcal{T}} \boldsymbol{\phi}(\mathbf{X}_i)^\top \boldsymbol{\lambda} + \underbrace{|\boldsymbol{\lambda}|^\top \boldsymbol{\delta}}_{\text{imbalance / regularisation}} \\ \boldsymbol{\gamma}^* &= \rho'(\boldsymbol{\phi}(\mathbf{X}_i)^\top \boldsymbol{\lambda}^*) \end{aligned}$$

Where  $\boldsymbol{\lambda}$  is a  $K$ -vector of dual variables (lagrange multipliers) from the constrained optimization problem, and  $\rho(\cdot)$  is a transformation of the loss function  $h(\cdot)$  from the constrained problem, and the final weights  $\boldsymbol{\gamma}$  are given by evaluating the  $\rho'(\cdot)$  function at the solution coefficients  $\boldsymbol{\lambda}^*$ . Special cases include  $h(x) = x \log x \rightarrow \rho(x) = \exp(x - 1) = \rho'(x)$  for entropy balancing and  $h(x) = (x - 1/n_c)^2 \implies \rho(x) = -x^2/4 + x/n_c, \rho'(x) = -x^2 + 1/n_c$  for quadratic balancing. The dual formulation can be solved using modern optimization tools using either first-order (gradient descent and related) as well as Gauss-Newton or quasi-Newton methods. We include a performant implementation in the accompanying R package `abal`.

**2.1.3. Hybrid.** A final category of methods combine Regression and Weighting methods and aim for consistency when either the outcome model or weights are well specified, which is well known as the *double-robustness* property (Bang and Robins, 2005). When stated in terms of score functions (i.e. moment conditions that characterize the solution for the estimand), an estimator is said to be *Neyman-Orthogonal* if its directional derivative with respect to its nuisance parameters (outcome model and propensity score) is zero (Chernozhukov et al., 2018). Doubly-robust estimators are based on Neyman-orthogonal scores.

The most well known of these is the Augmented Inverse-Propensity Weighting (AIPW) estimator for the ATE (Robins, Rotnitzky, and Zhao, 1994; Hahn, 1998), and its analogue for the ATT, which are constructed from Neyman-Orthogonal moment conditions (Chernozhukov et al., 2018). This approach proposes the following estimator for the ATT

$$\widehat{\tau}^{\text{AIPW}} = \underbrace{\frac{1}{n^{-1} \sum_i W_i} \frac{1}{n} \sum_i W_i \{Y_i - \widehat{\mu}^{(0)}(\mathbf{X}_i)\}}_{n_t/n =: \widehat{\rho}} - (1 - W_i) \frac{\widehat{\pi}(\mathbf{X}_i)}{1 - \widehat{\pi}(\mathbf{X}_i)} \{Y_i - \widehat{\mu}^{(0)}(\mathbf{X}_i)\} \quad (2.14)$$

where we define the share of treated units  $n_t/n$  as  $\widehat{\rho}$  and note that the inverse propensity weights  $\widehat{\pi}(\mathbf{X}_i)/(1 - \widehat{\pi}(\mathbf{X}_i))$  are used to re-weight the residuals for the control units.

The expression in 2.14 can be partitioned into estimators for the two average potential outcomes for the treated as follows

$$\widehat{\tau}^{\text{AIPW}} = \underbrace{\frac{1}{\widehat{\rho}} \frac{1}{n} \sum_i W_i Y_i}_{\widehat{\mathbb{E}}[Y^{(1)}|W=1]} - \underbrace{\frac{1}{\widehat{\rho}} \frac{1}{n} \sum_i \left[ \underbrace{W_i \widehat{\mu}^{(0)}(\mathbf{X}_i)}_{\text{Regression}} + \underbrace{(1 - W_i) \frac{\widehat{\pi}(\mathbf{X}_i)}{1 - \widehat{\pi}(\mathbf{X}_i)} \{Y_i - \widehat{\mu}^{(0)}(\mathbf{X}_i)\}}_{\text{Weighting}} \right]}_{\widehat{\mathbb{E}}[Y^{(0)}|W=1]} \quad (2.15)$$

$$\widehat{\xi}^{\text{AIPW}} = \underbrace{\frac{1}{\widehat{\rho}} \frac{1}{n} \sum_{i \in \mathcal{T}} \widehat{\mu}^{(0)}(\mathbf{X}_i)}_{\text{Reg}} + \underbrace{\frac{1}{\widehat{\rho}} \frac{1}{n} \sum_{i \in \mathcal{C}} \frac{\widehat{\pi}(\mathbf{X}_i)}{1 - \widehat{\pi}(\mathbf{X}_i)} \{Y_i - \widehat{\mu}^{(0)}(\mathbf{X}_i)\}}_{\text{Weighting}} \quad (2.16)$$

The final expression 2.16 clarifies that the AIPW estimator for the control potential outcome  $\widehat{\mathbb{E}}[Y^{(0)}|W = 1]$  takes on the form of a regression imputation piece for treated units and an re-weighting piece for control units, which is similar in structure to the bias-corrected matching estimator of Abadie and Imbens (2011) (wherein the latter applies uniform weights to units in the matched set for any treated observation  $i$ ).

While existing work typically uses propensity-score based weights  $\widehat{\pi}(\mathbf{X}_i)/(1 - \widehat{\pi}(\mathbf{X}_i))$  that solve for  $\widehat{\gamma}_i$  that solve the population balance condition 2.6, arguments favouring balancing weights 2.1.2.2 apply equally strongly here, which suggests a form for an *augmented*-balancing estimator.

### Defn 2.2 (Augmented Balancing Estimator).

The Augmented Balancing estimator estimates the average control potential outcome for the treated as

$$\widehat{\xi}^{\text{AUGBAL}} = \underbrace{\frac{1}{\widehat{\rho}} \frac{1}{n} \sum_{i \in \mathcal{C}} \gamma_i \{Y_i - \widehat{\mu}^{(0)}(\mathbf{X}_i)\}}_{\text{Balancing}} + \underbrace{\frac{1}{\widehat{\rho}} \sum_{i \in \mathcal{T}} \widehat{\mu}^{(0)}(\mathbf{X}_i)}_{\text{Reg augmentation}} \quad (2.17)$$

where balancing weights  $\widehat{\gamma}_i$  solve the finite-sample covariate balancing program 2.1 and the outcome model  $\widehat{\mu}^{(0)}$  is fit using flexible nonparametric regression methods.

This improves upon balancing-based approaches such as Hainmueller (2012), Imai and Ratkovic (2014), and Zhao and Percival (2016) which are doubly-robust for the ATT when the outcome model is linear, but not otherwise, and therefore hinge largely on the specification of the implied treatment model. The augmented balancing estimator 2.17 fits a separate outcome model flexibly, and therefore is doubly-robust when either the weights are correctly specified, or if the outcome model is correctly specified. It is asymptotically normal, semiparametrically efficient, and admits to a standard variance formula using the influence function (Ben-Michael et al., 2021). The standard nonparametric bootstrap, as well as multiplier bootstrap, are also available for variance estimation.

Bruns-Smith et al. (2023) show that augmented balancing that combines linear models (e.g. OLS, lasso, or ridge, potentially in some basis) with linear balancing (e.g.  $\ell_1$  or  $\ell_2$  linear loss but not entropy loss) can be interpreted as undersmoothed linear regression imputation estimator (for a given fixed choice of hyperparameter). shows that imposing the simplex constraint is equivalent to sample-trimming to units that don't have extreme propensity scores. This provides an additional motivation for a non-linear balancing function (entropy loss) and the simplex constraint; it cleanly separates the roles of the outcome model and inverse weights to extrapolate and interpolate respectively.

**2.2. Two-period Setting.** To establish connections with the estimators in 2.1, we begin with the simplest extension to cross-sectional data, wherein we have outcome data from two periods for each observation  $i$ , which is commonly known as the two-period difference-in-differences setting.

We write  $Y_{it}^{(w)}$  to denote the potential outcomes  $Y^{(0)}, Y^{(1)}$  for unit  $i$  at time  $t \in \{0, 1\}$  and  $Y_{it}$  to denote the realised outcome. We observe  $n$  IID copies of  $(Y_{i1}, Y_{i0}, W_i, \mathbf{X}_i)$ . We additionally define the difference in *realized* outcomes between the two periods for a given unit  $i$   $Y_{i1} - Y_{i0}$  as  $\Delta_i$ . Some fraction  $\rho$  of units are treated in the second period. The estimand is the ATT in the 2nd period

$$\tau^{\text{ATT2}} = \mathbb{E} \left[ Y_{i1}^{(1)} \mid W = 1 \right] - \mathbb{E} \left[ Y_{i1}^{(0)} \mid W = 1 \right]$$

As before, the treated counterfactual mean for treated units is observed, so the problem is effectively that of constructing an estimator for the average control potential outcome for treated units in the second period  $\hat{\mathbb{E}} \left[ Y_{i1}^{(0)} \mid W = 1 \right] =: \hat{\xi}$ .

Since control potential outcomes are observed for treated units for the pre-treatment period  $t = 0$ , intuitively, the identification problem seems somewhat easier, and doesn't require selection on observables assumptions such as 2. Instead, one can resort to assumptions about trends in potential outcomes.

**Assumption 4 (Parallel Trends).**

$$\mathbb{E} \left[ Y_{i1}^{(0)} - Y_{i0}^{(0)} \mid W = 1 \right] = \mathbb{E} \left[ Y_{i1}^{(0)} - Y_{i0}^{(0)} \mid W = 0 \right] \quad (2.18)$$

In words, this requires that the trends in control potential outcomes  $Y^{(0)}$  over time be identical across treatment and control groups.

**Assumption 5 (No Anticipation (Two-period version)).**

$$\mathbb{E} [Y_{i0} \mid W_i = 1] = \mathbb{E} \left[ Y_{i0}^{(0)} \mid W_i = 0 \right] \quad (2.19)$$

This requires that the treatment doesn't affect outcomes before coming into effect. This is frequently substantively justified by researchers with the claim that the timing of the intervention was as good as random.

Under assumptions 1, 4, and 5, the average control potential outcome for treated units in the second period is nonparametrically identified (Lechner, 2011). It can be estimated as

$$\widehat{\xi}^{\text{DID}} = \underbrace{\frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} Y_{i0}}_{\text{Baseline outcome for treated}} + \underbrace{\frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \Delta_i}_{\text{Trend for control}} \quad (2.20)$$

which is the baseline outcome for the treated group offset by the trend in the control group. This estimator's simplicity has led to widespread adoption in applied work. However, assumption 4 is a very strong one, and researchers frequently want to relax it or assume it hold conditional on observable covariates.

**Assumption 6 (Conditional Parallel Trends).**

$$\mathbb{E} \left[ Y_{i1}^{(0)} - Y_{i0}^{(0)} \mid W = 1, \mathbf{X} \right] = \mathbb{E} \left[ Y_{i1}^{(0)} - Y_{i0}^{(0)} \mid W = 0, \mathbf{X} \right] \quad (2.21)$$

This is a conditional version of 4, and as such requires that the trends in control potential outcomes  $Y^{(0)}$  over time be identical across treatment and control groups conditional on baseline covariates  $\mathbf{X}$ .

Abadie (2005) derives an alternative Difference-in-Differences estimator under assumptions 1, 3, 5, and 6 of the following form

$$\widehat{\tau}^{\text{ABADIE}} = \sum_{i=1}^n \Delta_i \frac{W_i - \widehat{\pi}(\mathbf{X}_i)}{1 - \widehat{\pi}(\mathbf{X}_i)} \quad (2.22)$$

$$= \underbrace{\frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \Delta_i}_{\text{Diff in Treated}} - \underbrace{\frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \Delta_i \frac{\widehat{\pi}(\mathbf{X}_i)}{1 - \widehat{\pi}(\mathbf{X}_i)}}_{\text{Wtd Diff in Control}} \quad (2.23)$$

which is clearly a re-weighted version of difference in differences. This scheme works by weighting-down the distribution of  $Y_{i1} - Y_{i0}$  for the control for those values of the covariates  $\mathbf{X}$  which are over-represented among the control (that is, with low  $\widehat{\pi}(\mathbf{X}_i)/(1 - \widehat{\pi}(\mathbf{X}_i))$ ), and conversely weighting-up  $Y_{i1} - Y_{i0}$  for those values of the covariates under-represented among the control (that is with high  $\widehat{\pi}(\mathbf{X}_i)/(1 - \widehat{\pi}(\mathbf{X}_i))$ ).

Further decomposing the second term in 2.23 gives us the following estimator for the counterfactual mean  $\widehat{\xi}$

$$\hat{\xi}^{\text{ABADIE}} = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} Y_{i0} + \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \frac{\hat{\pi}(\mathbf{X}_i)}{1 - \hat{\pi}(\mathbf{X}_i)} (Y_{i1} - Y_{i0}) \quad (2.24)$$

where the first term is the average outcome for the treated group in the pre-treatment period, and the second and third terms comprise of the trend in the control group re-weighted by IPW weights. Comparing 2.24 to the augmented AIPW estimator 2.16 is instructive: it suggests that the Abadie estimator can be viewed as an AIPW-style estimator with a compound outcome model that combines pre-treatment outcomes for the treatment units and both pre- and post-treatment outcomes for the control units. This uses both *between-unit* and *within-unit* variation for the inverse-propensity weights and outcomes respectively.

**2.2.1. Hybrid Estimators for Difference in Differences Designs.** Consistency of the Abadie estimator hinges on a well specified model for  $\hat{\pi}(\mathbf{X})$ ; its score function is not Neyman-orthogonal. Chang (2020) and Sant’Anna and Zhao (2020) propose doubly-robust difference-in-differences estimators based on Neyman-orthogonal scores of the following form:

$$\begin{aligned} \hat{\tau}^{\text{DR DID}} &= \sum_i (\Delta_i - \hat{\mu}^0(\mathbf{X}_i)) \cdot \frac{W_i - \hat{\pi}(\mathbf{X}_i)}{1 - \hat{\pi}(\mathbf{X}_i)} \\ &= \underbrace{\sum_{i \in \mathcal{T}} (\Delta_i - \hat{\mu}^0(\mathbf{X}_i))}_{\text{Diff in Treated - debiasing}} - \underbrace{\sum_{i \in \mathcal{C}} (\Delta_i - \hat{\mu}^0(\mathbf{X}_i)) \frac{\hat{\pi}(\mathbf{X}_i)}{1 - \hat{\pi}(\mathbf{X}_i)}}_{(\text{Diff in Control - debiasing}) \times \text{weights}} \end{aligned}$$

where  $\hat{\mu}^0(\cdot)$  is an imputation model for the trend  $\mathbb{E} [Y_{i1}^{(0)} - Y_{i0}^{(0)}]$  fit on control units only. This appends the Abadie estimator with an outcome model, which is an estimate of the imputed change  $\Delta_i$  learned from the control units. This implies the following counterfactual mean estimator is

$$\hat{\xi}^{\text{DR DID}} = \sum_{i \in \mathcal{T}} Y_{i0} - \hat{\mu}^0(\mathbf{X}_i) + \sum_{i \in \mathcal{C}} (\Delta_i - \hat{\mu}^0(\mathbf{X}_i)) \frac{\hat{\pi}(\mathbf{X}_i)}{1 - \hat{\pi}(\mathbf{X}_i)} \quad (2.25)$$

where the first term is the average outcome for the treated group in the pre-treatment period minus predicted change in the trend, and the second term is a weighted average of the difference between realized change  $\Delta_i$  minus predicted change in the trend.

As with the cross-sectional setting, the IPW weights  $\hat{\pi}(\mathbf{X})/(1 - \hat{\pi}(\mathbf{X}))$  in 2.24 and 2.25 guarantee asymptotic balance, but not necessarily in-sample balance. Furthermore, in difference in differences settings, treated units are frequently quite unusual in terms of covariates

relative to control units, which makes overlap challenging and means that the propensity score is difficult to specify correctly.

Therefore, an alternative would be to use the balancing program 2.1 to solve for weights that guarantee in-sample balance directly.

**Defn 2.3 (Augmented Balancing Estimator for Difference-in-Differences).**

The Augmented Balancing estimator estimates the average control potential outcome for the treated  $\widehat{\mathbb{E}} \left[ Y_{i1}^{(0)} \mid W = 1 \right]$  as

$$\widehat{\xi}^{\text{AUGBAL DID}} = \sum_{i \in \mathcal{T}} Y_{i0} - \widehat{\mu}^0(\mathbf{X}_i) + \sum_{i \in \mathcal{C}} \gamma_i (\Delta_i - \widehat{\mu}^0(\mathbf{X}_i)) \quad (2.26)$$

where balancing weights  $\widehat{\gamma}_i$  solve the finite-sample covariate balancing program 2.1 and the outcome model  $\widehat{\mu}^0$  is fit using flexible nonparametric regression methods.

The corresponding estimator for the ATT is

$$\widehat{\tau}^{\text{AUGBAL}} = \sum_{i \in \mathcal{T}} (\Delta_i - \widehat{\mu}^0(\mathbf{X}_i)) - \sum_{i \in \mathcal{C}} \gamma_i (\Delta_i - \widehat{\mu}^0(\mathbf{X}_i)) \quad (2.27)$$

The main limitation of two-period difference in differences is that the strategy hinges entirely on the untestable parallel trends assumption; conditional parallel trends is undoubtedly weaker but similarly untestable. When *panel data* is available (with  $T \geq 3$ ) researchers may use pre-treatment outcomes as a form of *placebo* check. One may evaluate the plausibility of this assumption by plotting pre-treatment differences between treated and control units to visually evaluate parallel trends, and run dynamic specifications of lags and leads of the treatment regressed on the outcome, with the belief that null effect on leads is suggestive of plausible parallel trends. If pre-treatment differences are present between treatment and control units, additional balancing methods are called for, which we turn to next.

Secondly, the parallel trends assumption is inherently tied to a specific functional form for the outcome  $Y$  (e.g. parallel trends levels implies non-parallel trends in logs, and so on). The Changes-in-Changes approach of Athey and Imbens (2006) relaxes the strong functional-form dependence of the parallel trends assumption in favour of a time-invariant monotonicity assumption, but has seen limited use among applied researchers (potentially because of the somewhat harder to interpret monotonicity assumption and challenges in the incorporation of covariates).

**2.3. Panel Data Setting.** Finally, we consider the panel setting where we have  $T > 2$  periods of data. We focus our attention on settings where the treatment is *absorbing* and treated units are assigned to treatment at one point in time  $T_0 + 1$ <sup>5</sup> For each unit, we observe

<sup>5</sup>we restrict our exposition to a single treatment time for notational simplicity, but the approach outlined below can readily be extended to the staggered setting where treatment is assigned at one of  $\mathcal{G}$  distinct adoption

$T$ -vectors of outcome  $\mathbf{Y}_{it}$ , treatment  $\mathbf{W}_{it}$ , and (an optional) time-invariant covariate vector  $\mathbf{Z}_i$ .

Since treatment is only ever implemented on a subset of observations at time  $T_0 + 1$ , in a slight abuse of notation we define a treatment indicator *without a time subscript*  $W_i := \sum_{t=1}^T \mathbb{1}_{W_{it} > 0}$  for each unit that takes a value of 1 only for units that were treated. Finally, we partition the observations into Controls  $\mathcal{C} := \{i : W_i = 0\}$  and Treated  $\mathcal{T} := \{i : W_i = 1\}$  with corresponding sizes  $N_0, N_1$  respectively. To define potential outcomes in this setting, we make a simplifying assumption that links treatments to outcomes in a restricted manner.

**Assumption 7 (No Carryover).**

We assume that for  $t$ -vectors  $\mathbf{W}_i$  and  $\mathbf{W}'_i$  such that the assignment in the  $t$ -th period is the same  $W_{it} = W'_{it}$ , the potential outcomes are the same  $W_{it} = W'_{it} \implies Y_{it}^{\mathbf{W}} = Y_{it}^{\mathbf{W}'}$ . This allows us to index the potential outcomes by a single binary argument  $w$  and write  $Y_{it}^{(w)}$  as opposed to the entire treatment history  $Y_{it}^{(\mathbf{w})}$  (Imbens and Arkhangelsky, 2021), which in turn lets us represent the realised outcome using the familiar switching equation  $Y_{it} = W_{it}Y_{it}^{(1)} + (1 - W_{it})Y_{it}^{(0)}$ . The corresponding contemporaneous treatment effect is the difference between two potential outcomes  $\tau_{it} = Y_{it}^{(1)} - Y_{it}^{(0)}$ , which is unidentifiable thanks to the FPCI. Instead, our estimand is the ATT at each time after treatment  $T_0 < t < T$ .

$$\tau_t^{\text{ATTp}} = \mathbb{E} \left[ Y_{it}^{(1)} - Y_{it}^{(0)} \mid W_i = 1 \right] ; T_0 < t \leq T$$

This is the estimand in a variety of panel data settings, including the difference-in-differences and synthetic control and matrix completion and event-study literature. These can be averaged over time for an analogue to the cross-sectional  $\tau^{\text{ATT}}$ .

$$\tau^{\text{ATT}} = \frac{\sum_{(i,t): W_{it}=1} [Y_{it}^{(1)} - Y_{it}^{(0)}]}{\sum_{i,t} W_{it}}$$

Since  $\mathbb{E}[Y_{it}^{(1)} \mid W_i = 1]$  is identified for each treated unit, the estimation problem involves estimating the counterfactual control potential outcome  $Y_{it}^{(0)}$ . The control potential outcomes can be written as a matrix  $\mathbf{Y}^0$  with blocked structure, where the bottom right of the matrix is missing for the treated units  $\mathcal{T}$ .

---

dates/cohorts  $G_i \in \mathcal{A}$ , where  $\mathcal{G} = \mathcal{T} \cup \infty = \{1, \dots, T, \infty\}$ , where  $G_i = \infty$  denotes units that were never treated.

$$\begin{pmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,T_0} & Y_{1,T} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,T_0} & Y_{2,T} \\ \vdots & & & & \vdots \\ Y_{N_0,1} & Y_{N_0,2} & \dots & Y_{N_0,T_0} & Y_{N_0,T} \\ \vdots & & & & ? \\ Y_{N,1} & Y_{N,2} & \dots & Y_{N,T_0} & ? \end{pmatrix} \equiv \begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,T_0} & \dots & Y_{1,T} \\ X_{2,1} & X_{2,2} & \dots & X_{2,T_0} & \dots & Y_{2,T} \\ \vdots & & & & \dots & \vdots \\ X_{N_0,1} & X_{N_0,2} & \dots & X_{N_0,T_0} & \dots & Y_{N_0,T} \\ \vdots & & & & \dots & ? \\ \underbrace{X_{N1} & X_{N2} & \dots & X_{NT_0}}_{\text{pre-treatment outcomes}} & \dots & ? \end{pmatrix} =: \left( \begin{array}{c|c} \mathbf{X}^0 & \mathbf{y}^n \\ \hline \mathbf{X}^1 & ? \end{array} \right) \quad (2.28)$$

where we stack the  $N_0$  outcome vectors for the control group followed by  $N_1$  outcome vectors for the treated group, where the last  $T - T_0$  elements of the treated groups' outcome vectors are missing. We label pre-treatment outcomes for the control and treatment groups as  $\mathbf{X}^0, \mathbf{X}^1$  respectively to emphasize that pre-treatment outcomes serve the role of covariates in the sequel, and control units' post-treatment outcomes are stacked to form a matrix  $\mathbf{y}^n$ .

To construct estimators to fill in missing entries in  $\mathbf{Y}^0$ , we must make one of three broad categories of assumptions: (1) parallel trends, (2) latent factor model for control potential outcomes, or (3) unconfoundedness. The first corresponds with the most natural extension of the two-period parallel trends assumption 4.

**Assumption 8 (Parallel trends and additive separability of time and unit effects).**

$$\mathbb{E} \left[ Y_{it}^{(0)} - Y_{it'}^{(0)} | W_i = 1 \right] = \mathbb{E} \left[ Y_{it}^{(0)} - Y_{it'}^{(0)} | W_i = 0 \right] \quad \forall t \neq t'$$

This assumption imposes that in the counterfactual where treatment had not been implemented, the average outcomes for ever-treated groups would have evolved in parallel with the outcomes for the never-treated groups. The parallel trends assumption is frequently paired with the following representation for the control potential outcome  $Y_{it}^{(0)} = \alpha_i + \gamma_t + \varepsilon_{it}$ , which asserts that the potential outcome is additively separable into unit effects  $\alpha_i$  and time effects  $\gamma_t$ .

Assumption 8 is the most widely used in applied research and is typically used to motivate the following two-way fixed effects regression

$$Y_{it} = \tau W_{it} + \alpha_i + \gamma_t + \varepsilon_{it} \quad (2.29)$$

where  $\tau$  is interpreted as the ATT. This representation implies the following estimator for the control potential outcome

$$\hat{\xi}_{it}^{\text{RegImpute}} := \hat{Y}_{it}^0 = \hat{\alpha}_i + \hat{\gamma}_t$$



In staggered adoption settings, several recent papers have pointed out that the estimate  $\hat{\tau}$  obtained from the above regression does not uncover the ATT or a convex-weighted average of ATTs when treatment effects  $\tau_{it}$  are heterogeneous and may in fact put negative weights on the ATT in some cohorts (Goodman-Bacon, 2018; Callaway and Sant’Anna, 2020). The inconvenient negative-weighting property of the two-way fixed effects specification under staggered adoption can be remedied by fitting the fixed-effects regression 2.29 on control observations alone  $\{i, t : W_{it} = 0\}$ , which purges the data of the comparisons that generate negative weighting in the naive two-way fixed effects regression (Borusyak, Jaravel, and Spiess, 2022; Liu, Wang, and Xu, 2021; Gardner, 2022).

**Assumption 9 (Latent Factor Model for control potential outcomes).**

Following the setup in Abadie, Diamond, and Hainmueller (2010), Xu (2017), and Ben-Michael, Feller, and Rothstein (2021), assume there are  $J$  unknown latent time-varying factors  $\boldsymbol{\mu}_t = \{\mu_{jt}\} \in \mathbb{R}^T, j = 1, \dots, J$ , and each unit has an unknown set of factor loadings  $\boldsymbol{\phi}_i \in \mathbb{R}^J$ . The latent factor model asserts that control potential outcomes are generated as

$$Y_{it}^{(0)} = \sum_{j=1}^J \phi_{ij} \mu_{jt} + \varepsilon_{it} \equiv \mathbf{Y}_i^{(0)} = \boldsymbol{\phi}_i \odot \boldsymbol{\mu}_t + \varepsilon_i$$

This is heuristically equivalent to a low-rank assumption on control potential outcomes  $\mathbf{Y}^{(0)}$  where  $J < N, T$ . These latent factors can be estimated using the complete data [i.e. entire time series for control units and pre-treatment data for treated units], with rank  $J$  estimated using a (tailored) cross-validation procedure, as proposed in Xu (2017), or via Nuclear-norm penalization as proposed in Athey et al. (2021). The latent factor model assumption nests the two-way fixed-effects assumption for the control potential outcome as a special case.

An influential series of papers uses this assumption to motivate the synthetic control method (Abadie and Gardeazabal, 2003; Abadie, Diamond, and Hainmueller, 2010), constructs an estimator for the control potential outcome using *between-unit* correlations across control potential outcomes in the pre-treatment period.

$$\boldsymbol{\gamma}^{\text{SC}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \left\| \bar{\mathbf{X}}_{.,t}^{1'} - \gamma_0 - \mathbf{X}^{0'} \boldsymbol{\gamma} \right\|_2^2 + \lambda_1 \|\boldsymbol{\gamma}\|_1 + \lambda_2 \|\boldsymbol{\gamma}\|_2$$

In words, we perform a regularized regression of the  $T_0$  vector of pre-treatment average outcome for the treated group  $(\bar{\mathbf{X}}_{.,t}^1)_{t=1}^{T_0}$  on the  $T_0 \times N_0$  matrix pre-treatment outcomes for the control group  $\mathbf{X}^{0\top}$ <sup>6</sup> with an optional intercept  $\gamma_0$ , which allows for level-differences between the two groups as in difference in differences. This approach is a leading example of the *vertical* regression (so named by Athey et al. (2021) because it uses vertical information in the potential outcome matrix : correlation between  $\mathbf{X}_0$  and  $\mathbf{X}_1$ , to impute potential outcomes). This then gives us the Vertical Regression (Synthetic Control) estimator for

<sup>6</sup>The original proposal constrains the coefficients  $\boldsymbol{\gamma}$  be on the simplex ( $\gamma_i \geq 0, \sum_{i \in C} \gamma_i = 1$ ) to avoid extrapolation bias

the control potential outcome which multiplies the synthetic control weights  $\gamma$  with post-treatment outcomes for control units  $\mathbf{y}^n$

$$\hat{\xi}_t^{\text{VR}} = \langle \mathbf{y}^n, \hat{\gamma} \rangle$$

**Assumption 10 (Unconfoundedness given pre-treatment outcomes).**

$$Y_{it}^{(0)} \perp\!\!\!\perp W_i | \mathbf{Y}_{i,1:T_0} \quad \forall t > T_0$$

where  $\mathbf{Y}_{i,1:T_0} = (Y_{i1}, \dots, Y_{iT_0})$  is a  $T_0$ -vector of pre-treatment outcomes for unit  $i$ .

Assumption 10 is equivalent to the unconfoundedness assumption from the cross-sectional setting 2 with pre-treatment outcomes as covariates. It is therefore used to motivate an unconfoundedness based approach that uses *horizontal* regression (so named because it uses over-time dependence in the outcome for control outcomes in the potential outcome matrix: correlation between  $\mathbf{y}_0^0$  and  $\mathbf{X}^0$ , to impute potential outcomes).

A horizontal regression approach involves solving the following regularized regression problem

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|\mathbf{y}^n - \mathbf{X}^0 \beta\|_2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2$$

In words, we regress the average post-treatment outcome for each control unit on  $S$ -th order lags (in the pre-treatment period). This then gives us the Horizontal Regression estimator for the control potential outcome. This then gives us the Horizontal Regression (Time Series) estimator for the control potential outcome which multiplies the autoregressive coefficients  $\beta_s$  with pre-treatment outcomes for treated units  $\mathbf{X}^1$

$$\hat{\xi}^{\text{HR}} = \langle \bar{\mathbf{X}}_{\cdot,t}^{-1'}, \hat{\beta} \rangle$$

**2.3.1. Augmented Balancing for Panel Data.** Two notable recent proposals attempt to combine the strengths of vertical and horizontal regression. The *augmented synthetic control method* (AugSynth) (Ben-Michael, Feller, and Rothstein, 2021) combines regularized (ridge) regressions for both Vertical and Horizontal Regression to construct a counterfactual  $\hat{\xi}^{\text{AugSynth}} = \hat{\mu}_{it}(\mathbf{X}_1) + \sum_{i \in \mathcal{C}, t > T_0} \hat{\gamma}_i^{\text{SC}} (Y_{it} - \hat{\mu}_{it}(\mathbf{X}_1))$ .

Similarly, the *synthetic difference in differences* (SDID) (Arkhangelsky et al., 2020) method fits two sets of weights: time weights that equalize average post-treatment outcomes for control units with average pre-treatment outcomes for control units, and (synthetic control) unit weights, that equalize average pre-treatment outcome for control units with average pre-treatment outcomes for treated units. Indeed, the latter is equivalent to the former with unregularized OLS as the outcome model. Both methods make substantial progress in

pooling the strengths of cross-sectional and over-time dependencies. However, both learn vertical and horizontal regression in isolation: the estimation of the horizontal regression (time weights) is independent of the estimation of the vertical regression (unit weights), which potentially misses out on useful information if there are dependencies across the two dimensions, for example if each unit's outcome follows an autoregressive process.

To accommodate this possibility, we propose estimating an outcome model that seeks to learn a low-rank approximation of *both* vertical and horizontal regressions, and estimate unit weights on the residuals to pick up weaker factors that were missed out by the outcome model.

**Defn 2.4 (Augmented Balancing estimator for Panel Data).**

We therefore use both pre-treatment data for the treated unit  $\mathbf{X}_1$  and pre and post-treatment  $\mathbf{X}_0, \mathbf{y}_n$  data for the control unit to train our outcome model: a matrix completion estimator that uses nuclear-norm minimization to estimate a low-rank approximation of the true untreated potential outcome matrix  $\mathbf{Y}^0$ . This motivates the following estimator which combines the Matrix completion outcome model (Athey et al., 2021) with balancing weights.

$$\hat{\xi}^{\text{AugBal}} = \overbrace{\hat{\mu}_{it}^{\text{MC}}(\mathbf{X}, \mathbf{y}_n)}^{\text{MC Imputation}} + \sum_{i \in \mathcal{C}, t > T_0} \overbrace{\hat{\gamma}_i (Y_{it} - \hat{\mu}_{it}^{\text{MC}}(\mathbf{X}, \mathbf{y}_n))}^{\text{Reweighted debiasing term}} \quad \text{where}$$

$$\hat{\mu}_{it}^{\text{MC}} := \underset{\mathbf{L}, \Gamma, \Delta}{\text{argmin}} \left[ \frac{1}{\mathcal{O}} \|\mathbf{Y} - \mathbf{L} - \Gamma \mathbf{1}'_T + \mathbf{1}'_N \Delta\|_F^2 + \lambda \|\mathbf{L}\|_* \right]$$

where our outcome model imputation  $\hat{\mu}(\cdot)$  combines low-rank imputation  $\mathbf{L}$ , unit, and time fixed effects  $\Gamma, \Delta$  to construct an imputed  $\hat{\mathbf{Y}}^0$ , and  $\gamma^i$  are weights constructed using vertical regression.

Notable cases of  $\gamma_i$  include (1) synthetic control, which uses simplex regression to approximately equate average outcomes in the treatment group before treatment adoption with weighted average outcomes in the control group, and (2) entropy regularized balancing weights, which uses a covariate balancing propensity score to approximately equate average outcomes in the treatment group before treatment adoption with weighted average outcomes in the control group, and collapses to difference in differences if parallel trends does indeed hold. The latter is true because entropy loss  $-\sum_{i \in \mathcal{C}} \gamma_i \log \gamma_i$  is minimised when each unit receives equal weight  $1/n_0$ ; so minimising entropy loss subject to pre-treatment balance constraints effectively modifies difference-in-differences weights as little as possible to satisfy pre-treatment balance constraints. In stark contrast, synthetic control weights are generically sparse and mimic a matching estimator by choosing a handful of units with positive weights. The choice between (1) synthetic control and (2) entropy regularized weights should be based on whether the researcher believes sparsity is plausible and whether they prefer a small number of interpretable weights (which favors synthetic control) versus minimal modifications to difference in differences (which favours entropy weights).

### 3. Simulation Studies

In the following section, we conduct simulation studies for the cross-sectional, difference in differences, and panel setting to evaluate the performance of augmented balancing estimators relative to other standard approaches in each setting.

#### 3.1. Cross-sectional Setting.

**3.1.1. Low-Dimensional Covariates.** We generate data with a variety of overlap, error variance, and sample sizes to thoroughly evaluate the performance of estimators proposed in sec 2.1. The DGP is adapted from the simulation DGPs in Frölich (2007) and Hainmueller (2012) with additional non-linearity. We use 6 covariates where  $X_1, \dots, X_3$  are generated from the following multivariate normal

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & -0.5 \\ -1 & -0.5 & 1 \end{pmatrix} \right)$$

$X_4 \sim \text{U}[-3, 3]$ ,  $X_5 \sim \chi_1^2$ , and  $X_6$  is Bernoulli with mean 0.5. The treatment is generated using the linear function

$$W = 1[X_1 + 2X_2 - 2X_3 - X_4 - 0.5X_5 + X_6 + \varepsilon > 0]$$

where  $\varepsilon$  is drawn from one of three distributions: (1, Strong Separation)  $\varepsilon \sim \mathcal{N}(0, 30)$ , (2, Medium Separation, Leptokurtic)  $\varepsilon \sim \chi_5^2$  scaled to mean 0.5 and variance 67.6, and (3, Weak Separation)  $\varepsilon \sim \mathcal{N}(0, 100)$ . The first shows strong separation between treatment and control group and provides a challenging case for reweighting or balancing, the second has medium separation but heavier tails, and the the third is the most favourable case for reweighting by virtue of weak separation. The outcome is generated according to one of three functional forms: (1, Linear)  $Y = X_1 + X_2 + X_3 - X_4 + X_5 + X_6 + \eta$ , (2, Quadratic)  $Y = (X_1 + X_2 + X_5)^2 + \eta$ , or (3, Non-linear)  $Y = X_1 + \sin(X_2) + 0.2X_3X_4 + \sqrt{X_5} + \eta$ , where  $\eta \sim \mathcal{N}(0, 1)$ . The treatment effect is 0.

With this DGP in hand, we compare the performance of several estimators in recovering the true effect of 0. Augmented balancing involves two choices: the loss function for balancing and functional form for the outcome model. We implement three choices: entropy balancing with a  $\ell_1$  an OLS outcome model (`Augbal(EB, OLS)`) as well as a Random Forest Outcome Model (`Augbal(EB, RF)`), as well as  $\ell_2$  loss with regularized linear regression (proposed as ‘Approximate Residual Balancing’ by Athey, Imbens, and Wager (2018),

Augbal(L2, OLS)). We benchmark these augmented balancing approaches against an extensive set of estimators: standard entropy balancing (Bal(EB), Hainmueller (2012)), standard L2 balancing (Bal(L2), Zubizarreta (2015)) augmented IPW implemented with random forest outcome and propensity models (aipw), inverse propensity weighting implemented with parametric logistic regression (ipw(logit)) and random forests (ipw(rf)), outcome modelling with regression (OM(OLS)) and random forests (OM(RF)), and naive difference in means (Diff-Means).

We plot the boxplot of the distribution of estimates and overlay Root Mean Square Error (RMSE) and Mean Absolute Deviation (MAD) in fig 1. We find that augmented balancing with entropy loss (Augbal(EB, OLS)) is the best performing of the set considered across all three of outcome and error designs, closely followed by augmented balancing with L2 loss. We report analogous results for smaller ( $n = 300$ ) and larger ( $n = 1500, n = 5000$ ) datasets in figures A1, A2, A3 in appendix A and find that the rank ordering of estimators' performance remains largely stable, except that augmented balancing incorporating more flexible outcome models (Augbal(EB, RF)) begins to outperform its OLS counterpart with larger datasets in some settings.

**3.1.2. High-Dimensional Covariates.** In appendix A.1, we apply a wide variety of cross-sectional methods to the 2016 ACIC DGPs, which vary a wide variety of parameters pertaining to the outcome and treatment assignment models including functional form, overlap, covariate importance in one or both models, and treatment effect heterogeneity in a large, high dimensional dataset with approximately 5000 observations and 60 covariates. We report mean absolute bias and RMSE in tables A1 and A2 respectively, and find that augmented IPW with flexible learners for nuisance functions and augmented balancing typically perform best across most simulation settings.

## Simulation Study

$n = 600$ ; True effect is 0

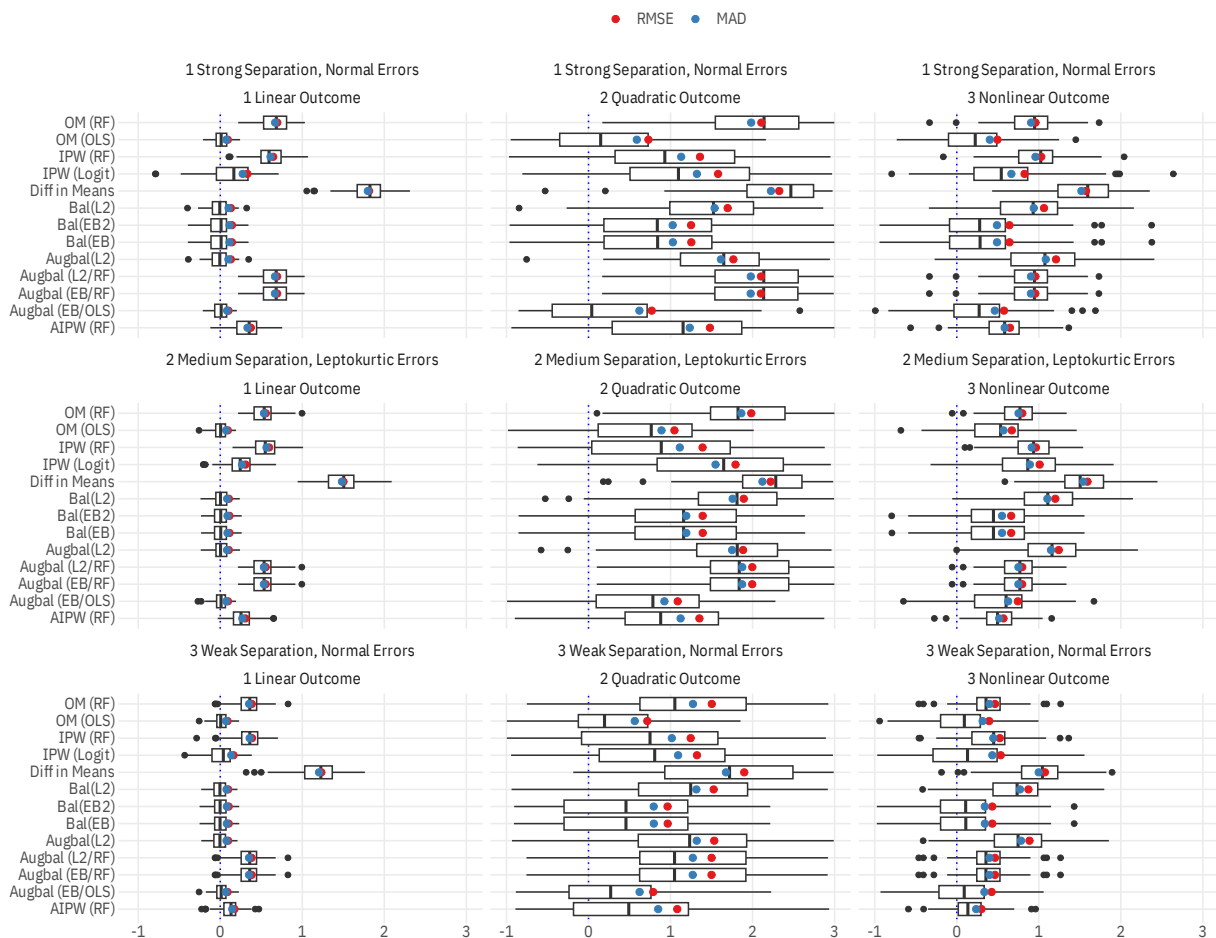


FIGURE 1. Cross-sectional simulation study with  $n = 600$

**3.2. Two-Period Simulation.** Next, we perform a simulation study for the difference-in-differences setting to evaluate the performance of estimators outlined in sec 2.2. We generate covariates  $\mathbf{X}$  from a  $p$ -variate normal (where  $p = 10, 50$  for the low and high-dimensional settings respectively) with mean zero and covariance matrix  $\Sigma$  with non-zero off-diagonal elements that are decreasing in distance  $|i - j|$ <sup>7</sup>, which emulates realistic data generating processes with correlated covariates as in the previous section.

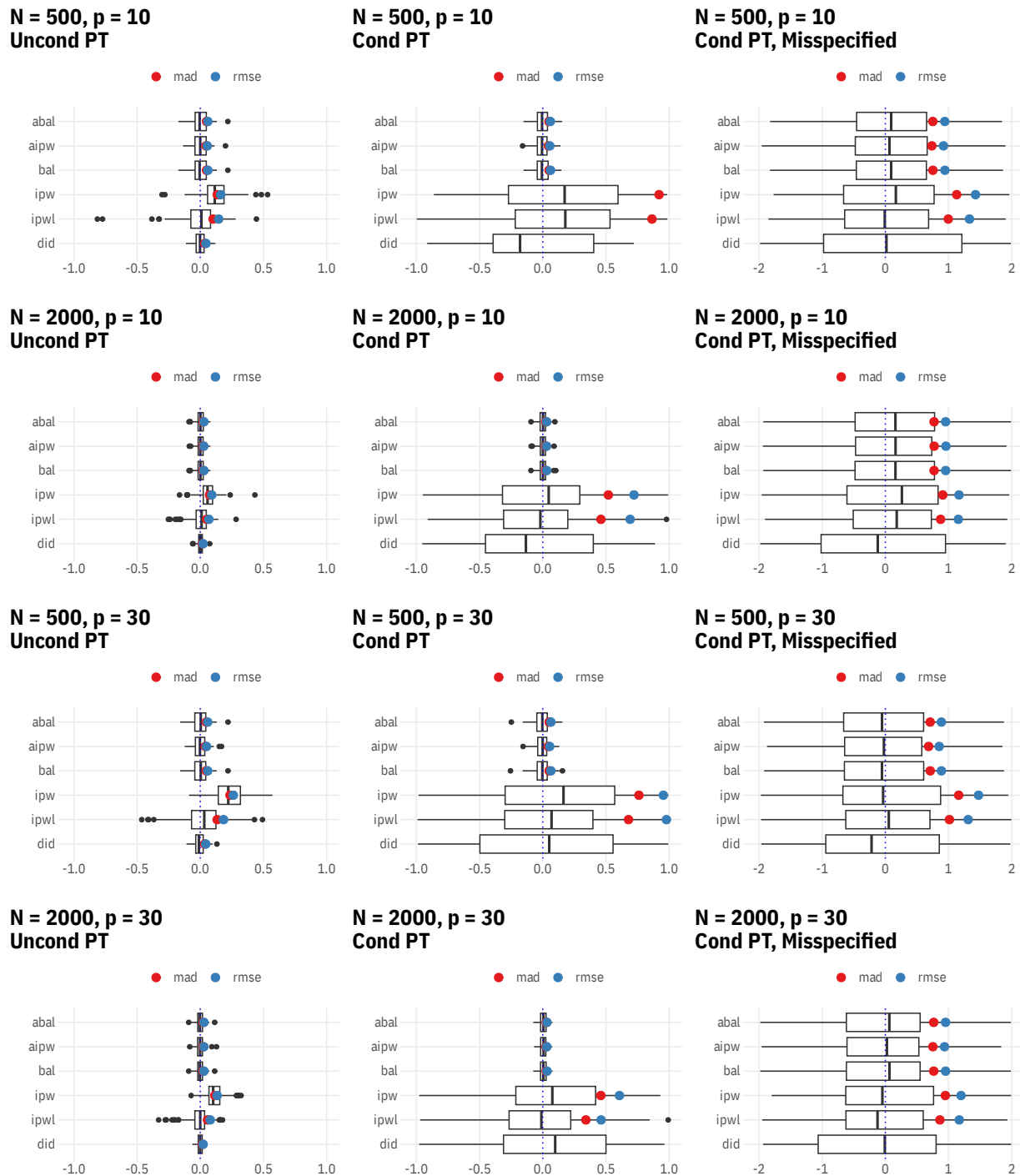
We pick  $k$  of the  $p$  covariates to be ‘active’ as a function of covariates  $\mathbf{X}$  for the outcome and selection processes, with coefficients drawn from  $U[-1, 1]$ ,  $U[-1, -1] \cup [0, 2]$ , respectively for the selection, baseline outcome  $Y_{i0}$  in the control and treated groups respectively (where the latter being drawn from  $U[0, 2]$  implies that treated units begin with a level difference in outcomes in the first period  $Y_{i0}$  relative to the control units, which is fairly typical in

<sup>7</sup>more specifically,  $\Sigma$  is characterized by a symmetric Toeplitz matrix with first row  $0.5^{(p-1)}$ . For  $p = 20$ , this implies that the covariance between  $X_1$  and  $X_2$  is 0.5,  $X_1$  and  $X_3$  is 0.25, and so on, such that a covariate  $X_j$  is non-trivially correlated with approximately 6 of its neighbouring covariates.

difference in differences applications and implies that the naive cross-sectional comparison in the second period is badly biased). Next, we set the trend between the two periods  $Y_{i1} - Y_{i0}$  to be a function of  $\rho \times U[0, 4]$ , where  $\rho = 0$  implies that unconditional parallel trends holds between the two groups, while  $\rho \neq 0$  implies that parallel trends only holds conditional on covariates. This is the setting where the way that we adjust for covariates matters a lot for consistency and efficiency. The true ATT is the difference in potential outcomes for the treated group in the second period and is set to 3.

We vary the size of the dataset  $N$ , the number of covariates  $p$ , and the magnitude of parallel trend violation  $\rho$  across simulations to evaluate the methods under consideration. We benchmark Augmented Balancing (`abal` with Entropy loss and LASSO regression for outcome modelling) against an extensive list of estimators: Naive post-treatment comparison (`naive`), Difference in Differences using cell averages (`did`), Inverse Propensity-Weighted Difference in Differences (using logistic regression `ipwl` and LASSO logit `ipw`, Abadie (2005)), Outcome modelling for the four cell means (`om` using LASSO for each cell, Heckman, Ichimura, and Todd (1998)), Augmented IPW (`aipw` with LASSO for both propensity and outcome models, Chang (2020)). We report estimates in figure 2, and find that augmented balancing collapses to OM (difference in differences) when parallel trends is true, but is adaptive to covariate adjustment when parallel trends only holds conditional on covariates. IPW has considerable bias even when unconditional parallel trends holds, which suggests that inverting a logistic propensity score often performs worse than standard difference in differences.

FIGURE 2. Difference in Differences simulation study.  $N = 500$  (Top) and  $N = 2000$  (Bottom) panel, and  $p = 10$  (low dimensional covariates) and  $p = 30$  (high dimensional covariates)





**3.3. Panel Simulation.** For the panel setting, we focus on benchmarking augmented balancing approaches with a variety of related panel data estimators using realistic data generating processes frequently encountered in the social science.

**3.3.1. Simulations with latent factors.** First, we study the properties of several standard estimators under a DGP with a single strong latent factor that influences both the outcome time series and treatment assignment under presence and absence of parallel trends with varying amounts of noise in the outcome DGP. We simulate data for  $N$  units for  $T$  periods, where the treatment applies only in the last period with a true treatment effect of zero. This means that prediction error of each estimator for the final period outcome for treated units is equal to the bias in estimating the ATT. The unobserved factor is generated as  $\mu_i \sim \mathcal{N}(i/N - 0.5, 0.5)$ , with treatment assignment following  $W_i \sim \text{Bern}(\Lambda(\mu_i))$ . The outcome time series is constructed as  $Y_{it} = \mu_i + 0.1t + \varepsilon_i$ ,  $\varepsilon_i \sim \mathcal{N}(0, \sigma)$  under parallel trends and  $Y_{it} = \mu_i \alpha_t t + \varepsilon_i$ ,  $\varepsilon_i \sim \mathcal{N}(0, \sigma)$ ,  $\alpha_t \sim \text{U}[0.05, 0.1]$  under violations of parallel trends. The latter allows for each unit to follow its own time trend scaled by its unobserved factor  $\mu_i$ , which means that in the aggregate, parallel trends does not hold. We evaluate all our methods in terms of mean absolute deviation/bias (MAD) and root mean squared error (RMSE) recovering the masked entries of the outcome matrix

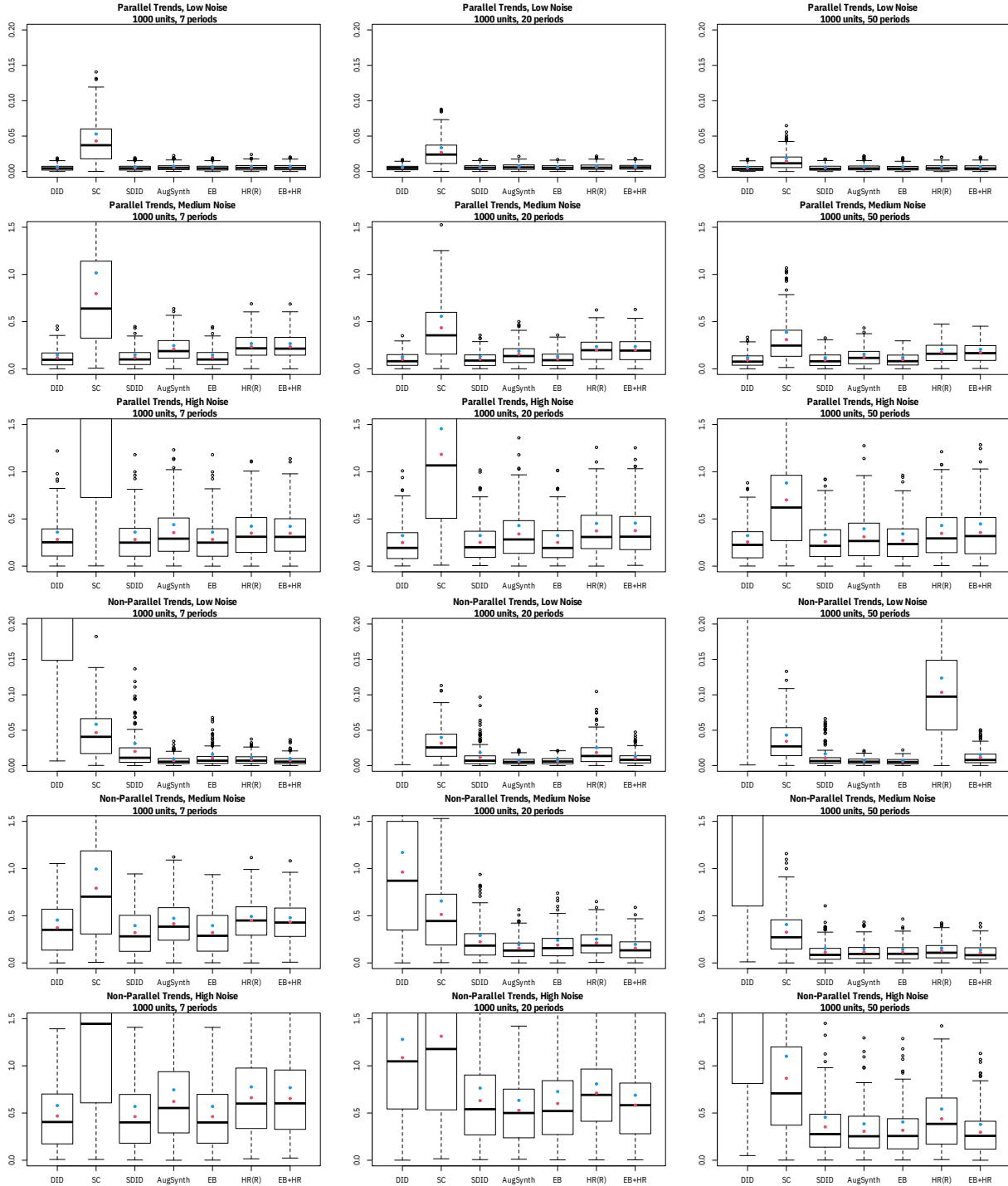
$$\text{MAD} = \frac{1}{|\mathcal{T}|} \sum_{i,t \in \mathcal{T}} |Y_{it} - \hat{Y}_{it}|$$

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{i,t \in \mathcal{T}} (Y_{it} - \hat{Y}_{it})^2}$$

We compare the performance of several popular estimators: Difference in differences estimated using two-way fixed effects (DID), Synthetic Control (SC), Synthetic Difference in Differences (SDID), Augmented Synthetic Control (AugSynth), Difference in Differences with Entropy Balancing weights (EB), Horizontal Ridge regression (HR(R)), and Augmented balancing with Horizontal Ridge regression outcome model and Entropy Balancing unit weights (EB+HR).

We report results in figure 3. We find that under parallel trends, Difference in Differences is unbiased as expected, synthetic control is not (because it erroneously throws away information by choosing non-uniform weights), while hybrid estimators such as SDID, Augsynth, and Entropy Balancing collapse to difference in differences. When the noise level is zero, estimating the latent factor  $\mu_i$  is relatively easy, while for higher levels of noise, it is considerably more challenging, and therefore the gap between SC and other methods shrinks. In the absence of parallel trends (bottom three rows), Difference in Differences is highly biased as expected, Synthetic Control has lower bias, but is strictly outperformed by other hybrid estimators. Across the board, we find that augmented balancing performs well, while with short panels, synthetic difference in differences and entropy balancing perform best (with the latter narrowly outperforming the former for short panels with large parallel trend violations and low noise).

FIGURE 3. Panel Simulation with factor structure : Absolute bias of estimators over 1000 replications. The first three rows correspond to simulations where parallel trends holds, while the next three correspond to simulations without parallel trends. In each setting, each row corresponds with different levels of noise in the outcome generating process  $\sigma$ , and each column corresponds with different lengths of panel data  $T$ . Average absolute bias is indicated by the red dot, while RMSE is indicated by the blue dot.



**3.3.2. Simulations with strong autocorrelation.** For each unit, we generate each time series with the following structure

$$Y_{it}^{(0)} = \underbrace{\sum_{p=1}^P (a_p Y_{i,t-p} + b_p \varepsilon_{i,t-p})}_{\text{ARMA piece}} + \underbrace{\alpha_i + \delta_t}_{\text{Unit/Time FE}} + \underbrace{t \cdot \psi_j}_{\text{time trend}} + \varepsilon_{it}$$

$$a_p \sim \text{U}[-\varphi, \varphi] \quad \text{Autoregressive coefficients} \quad \varphi \in \{0.01, 3\}$$

$$b_p \sim \text{U}[-\varpi, \varpi] \quad \text{Moving Average coefficients} \quad \varpi \in \{0.01, 3\}$$

$$\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2) \quad \text{Unit FE} \quad \sigma_\alpha = 2$$

$$\delta_t \sim \mathcal{N}(0, \sigma_\delta^2) \quad \text{Time FE} \quad \sigma_\delta = 5$$

$$\psi_j \sim \text{U}[-\varsigma, \varsigma]^J \quad \text{Time trend coefficient (cluster structure)} \quad \varsigma = 1, i \in [J], J \in \{5, 50\}$$

Fixing  $N = 50$  and varying the panel length  $T \in \{50, 100, 300\}$ , we generate each time series has an autoregressive moving average component (which encodes time series dependency), which can either be very strong (with  $\varpi = \varphi = 3$ ) or very weak (with  $\varpi = \varphi = 0.01$ ), unit and time-specific shocks, and time trends, which is drawn from a mixture with low rank ( $J = 5$ ) or idiosyncratic ( $J = 50$ ). This flexible configuration allows us to control whether vertical or horizontal regression is more useful in forecasting missing potential outcomes. When  $\varpi, \varphi = 3$  and time trends are not low rank ( $J = N = 50$ ), horizontal regression can be expected to perform relatively well, while when  $\varpi, \varphi = 0.01$  and  $J = 5$ , the data has low-rank structure that means vertical regression can be expected to perform better. Once we generate untreated potential outcomes  $Y_{it}^{(0)}$ , we randomly select  $K \in \{1, 10, 25\}$  units to be 'treated' at  $0.8T$  and mask the corresponding unit  $\times$  time period observations from the input matrix  $\mathbf{Y}^0$ , so the treatment effect is 0.

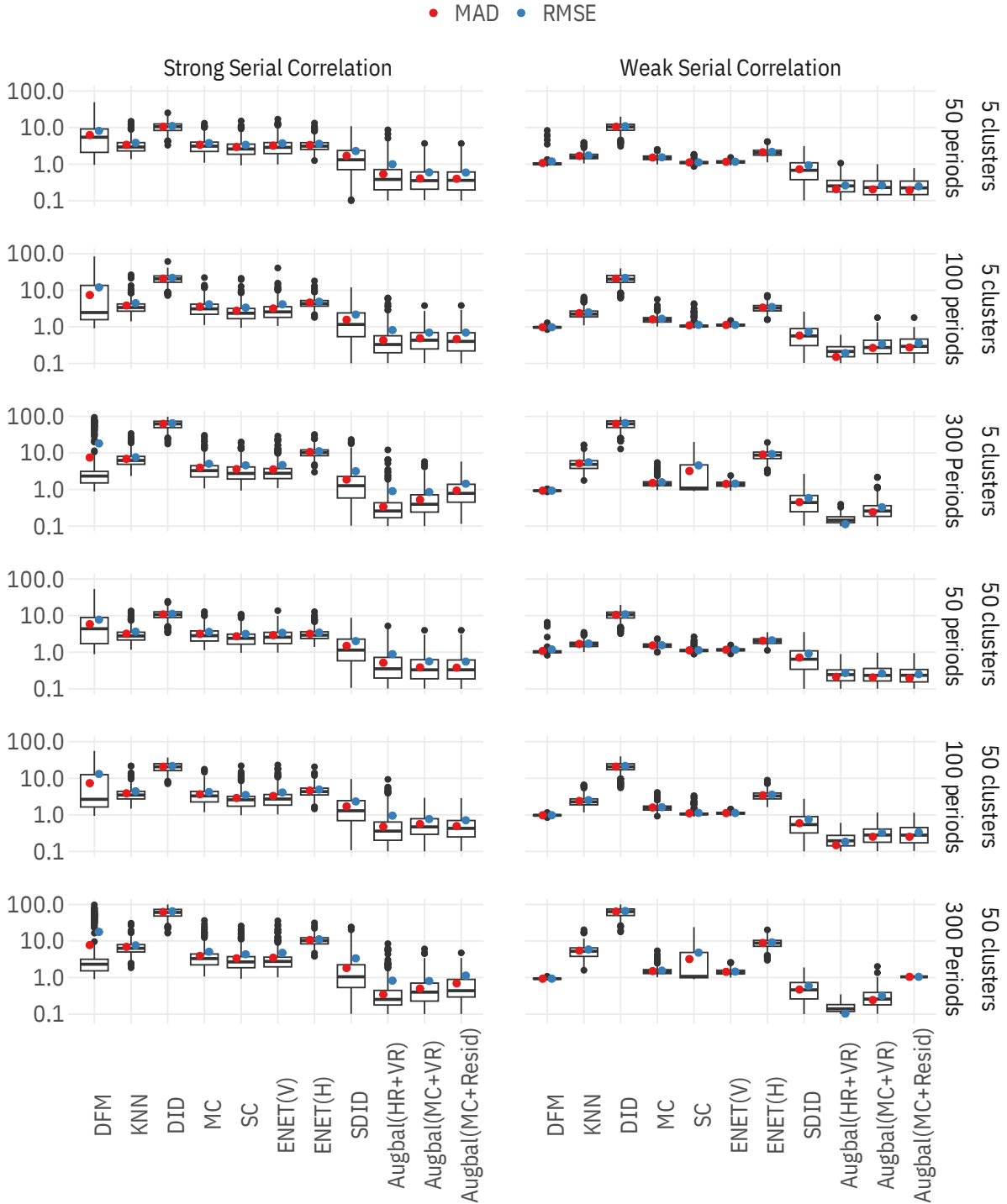
We compare several widely used panel data methods along these metrics. The methods can be classified as belonging to *vertical* methods: K Nearest Neighbours (KNN), Synthetic Control (SC), Vertical Elastic Net (ENET(V)), *horizontal* methods: Horizontal Elastic Net (ENET(H)), and *hybrid* methods: Difference-in-Differences (DID), Dynamic Factor Models (DFM), Synthetic Difference-in-Differences (SDID), Augmented Synthetic Control (Augbal (HR + VR)), and Augmented Balancing with a Matrix Completion outcome model paired with synthetic control weights fit on raw pre-treatment data (Augbal (MC + VR)) and residualized pre-treatment data (Augbal (MC + Resid)).

We report results from our simulations in 4, where we fix  $N = 50$  and 10 units ever treated<sup>8</sup>. We find that across a wide range of settings, hybrid methods such as SDID, and abal outperform pure vertical and horizontal regression methods, and augmented balancing estimators outperform pure outcome modelling methods (including those that can harness both vertical and horizontal patterns, such as DFM, MC, and DID). Among the augmented

<sup>8</sup>We report similar results for 1 and 25 treated units fixing  $N = 50$  in A5, A6, and larger  $N = 100$  units simulations in A7, A8, and A9.

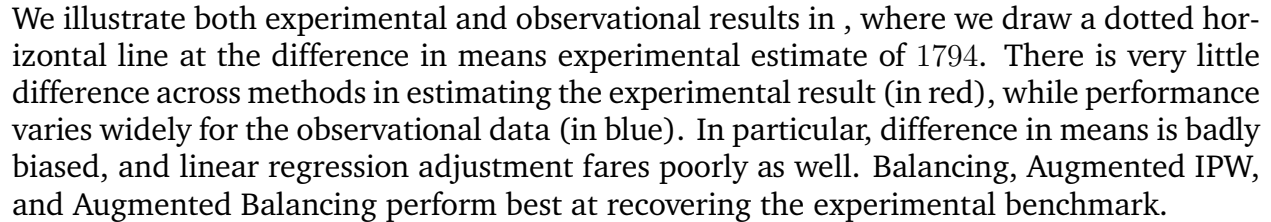
balancing methods, we find that when time series component is low-rank, a matrix completion outcome model performs better than a ridge time series outcome model, while the opposite is true when time series component is high rank, with less separation between them when the data is strongly serially autocorrelated. These simulations suggest that augmented balancing substantially outperforms a wide variety of existing methods.

FIGURE 4. Panel Simulation with Serial Correlation: Absolute bias of estimators over 1000 replications Columns correspond to *strong* and *weak* serial correlation (i.e. large and small values of the ARMA coefficients). The top three rows correspond to a *low-rank* structure in the time trends (5 clusters), while the bottom three correspond with a *high-rank* structure on time trends, with the three rows corresponding to the panel length (50, 100, and 300 time periods respectively). Average absolute bias is indicated by the red dot, while RMSE is indicated by the blue dot.



## 4. Empirical Applications

In this section, we illustrate the use of augmented balancing alongside some well known estimators.

**4.1. Application to cross-sectional data: LaLonde (1986).** We begin with an application to the celebrated LaLonde (1986) and Dehejia and Wahba (1999) job-training program data, where the experimental benchmark is known and observational estimators are frequently evaluated on their performance at uncovering the experimental benchmark when experimental controls are replaced with observational controls. We focus on the Lalonde observational sample where control units are drawn from the Panel Study of Income Dynamics (PSID), which gives us access to 2490 control units to estimate the counterfactual potential outcome under control for the 185 experimental units in the sample. We illustrate both experimental and observational results in , where we draw a dotted horizontal line at the difference in means experimental estimate of 1794. There is very little difference across methods in estimating the experimental result (in red), while performance varies widely for the observational data (in blue). In particular, difference in means is badly biased, and linear regression adjustment fares poorly as well. Balancing, Augmented IPW, and Augmented Balancing perform best at recovering the experimental benchmark.

## Estimating the ATT on Lalonde (1986) JTPA Data

Dashed line denotes difference in means estimate from the experiment

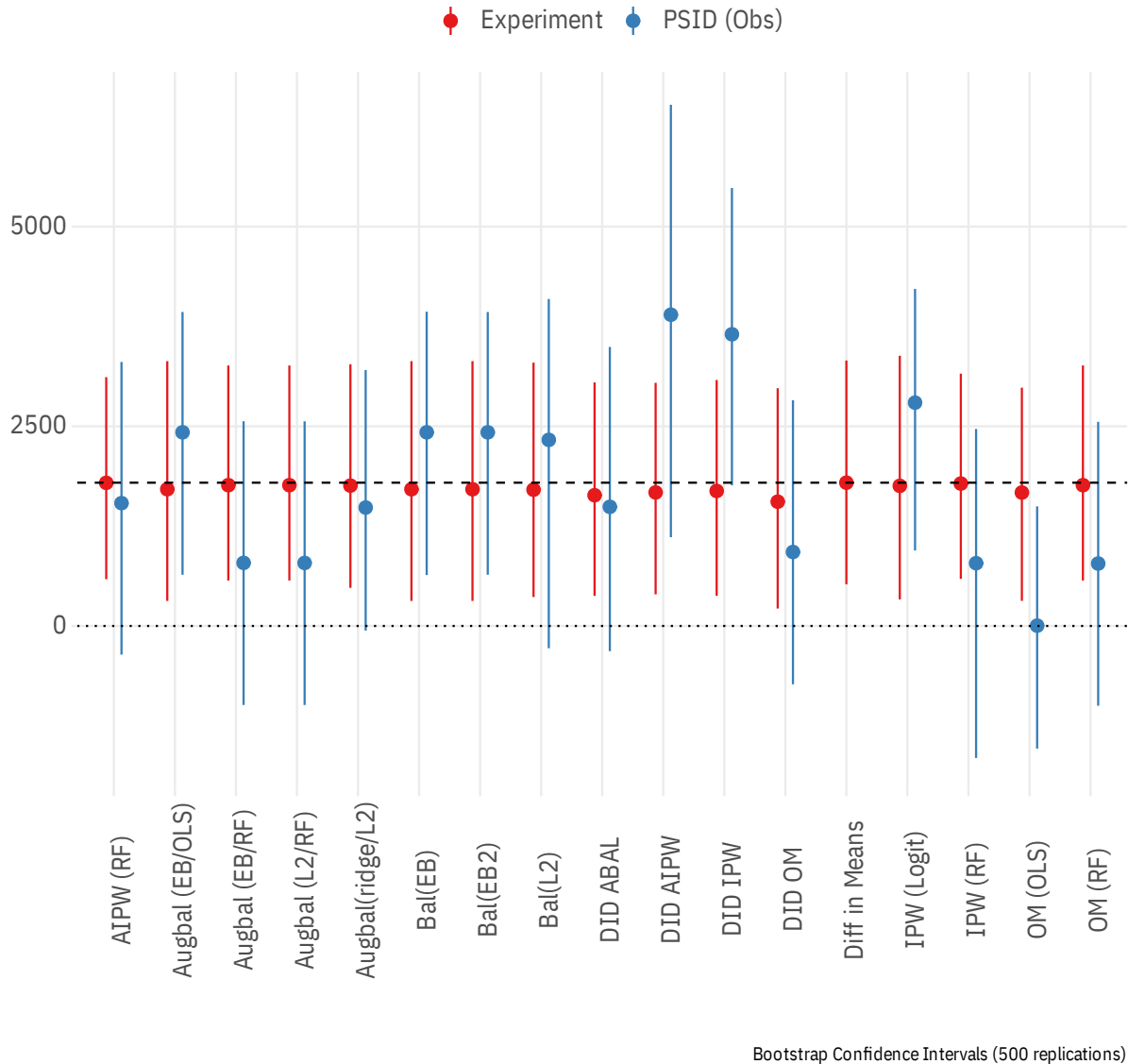


FIGURE 5. Estimates from Lalonde experimental and observational samples. Relative to the experimental benchmark of 1794, Augmented Balancing with Entropy loss and Augmented IPW have lowest bias, with the former having substantially lower variance.

### 4.2. Application to Difference in Differences: Ladd and Lenz (2009).

Next, we apply several existing methods and augmented balancing to the two-period difference in differences study of Ladd and Lenz (2009), who study the effects of newspaper endorsements on voting by leveraging a unique dataset of individual level voter behaviour and newspaper readership paired with the unexpected endorsement of Tony Blair by prominent English news papers. Ladd and Lenz study the effects of the Blair endorsement by the Sun newspaper by comparing the difference in labour voting rates in 1997 and 1992 among

## Difference in Differences Estimates : Ladd-Lenz (2009)

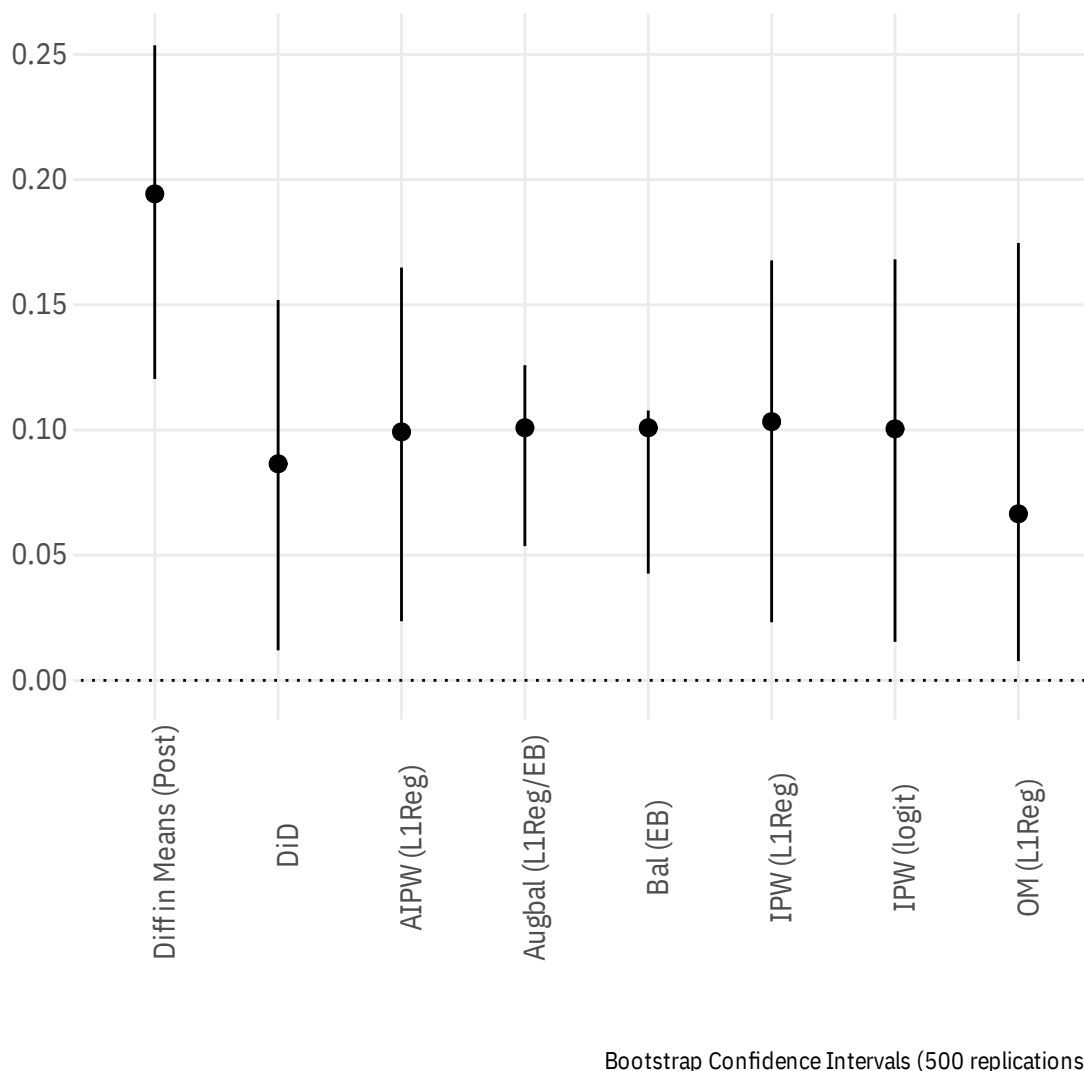


FIGURE 6. Estimates of difference in differences estimates for Ladd and Lenz (2009).

Sun readers with non-readers. The original study reports estimates from difference in differences using linear regression, and lagged dependent variable adjustment. We report estimates in figure 6, and find that the estimators largely agree with the original point estimate of increasing the probability of voting for the labour party by 10 percentage points.

### 4.3. Application to Panel Data: Heersink, Peterson, and Jenkins (2017).

Finally, we apply several existing methods and panel augmented balancing to the panel data analysis in Heersink, Peterson, and Jenkins (2017), who study retrospective voting in the aftermath of a natural disaster by estimating the effect of the great Mississippi flood of 1927 on county level vote shares in the South for the Republican Party in 1928. They use difference in differences as well as synthetic control methods in their paper. We apply several



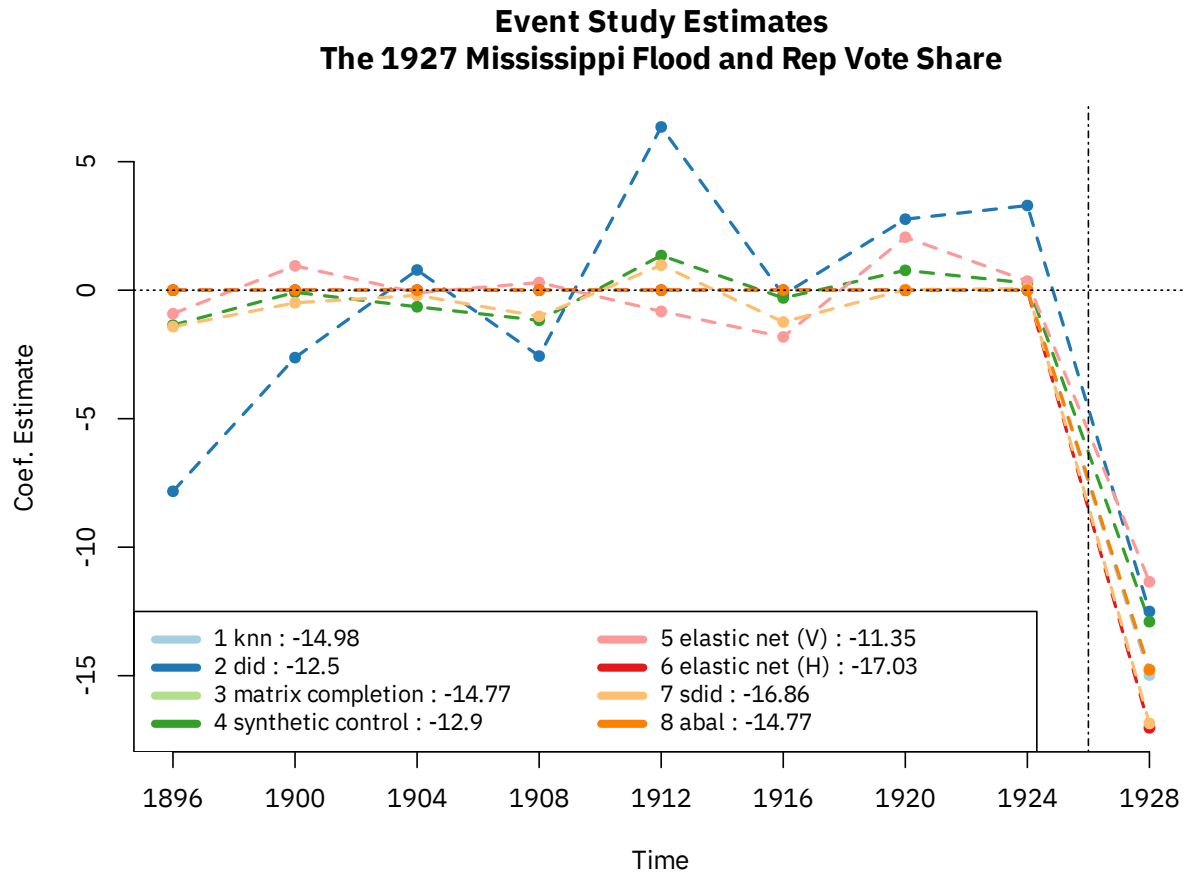


FIGURE 7. MS flood

estimators including the augmented balancing estimator proposed in the previous section to this setting and report the event study results in figure 7. We find that difference in differences may be biased by the presence of erratic pre-trends, while imputation, balancing, and augmented balancing largely agree on the estimate of between 10 to 15% decrease in republican vote shares in flooded counties.

## 5. Conclusion

In this paper, we propose a comprehensive framework that nests several disparate causal inference strategies into a common structure that incorporates flexible machine-learning based outcome modelling augmented with covariate balancing weights. We provide an overview of several specific instantiations of these ideas in the cross-sectional and panel data settings, and extend this approach to the two-period difference in differences. Finally, we perform extensive simulation studies for cross-sectional, difference-in-differences, and panel data settings to benchmark the performance of several state-of-the-art estimators and find that augmented balancing based estimators weakly outperform pure outcome

modelling and inverse-propensity weighting based estimators. We provide performant implementations of these estimators in the `abal` R package.

## References

- Abadie, Alberto (Jan. 2005). “Semiparametric Difference-in-Differences Estimators”. en. *The Review of economic studies* 72.1, pp. 1–19 (cit. on pp. [2](#), [12](#), [23](#)).
- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010). “Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program”. *Journal of the American statistical Association* 105.490, pp. 493–505 (cit. on pp. [2](#), [17](#)).
- Abadie, Alberto and Javier Gardeazabal (2003). “The economic costs of conflict: A case study of the Basque Country”. *American economic review* 93.1, pp. 113–132 (cit. on pp. [2](#), [17](#)).
- Abadie, Alberto and Guido Imbens (2006). “Large sample properties of matching estimators for average treatment effects”. *econometrica* 74.1, pp. 235–267 (cit. on p. [6](#)).
- Abadie, Alberto and Guido W Imbens (2011). “Bias-corrected matching estimators for average treatment effects”. *Journal of Business & Economic Statistics* 29.1, pp. 1–11 (cit. on p. [10](#)).
- Arkhangelsky, Dmitry, Susan Athey, David A Hirshberg, Guido Imbens, and Stefan Wager (2020). “Synthetic Difference in Differences”. *American Economic Review* (cit. on p. [18](#)).
- Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi (2021). “Matrix Completion Methods for Causal Panel Data Models”. *Journal of the American Statistical Association* (cit. on pp. [2](#), [17](#), [19](#)).
- Athey, Susan and Guido W Imbens (Mar. 2006). “Identification and Inference in Nonlinear Difference-in-Differences Models”. *Econometrica: journal of the Econometric Society* 74.2, pp. 431–497 (cit. on p. [14](#)).
- Athey, Susan, Guido W Imbens, and Stefan Wager (Sept. 2018). “Approximate residual balancing: debiased inference of average treatment effects in high dimensions”. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 80.4, pp. 597–623 (cit. on p. [20](#)).
- Bang, Heejung and James M Robins (2005). “Doubly robust estimation in missing data and causal inference models”. *Biometrics* 61.4, pp. 962–973 (cit. on p. [9](#)).
- Ben-Michael, Eli, Avi Feller, David A Hirshberg, and José R Zubizarreta (Oct. 2021). “The Balancing Act in Causal Inference”. arXiv: [2110.14831 \[stat.ME\]](#) (cit. on pp. [6](#), [10](#)).
- Ben-Michael, Eli, Avi Feller, and Jesse Rothstein (2021). “The augmented synthetic control method”. *Journal of the American Statistical Association* 116.536, pp. 1789–1803 (cit. on pp. [2](#), [17](#), [18](#)).
- Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2022). “Revisiting Event Study Designs: Robust and Efficient Estimation”. Available at SSRN [2826228](#) (cit. on p. [17](#)).
- Bruns-Smith, David, Oliver Dukes, Avi Feller, and Elizabeth L Ogburn (2023). “Augmented balancing weights as undersmoothed regressions” (cit. on pp. [5](#), [11](#)).
- Callaway, Brantly and Pedro HC Sant’Anna (2020). “Difference-in-differences with multiple time periods”. *Journal of Econometrics* (cit. on p. [17](#)).

- Chang, Neng-Chieh (Feb. 2020). “Double/debiased machine learning for difference-in-differences models”. en. *The econometrics journal* 23.2, pp. 177–191 (cit. on pp. 13, 23).
- Chattopadhyay, Ambarish and Jose R Zubizarreta (2021). “On the implied weights of linear regression for causal inference”. *arXiv preprint arXiv:2104.06581* (cit. on p. 5).
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (Feb. 2018). “Double/debiased machine learning for treatment and structural parameters”. *The econometrics journal* 21.1 (cit. on pp. 2, 9).
- Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins (2022). “Locally robust semiparametric estimation”. en. *Econometrica: journal of the Econometric Society* 90.4, pp. 1501–1535 (cit. on p. 2).
- Dehejia, Rajeev H and Sadek Wahba (1999). “Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs”. *Journal of the American statistical Association* 94.448, pp. 1053–1062 (cit. on p. 30).
- Doudchenko, Nikolay and Guido W Imbens (2016). *Balancing, regression, difference-in-differences and synthetic control methods: A synthesis*. Tech. rep. National Bureau of Economic Research (cit. on p. 2).
- Frölich, Markus (2007). “Propensity score matching without conditional independence assumption—with an application to the gender wage gap in the United Kingdom”. *The Econometrics Journal* 10.2, pp. 359–407 (cit. on p. 20).
- Gardner, John (2022). “Two-stage differences in differences”. *arXiv preprint arXiv:2207.05943* (cit. on p. 17).
- Goodman-Bacon, Andrew (2018). *Difference-in-differences with variation in treatment timing*. Tech. rep. National Bureau of Economic Research (cit. on p. 17).
- Hahn, Jinyoung (1998). “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects”. *Econometrica* (cit. on pp. 2, 3, 9).
- Hainmueller, Jens (2012). “Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies”. *Political analysis*, pp. 25–46 (cit. on pp. 2, 8, 10, 20, 21).
- Heckman, James J., Hidehiko Ichimura, and Petra Todd (Apr. 1998). “Matching As An Econometric Evaluation Estimator”. *The Review of Economic Studies* 65.2, pp. 261–294. DOI: [10.1111/1467-937X.00044](https://doi.org/10.1111/1467-937X.00044) (cit. on pp. 5, 6, 23).
- Heersink, Boris, Brenton D Peterson, and Jeffery A Jenkins (2017). “Disasters and elections: Estimating the net effect of damage and relief in historical perspective”. *Political Analysis* 25.2, pp. 260–268 (cit. on p. 32).
- Hirshberg, David A and Stefan Wager (2021). “Augmented minimax linear estimation”. *The Annals of Statistics* 49.6, pp. 3206–3227 (cit. on p. 2).
- Imai, Kosuke and Marc Ratkovic (2014). “Covariate balancing propensity score”. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pp. 243–263 (cit. on pp. 6, 7, 10).
- Imbens, Guido and Dmitry Arkhangelsky (2021). “Double-Robust Identification for Causal Panel Data Models”. *NBER* (cit. on p. 15).
- Imbens, Guido W (Feb. 2004). “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review”. *The review of economics and statistics* 86.1, pp. 4–29 (cit. on p. 4).

- Kang, Joseph DY and Joseph L Schafer (2007). “Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data”. *Statistical Science* (cit. on p. 2).
- Kline, Patrick (2011). “Oaxaca-Blinder as a reweighting estimator”. *American Economic Review* 101.3, pp. 532–37 (cit. on p. 5).
- Ladd, Jonathan McDonald and Gabriel S Lenz (2009). “Exploiting a rare communication shift to document the persuasive power of the news media”. *American Journal of Political Science* 53.2, pp. 394–410 (cit. on p. 31).
- LaLonde, Robert J (1986). “Evaluating the econometric evaluations of training programs with experimental data”. *The American economic review*, pp. 604–620 (cit. on p. 30).
- Lechner, Michael (2011). “The Estimation of Causal Effects by Difference-in-Difference Methods”. *Foundations and Trends® in Econometrics* 4.3, pp. 165–224 (cit. on p. 12).
- Liu, Licheng, Ye Wang, and Yiqing Xu (July 2021). “A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data”. arXiv: [2107.00856](https://arxiv.org/abs/2107.00856) [stat.ME] (cit. on p. 17).
- Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao (1994). “Estimation of regression coefficients when some regressors are not always observed”. *Journal of the American statistical Association* 89.427, pp. 846–866 (cit. on pp. 2, 9).
- Robinson, Peter M (1988). “Root-N-consistent semiparametric regression”. *Econometrica: Journal of the Econometric Society*, pp. 931–954 (cit. on p. 2).
- Rosenbaum, Paul R and Donald B Rubin (1983). “The central role of the propensity score in observational studies for causal effects”. *Biometrika* 70.1, pp. 41–55 (cit. on pp. 5–7).
- Sant’Anna, Pedro H C and Jun Zhao (Nov. 2020). “Doubly robust difference-in-differences estimators”. *Journal of econometrics* 219.1, pp. 101–122 (cit. on pp. 2, 13).
- Smith, Jeffrey A and Petra E Todd (2005). “Does matching overcome LaLonde’s critique of nonexperimental estimators?” *Journal of econometrics* 125.1-2, pp. 305–353 (cit. on p. 6).
- Wang, Yixin and Jose R Zubizarreta (Oct. 2019). “Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations”. en. *Biometrika* (cit. on p. 8).
- Xu, Yiqing (Jan. 2017). “Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models”. *Political analysis: an annual publication of the Methodology Section of the American Political Science Association* 25.1, pp. 57–76 (cit. on p. 17).
- Zhao, Qingyuan and Daniel Percival (2016). “Entropy balancing is doubly robust”. *Journal of Causal Inference* 5.1 (cit. on pp. 2, 10).
- Zubizarreta, Jose R (July 2015). “Stable weights that balance covariates for estimation with incomplete outcome data”. en. *Journal of the American Statistical Association* 110.511, pp. 910–922 (cit. on pp. 2, 8, 21).

## Appendix A. Additional Simulation Studies

**A.1. cross-sectional: ACIC Studies for ebal.** In this section, we apply the set of estimators in 2.1 to the 2016 ACIC data analysis competition (Dorie et al., 2019) [appdx A.1] synthetic datasets, which provide a wide variety of realistic data settings that researchers face in applied work. The simulations all generate outcome and treatment assignment using a dataset with 4802 observations and 58 covariates. The 77 simulation settings vary the following six parameters over 100 replications.

- **Treatment model** ( $\text{trtM}$ )  $\in \{ \text{Linear, polynomial, step} \}$ , which varies the treatment assignment model to either be linear in covariates, incorporate polynomial terms, or step functions.
- **Response model** ( $\text{yM}$ )  $\in \{ \text{Linear, exponential, step} \}$  which varies the outcome model to either be linear in covariates, incorporate polynomial terms passed through  $\exp$ , or step functions.
- **Treatment/Response Alignment** ( $\text{algn}$ )  $\in \{ \text{None, Low, High} \}$  governs the probability with which covariates entering the treatment model also enter the outcome model, where *low* gives 25% probability and *high* gives 75% probability
- **Heterogeneity** ( $\text{het}$ )  $\in \{ \text{None, Low, High} \}$  governs with treatment effect heterogeneity, where *None* corresponds with a constant effect, *low* corresponds with heterogeneity along 3 covariates, and *high* corresponds with heterogeneity along 6 covariates.
- **Overlap** ( $\text{ovr}$ )  $\in \{ \text{Full, Penalty} \}$  *Full* indicates moderate coefficients in treatment assignment model  $\text{logit}(\Pr(W = 1 | X))$ , while *penalty* adds large negative values of randomly chosen covariates such that some combinations have deterministic assignment
- **Treated %** ( $\text{trtP}$ )  $\in [\text{low} = 35\%, \text{high} = 65\%]$  is the share of treated units in the population

For each simulation, we compute the ATT using

- **ols**: Linear regression
- **ipw**: Inverse Propensity Weighting using logistic regression that incorporates all covariates
- **aipwRF**: Augmented IPW that uses Honest Random Forests to estimate both nuisance functions
- **eba101**: Entropy balancing on the first moment of all covariates using the NR-implementation of Entropy Balancing
- **eba1New**: Entropy balancing on first and second moments of all covariates using an autodiff-based implementation of Entropy Balancing. This problem is infeasible with the old implementation.
- **hba1**: Hierarchically regularized entropy balancing as proposed by Xu and Yang (2022) on first and second moments
- **ba1HD**: Augmented Balancing pairing regularized (elastic net) regression with L2 balancing weights as proposed by Athey, Imbens, and Wager (2018)

- `augbalE`: Augmented balancing pairing random forest with Entropy balancing on first and second moments of all covariates

We report the Bias and RMSE across 100 simulations under each of the 77 combinations in tables A1 and A2 (with best performers bolded in each row). We find that `aipwRF`, `augbalE`, and `ebalNew` perform best in terms of Bias and RMSE. `aipwRF` performs well across a variety of settings because the random forest adapts to different treatment and outcome model functional forms. When one of the outcome and treatment models are smooth (linear or polynomial), the augmented balancing estimator also performs well and is typically on par with the AIPW estimator.

TABLE A1. Bias across ACIC simulation settings. Best performers within order of  $10^{-2}$  bolded.

trtM	yM	algn	het	ovr	trtP	ols	ipw	aipwRF	ebalOld	ebalNew	hbal	balHD	augbalE
lin	exp	high	high	pen	low	0.787	0.998	<b>0.184</b>	0.578	0.315	2.46	0.437	0.438
lin	exp	low	high	pen	low	0.727	1.114	<b>0.117</b>	0.59	0.274	2.563	0.447	0.348
lin	exp	high	high	pen	high	0.877	0.74	<b>0.157</b>	0.319	0.198	4.058	0.246	<b>0.157</b>
lin	exp	low	high	pen	high	0.827	1.108	<b>0.109</b>	0.428	0.238	4.78	0.303	0.162
lin	lin	high	high	pen	low	0.249	0.221	0.147	0.176	<b>0.123</b>	2.214	0.149	0.417
lin	lin	high	none	pen	low	0.174	0.242	0.176	0.179	<b>0.121</b>	2.187	0.152	0.474
pol	exp	high	high	full	low	0.833	0.715	0.29	0.705	0.26	3.66	0.704	<b>0.187</b>
pol	exp	high	high	pen	low	0.828	0.73	0.267	0.711	<b>0.248</b>	2.433	0.686	0.372
pol	exp	none	high	pen	low	0.169	0.173	<b>0.075</b>	0.139	0.096	1.357	0.132	0.165
pol	exp	low	high	pen	low	0.719	0.798	<b>0.187</b>	0.631	0.247	2.335	0.551	0.314
pol	exp	high	high	pen	high	0.893	0.764	<b>0.216</b>	0.532	0.241	4.099	0.504	<b>0.207</b>
pol	exp	none	high	pen	high	0.215	0.575	0.076	0.113	0.077	4.614	0.098	<b>0.056</b>
pol	exp	low	high	pen	high	0.796	0.89	<b>0.158</b>	0.479	0.246	4.534	0.389	<b>0.168</b>
pol	exp	low	high	full	low	0.779	0.641	0.249	0.625	0.224	3.117	0.631	<b>0.176</b>
pol	exp	low	high	full	high	0.824	0.55	0.205	0.605	<b>0.178</b>	4.985	0.543	<b>0.169</b>
pol	exp	high	high	full	high	0.868	0.619	0.244	0.514	<b>0.177</b>	4.807	0.601	0.216
pol	exp	low	low	pen	low	0.235	0.254	<b>0.095</b>	0.227	0.115	1.966	0.198	0.199
pol	exp	high	low	pen	low	0.399	0.501	0.177	0.374	<b>0.157</b>	2.167	0.372	0.271
pol	exp	low	low	full	low	0.252	0.232	<b>0.108</b>	0.226	<b>0.103</b>	2.18	0.235	<b>0.106</b>
pol	exp	high	low	full	low	0.516	0.468	<b>0.186</b>	0.454	<b>0.176</b>	3.474	0.467	<b>0.183</b>
pol	exp	low	low	pen	high	0.36	0.571	<b>0.1</b>	0.304	0.159	5.368	0.221	0.121
pol	exp	high	low	pen	high	0.466	0.404	0.173	0.317	0.194	4.736	0.372	<b>0.147</b>
pol	exp	low	low	full	high	0.308	0.243	<b>0.106</b>	0.186	<b>0.109</b>	5.316	0.233	<b>0.102</b>
pol	exp	high	low	full	high	0.507	0.407	<b>0.181</b>	0.339	<b>0.177</b>	4.975	0.399	<b>0.182</b>
pol	exp	high	none	pen	low	0.283	0.723	<b>0.159</b>	0.328	0.196	1.526	0.291	0.314
pol	lin	high	high	pen	low	0.881	0.875	<b>0.294</b>	0.81	0.313	2.538	0.821	0.406
pol	lin	low	high	pen	low	0.835	0.902	<b>0.208</b>	0.754	0.298	2.382	0.71	0.432
pol	lin	high	high	pen	high	0.872	2.169	0.24	0.694	0.204	4.068	0.617	<b>0.183</b>
pol	lin	low	high	pen	high	0.88	0.922	<b>0.137</b>	0.439	0.222	4.32	0.398	0.181
pol	step	low	high	pen	low	0.826	0.954	<b>0.182</b>	0.72	0.327	2.397	0.652	0.361
pol	step	high	high	pen	low	0.898	0.921	<b>0.277</b>	0.814	0.302	2.346	0.835	0.441
pol	step	low	high	full	low	0.803	0.732	0.216	0.693	0.254	3.259	0.725	<b>0.155</b>
pol	step	high	high	full	low	0.873	0.817	0.258	0.808	0.27	3.897	0.806	<b>0.175</b>
pol	step	low	high	pen	high	0.85	0.894	<b>0.134</b>	0.66	0.309	4.149	0.504	0.2
pol	step	high	high	pen	high	0.834	1.164	0.236	0.775	0.291	4.943	0.662	<b>0.218</b>
pol	step	low	high	full	high	0.784	0.583	<b>0.163</b>	0.598	<b>0.173</b>	5.262	0.56	<b>0.17</b>
pol	step	high	high	full	high	0.91	0.759	<b>0.235</b>	0.689	<b>0.237</b>	4.965	0.742	0.245
pol	step	low	low	pen	low	0.217	1.093	<b>0.087</b>	0.235	0.133	1.823	0.205	0.151
pol	step	high	low	pen	low	0.416	0.439	<b>0.159</b>	0.432	0.191	2.263	0.403	0.241
pol	step	low	low	full	low	0.271	0.275	<b>0.104</b>	0.262	0.138	2.201	0.269	<b>0.104</b>
pol	step	high	low	full	low	0.493	0.472	0.154	0.464	0.158	3.65	0.477	<b>0.136</b>
pol	step	low	low	pen	high	0.266	0.344	<b>0.089</b>	0.242	0.158	4.843	0.205	0.101
pol	step	high	low	pen	high	0.41	0.574	<b>0.163</b>	0.494	0.223	4.662	0.388	<b>0.157</b>
pol	step	low	low	full	high	0.254	0.228	<b>0.093</b>	0.216	0.134	4.993	0.225	<b>0.098</b>

TABLE A1. Bias across ACIC simulation settings. Best performers within order of  $10^{-2}$  bolded. (continued)

trtM	yM	algn	het	ovr	trtP	ols	ipw	aipwRF	eбалOld	eбалNew	hbal	balHD	augbalE
pol	step	high	low	full	high	0.468	0.443	<b>0.163</b>	0.494	0.196	5.105	0.444	0.204
step	exp	low	high	pen	low	0.803	0.901	<b>0.145</b>	0.712	0.364	2.767	0.587	0.335
step	exp	high	high	pen	low	0.878	1.006	<b>0.171</b>	0.826	0.515	2.356	0.761	0.293
step	exp	low	high	full	low	0.928	0.69	<b>0.111</b>	0.682	0.456	3.836	0.71	<b>0.115</b>
step	exp	high	high	full	low	0.91	0.733	0.153	0.739	0.497	3.598	0.743	<b>0.132</b>
step	exp	low	high	pen	high	0.878	0.662	<b>0.121</b>	0.478	0.27	4.778	0.366	0.149
step	exp	high	high	pen	high	0.959	0.858	<b>0.17</b>	0.604	0.366	4.678	0.548	0.186
step	exp	low	high	full	high	0.876	0.507	<b>0.152</b>	0.599	0.375	4.838	0.52	0.177
step	exp	high	high	full	high	0.978	0.659	<b>0.171</b>	0.697	0.414	4.855	0.675	0.199
step	exp	low	low	pen	low	0.275	0.244	<b>0.107</b>	0.202	0.157	2.02	0.22	0.139
step	exp	high	low	pen	low	0.42	0.337	<b>0.103</b>	0.345	0.237	2.29	0.328	0.151
step	exp	low	low	full	low	0.218	0.172	<b>0.074</b>	0.17	0.096	1.944	0.17	0.086
step	exp	high	low	full	low	0.616	0.538	0.141	0.479	0.354	3.341	0.541	<b>0.129</b>
step	exp	low	low	pen	high	0.403	0.478	<b>0.111</b>	0.286	0.188	4.502	0.266	<b>0.106</b>
step	exp	high	low	pen	high	0.518	0.618	<b>0.134</b>	0.404	0.281	4.589	0.366	0.148
step	exp	low	low	full	high	0.337	0.233	<b>0.09</b>	0.197	0.168	5.339	0.233	<b>0.098</b>
step	exp	high	low	full	high	0.586	0.414	<b>0.161</b>	0.31	0.254	4.928	0.413	0.192
step	step	high	high	pen	low	0.846	0.787	<b>0.158</b>	0.803	0.532	2.189	0.771	0.325
step	step	high	high	pen	high	0.934	1.403	<b>0.182</b>	0.665	0.507	4.28	0.644	0.259
step	step	low	high	pen	low	0.794	0.722	<b>0.118</b>	0.685	0.492	2.21	0.664	0.264
step	step	low	high	full	low	0.777	0.685	<b>0.131</b>	0.689	0.467	3.19	0.697	0.152
step	step	high	high	full	low	0.962	0.846	<b>0.167</b>	0.859	0.595	4.508	0.861	<b>0.177</b>
step	step	low	high	pen	high	0.805	0.876	<b>0.086</b>	0.519	0.415	4.333	0.487	0.195
step	step	low	high	full	high	0.829	0.49	<b>0.121</b>	0.576	0.349	4.76	0.492	0.155
step	step	high	high	full	high	0.923	0.606	<b>0.128</b>	0.713	0.39	4.909	0.64	0.183
step	step	low	low	pen	low	0.283	0.285	<b>0.078</b>	0.271	0.205	1.919	0.259	0.17
step	step	high	low	pen	low	0.461	0.445	<b>0.117</b>	0.398	0.298	2.631	0.421	0.17
step	step	low	low	full	low	0.25	0.226	<b>0.054</b>	0.225	0.162	2.247	0.234	0.068
step	step	high	low	full	low	0.483	0.455	<b>0.117</b>	0.472	0.323	3.371	0.456	<b>0.121</b>
step	step	low	low	pen	high	0.324	0.292	<b>0.073</b>	0.25	0.204	5.156	0.237	0.108
step	step	high	low	pen	high	0.46	0.467	<b>0.12</b>	0.442	0.271	4.707	0.365	0.139
step	step	low	low	full	high	0.193	0.175	<b>0.067</b>	0.113	0.126	4.756	0.17	<b>0.075</b>
step	step	high	low	full	high	0.483	0.474	<b>0.105</b>	0.552	0.346	4.723	0.455	0.202

TABLE A2. rmse across ACIC simulation settings. Best performers within order of  $10^{-2}$  bolded.

trtM	yM	algn	het	ovr	trtP	ols	ipw	aipwRF	eбалOld	eбалNew	hbal	balHD	augbalE
lin	exp	high	high	pen	low	0.828	2.232	<b>0.232</b>	0.863	0.502	3.274	0.524	0.624
lin	exp	low	high	pen	low	0.754	2.693	<b>0.151</b>	0.745	0.41	3.389	0.531	0.472
lin	exp	high	high	pen	high	0.935	1.897	<b>0.205</b>	0.458	0.277	4.565	0.307	<b>0.21</b>
lin	exp	low	high	pen	high	0.871	3.204	<b>0.142</b>	0.665	0.307	5.135	0.382	0.208
lin	lin	high	high	pen	low	0.329	0.316	<b>0.188</b>	0.253	<b>0.188</b>	2.705	0.211	0.536
lin	lin	high	none	pen	low	0.272	0.604	<b>0.222</b>	0.291	<b>0.231</b>	2.608	0.245	0.684
pol	exp	high	high	full	low	0.912	0.793	0.344	0.77	0.343	4.225	0.778	<b>0.245</b>
pol	exp	high	high	pen	low	0.896	0.859	0.328	0.829	<b>0.315</b>	2.942	0.767	0.536
pol	exp	none	high	pen	low	0.262	0.274	<b>0.099</b>	0.216	0.145	1.723	0.205	0.265
pol	exp	low	high	pen	low	0.766	1.288	<b>0.231</b>	0.757	0.331	3.184	0.628	0.453
pol	exp	high	high	pen	high	0.987	1.048	<b>0.262</b>	0.641	0.302	4.321	0.6	<b>0.262</b>
pol	exp	none	high	pen	high	0.297	3.302	0.092	0.163	0.104	4.989	0.136	<b>0.073</b>
pol	exp	low	high	pen	high	0.838	1.731	<b>0.21</b>	0.591	0.333	4.822	0.463	<b>0.212</b>
pol	exp	low	high	full	low	0.851	0.712	0.303	0.705	0.306	3.618	0.704	<b>0.245</b>
pol	exp	low	high	full	high	0.875	0.658	0.247	0.735	0.248	5.151	0.642	<b>0.209</b>
pol	exp	high	high	full	high	0.929	0.692	0.297	0.577	<b>0.255</b>	5.108	0.68	<b>0.259</b>
pol	exp	low	low	pen	low	0.299	0.372	<b>0.128</b>	0.302	0.163	2.411	0.262	0.298
pol	exp	high	low	pen	low	0.495	1.06	0.222	0.48	<b>0.211</b>	2.683	0.461	0.467

TABLE A2. rmse across ACIC simulation settings. Best performers within order of  $10^{-2}$  bolded. (continued)

trtM	yM	algn	het	ovr	trtP	ols	ipw	aipwRF	eбалOld	eбалNew	hбал	балHD	augбалE
pol	exp	low	low	full	low	0.387	0.365	<b>0.156</b>	0.364	0.169	2.704	0.371	<b>0.16</b>
pol	exp	high	low	full	low	0.68	0.648	<b>0.239</b>	0.628	0.262	3.968	0.646	0.267
pol	exp	low	low	pen	high	0.529	1.488	<b>0.128</b>	0.417	0.214	5.59	0.289	0.173
pol	exp	high	low	pen	high	0.627	0.526	0.226	0.388	0.287	5.071	0.506	<b>0.202</b>
pol	exp	low	low	full	high	0.425	0.388	0.147	0.3	0.16	5.45	0.351	<b>0.13</b>
pol	exp	high	low	full	high	0.634	0.524	<b>0.234</b>	0.414	0.241	5.199	0.52	<b>0.226</b>
pol	exp	high	none	pen	low	0.374	2.923	<b>0.2</b>	0.443	0.263	1.828	0.389	0.426
pol	lin	high	high	pen	low	0.943	1.054	<b>0.332</b>	0.898	0.437	3.094	0.904	0.703
pol	lin	low	high	pen	low	0.886	1.33	<b>0.264</b>	0.852	0.419	3.027	0.788	0.57
pol	lin	high	high	pen	high	0.921	14.358	0.272	0.853	0.28	4.461	0.706	<b>0.232</b>
pol	lin	low	high	pen	high	0.946	2.011	<b>0.18</b>	0.566	0.319	4.728	0.517	0.268
pol	step	low	high	pen	low	0.894	2.061	<b>0.24</b>	0.833	0.439	2.906	0.727	0.607
pol	step	high	high	pen	low	0.93	1.409	<b>0.321</b>	0.859	0.407	2.929	0.879	0.605
pol	step	low	high	full	low	0.839	0.79	0.251	0.739	0.357	3.917	0.786	<b>0.205</b>
pol	step	high	high	full	low	0.926	0.878	0.299	0.874	0.367	4.355	0.87	<b>0.241</b>
pol	step	low	high	pen	high	0.907	1.383	<b>0.177</b>	0.794	0.405	4.389	0.58	0.31
pol	step	high	high	pen	high	0.87	2.269	<b>0.28</b>	0.877	0.372	5.117	0.733	0.314
pol	step	low	high	full	high	0.822	0.677	<b>0.195</b>	0.707	0.251	5.755	0.635	<b>0.202</b>
pol	step	high	high	full	high	0.971	0.842	<b>0.279</b>	0.762	0.337	5.356	0.82	0.292
pol	step	low	low	pen	low	0.287	8.838	<b>0.112</b>	0.344	0.173	2.503	0.28	0.252
pol	step	high	low	pen	low	0.52	0.553	<b>0.193</b>	0.534	0.257	2.817	0.509	0.392
pol	step	low	low	full	low	0.426	0.432	<b>0.147</b>	0.415	0.211	2.595	0.43	0.159
pol	step	high	low	full	low	0.622	0.608	0.199	0.597	0.253	4.228	0.616	<b>0.185</b>
pol	step	low	low	pen	high	0.382	0.476	<b>0.115</b>	0.323	0.204	5.484	0.286	0.129
pol	step	high	low	pen	high	0.534	1.524	<b>0.207</b>	0.861	0.283	4.994	0.524	<b>0.197</b>
pol	step	low	low	full	high	0.347	0.334	<b>0.131</b>	0.302	0.22	5.34	0.33	0.157
pol	step	high	low	full	high	0.575	0.543	<b>0.197</b>	0.573	0.307	5.667	0.542	0.249
step	exp	low	high	pen	low	0.929	1.45	<b>0.226</b>	0.883	0.472	3.681	0.677	0.94
step	exp	high	high	pen	low	0.956	2.224	<b>0.242</b>	0.912	0.602	2.901	0.841	0.515
step	exp	low	high	full	low	0.993	0.797	<b>0.163</b>	0.751	0.555	4.58	0.808	0.199
step	exp	high	high	full	low	0.994	0.843	0.23	0.846	0.61	4.209	0.855	<b>0.213</b>
step	exp	low	high	pen	high	0.933	0.942	<b>0.166</b>	0.558	0.365	5.001	0.449	0.229
step	exp	high	high	pen	high	1.037	1.195	<b>0.215</b>	0.728	0.456	5.06	0.654	0.251
step	exp	low	high	full	high	0.947	0.623	<b>0.214</b>	0.758	0.496	5.071	0.636	0.256
step	exp	high	high	full	high	1.064	0.777	<b>0.229</b>	0.811	0.496	5.295	0.792	0.263
step	exp	low	low	pen	low	0.367	0.348	<b>0.14</b>	0.297	0.232	2.437	0.322	0.203
step	exp	high	low	pen	low	0.556	0.487	<b>0.14</b>	0.491	0.332	2.787	0.461	0.202
step	exp	low	low	full	low	0.308	0.248	<b>0.095</b>	0.24	0.149	2.485	0.249	0.114
step	exp	high	low	full	low	0.805	0.731	<b>0.236</b>	0.642	0.497	3.949	0.743	<b>0.24</b>
step	exp	low	low	pen	high	0.51	0.96	<b>0.157</b>	0.401	0.275	4.852	0.376	<b>0.149</b>
step	exp	high	low	pen	high	0.711	1.332	<b>0.183</b>	0.528	0.367	4.999	0.477	0.241
step	exp	low	low	full	high	0.468	0.366	<b>0.123</b>	0.313	0.262	5.529	0.369	0.142
step	exp	high	low	full	high	0.734	0.531	<b>0.253</b>	0.441	0.345	5.273	0.523	0.266
step	step	high	high	pen	low	0.894	0.9	<b>0.217</b>	0.885	0.634	2.641	0.856	0.516
step	step	high	high	pen	high	0.993	4.681	<b>0.247</b>	0.785	0.601	4.708	0.744	0.353
step	step	low	high	pen	low	0.828	0.814	<b>0.158</b>	0.73	0.562	2.673	0.713	0.367
step	step	low	high	full	low	0.809	0.759	<b>0.187</b>	0.763	0.555	3.731	0.778	0.253
step	step	high	high	full	low	1.012	0.926	<b>0.223</b>	0.935	0.677	5.012	0.94	0.286
step	step	low	high	pen	high	0.853	1.549	<b>0.124</b>	0.627	0.51	4.679	0.567	0.275
step	step	low	high	full	high	0.877	0.608	<b>0.181</b>	0.747	0.462	5.212	0.612	0.215
step	step	high	high	full	high	0.978	0.711	<b>0.185</b>	0.798	0.478	5.202	0.734	0.242
step	step	low	low	pen	low	0.416	0.463	<b>0.105</b>	0.407	0.301	2.31	0.387	0.283
step	step	high	low	pen	low	0.566	0.547	<b>0.158</b>	0.494	0.39	3.163	0.521	0.23
step	step	low	low	full	low	0.367	0.355	<b>0.075</b>	0.372	0.253	2.766	0.368	0.101
step	step	high	low	full	low	0.636	0.623	<b>0.187</b>	0.664	0.469	4.004	0.634	0.204
step	step	low	low	pen	high	0.444	0.391	<b>0.104</b>	0.314	0.264	5.426	0.323	0.147
step	step	high	low	pen	high	0.596	0.648	<b>0.164</b>	0.532	0.378	5.137	0.47	0.19
step	step	low	low	full	high	0.284	0.26	<b>0.096</b>	0.157	0.184	4.952	0.25	0.114



**Simulation Study**

$n = 300$ ; True effect is 0

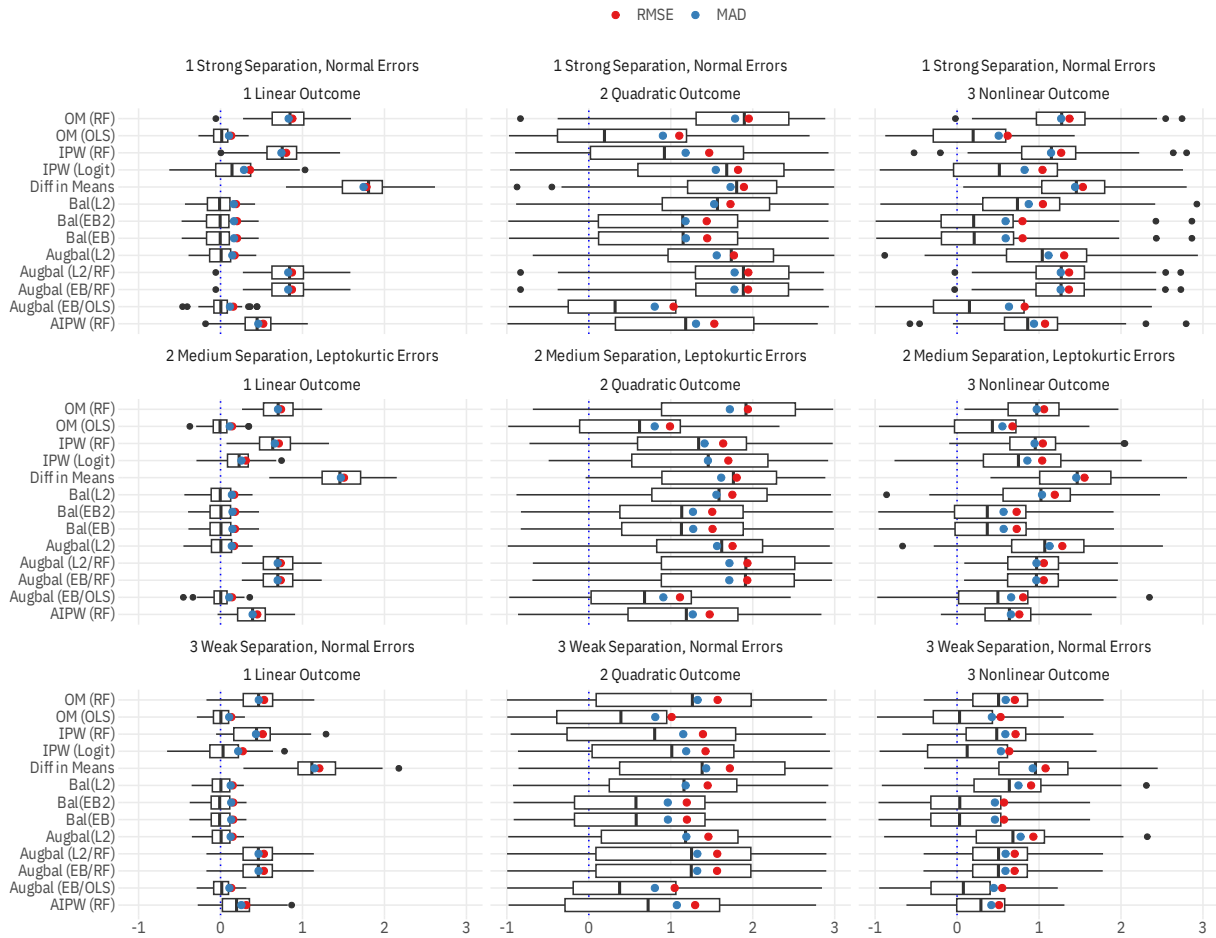


FIGURE A1. Cross-sectional simulation study with  $n = 300$

step step high low full high 0.636 0.613 0.147 0.688 0.454 5.382 0.591 0.294

**A.2. cross-sectional: Hainmueller (2012) simulation.**

**Simulation Study**

$n = 1500$ ; True effect is 0

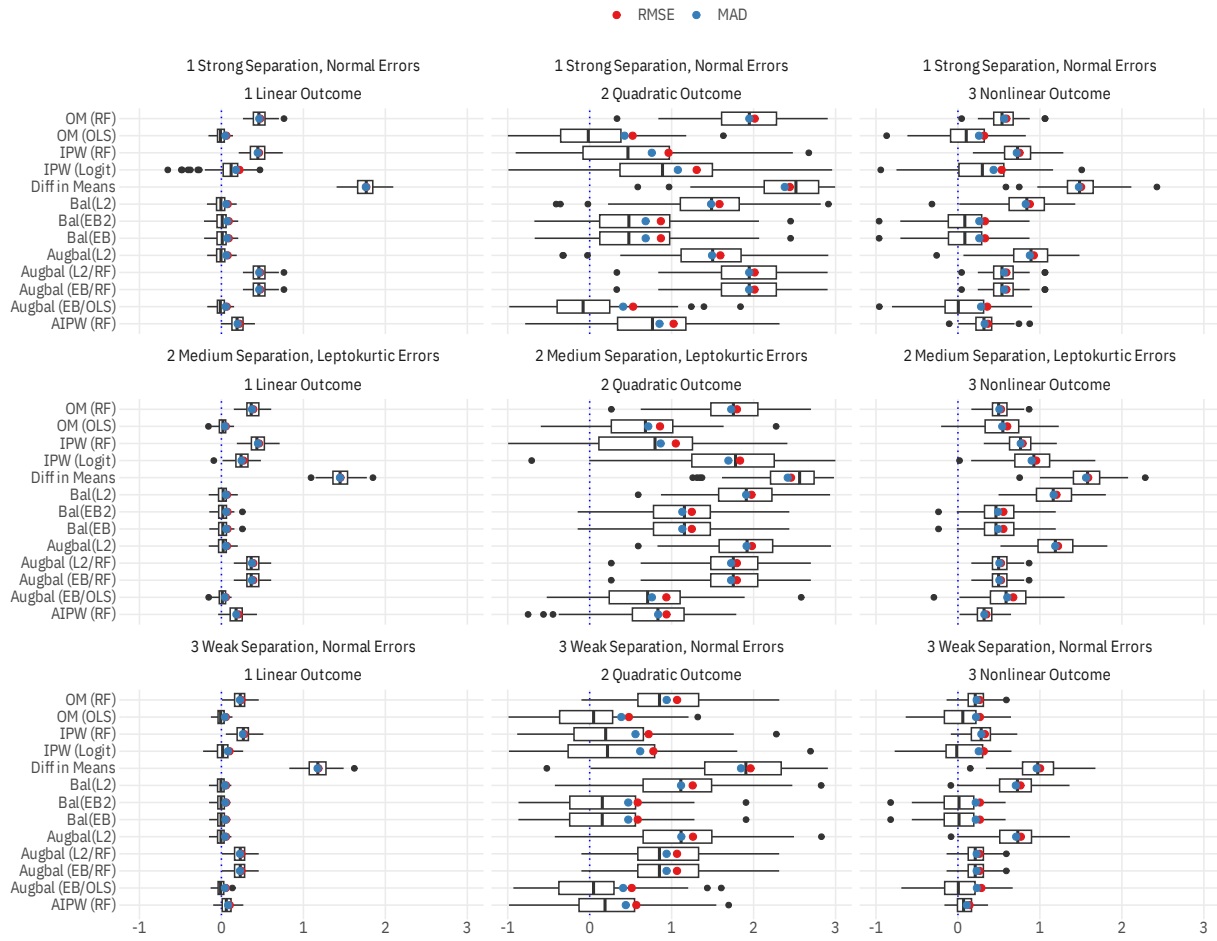


FIGURE A2. Cross-sectional simulation study with  $n = 1500$

**Simulation Study**

$n = 5000$ ; True effect is 0

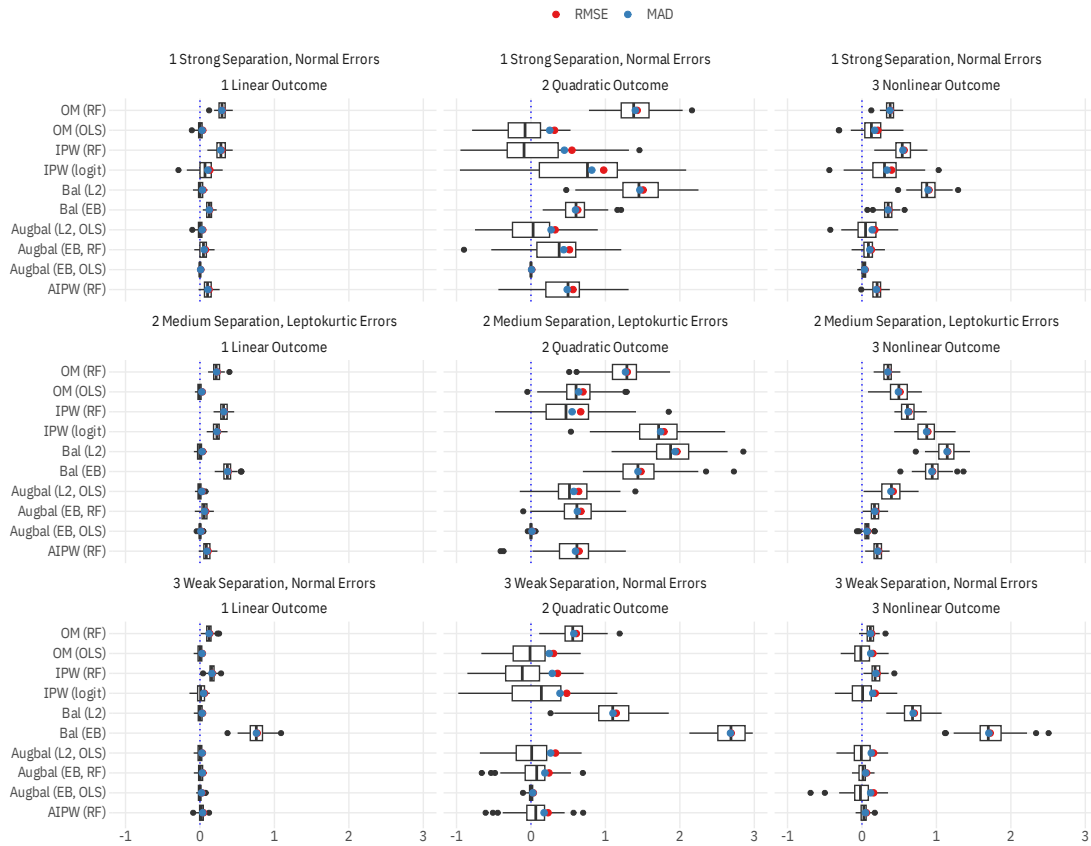
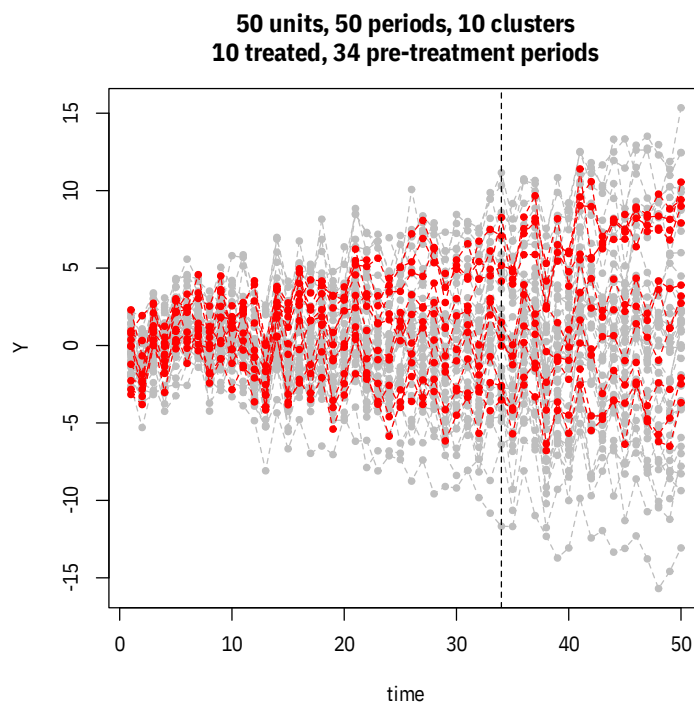
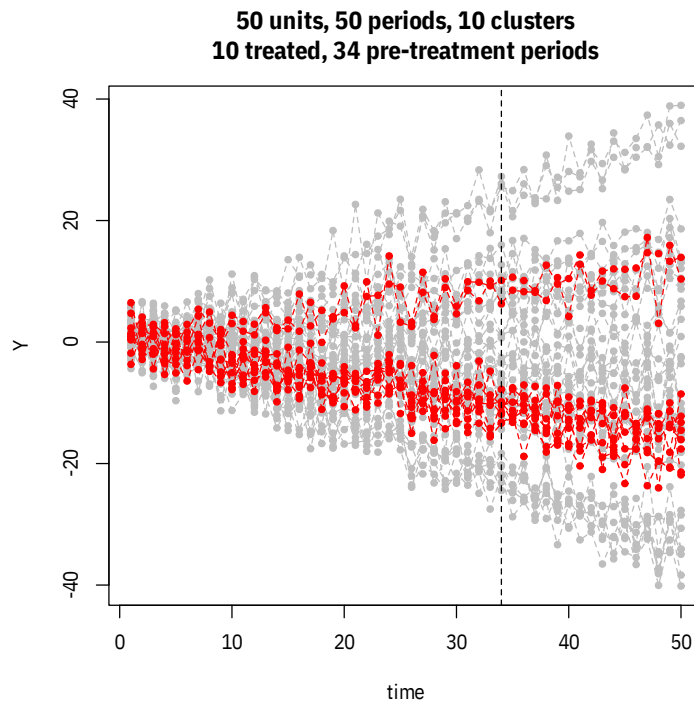


FIGURE A3. Cross-sectional simulation study with  $n = 5000$

### **A.3. Additional Panel Simulation Results.**

#### **A.3.1. DGPs visualised.**

FIGURE A4. Simulation examples under strong and weak autocorrelation



A.3.2.  $N = 50$  simulations.

FIGURE A5. 1 Treated Unit Simulation Results

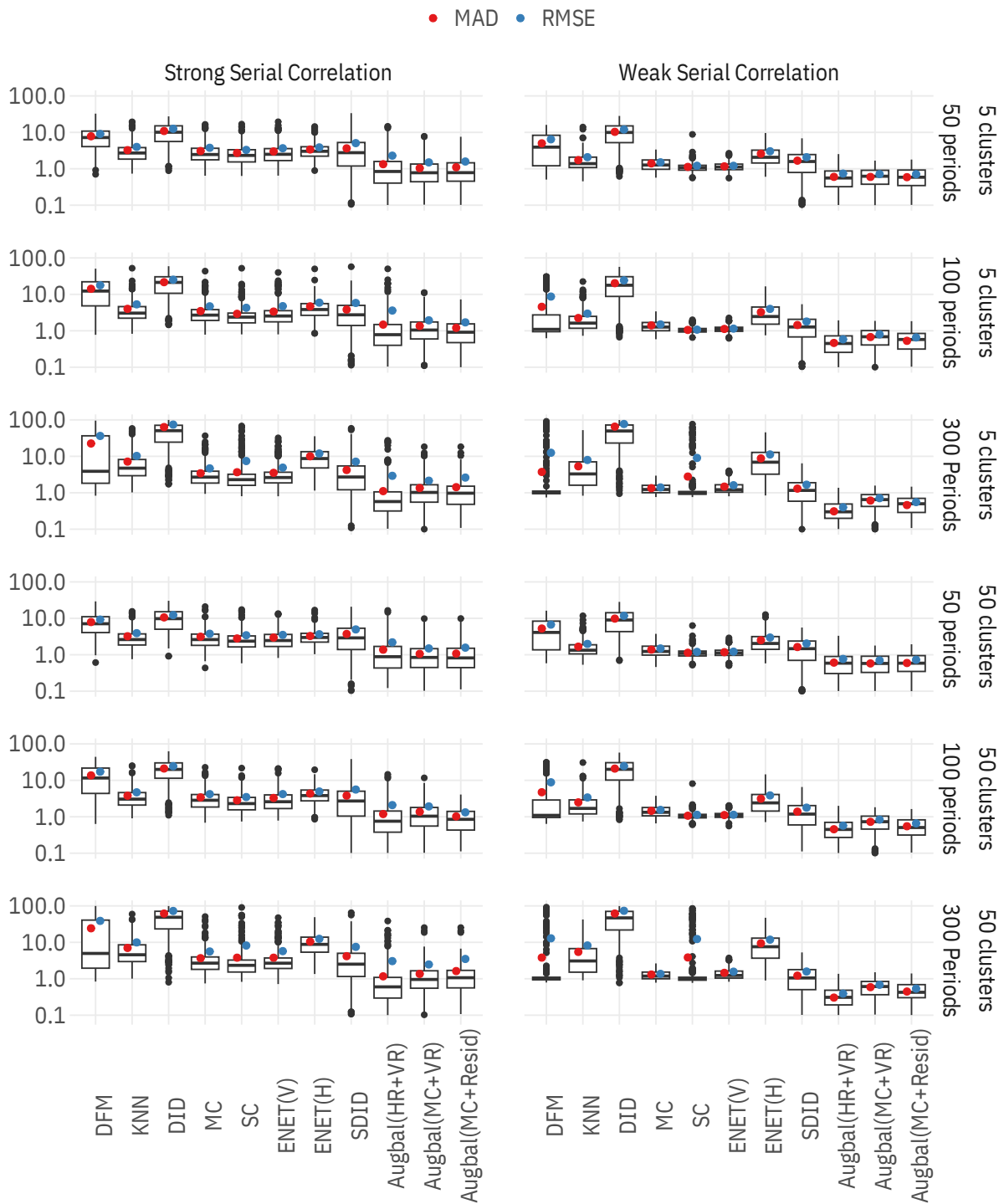
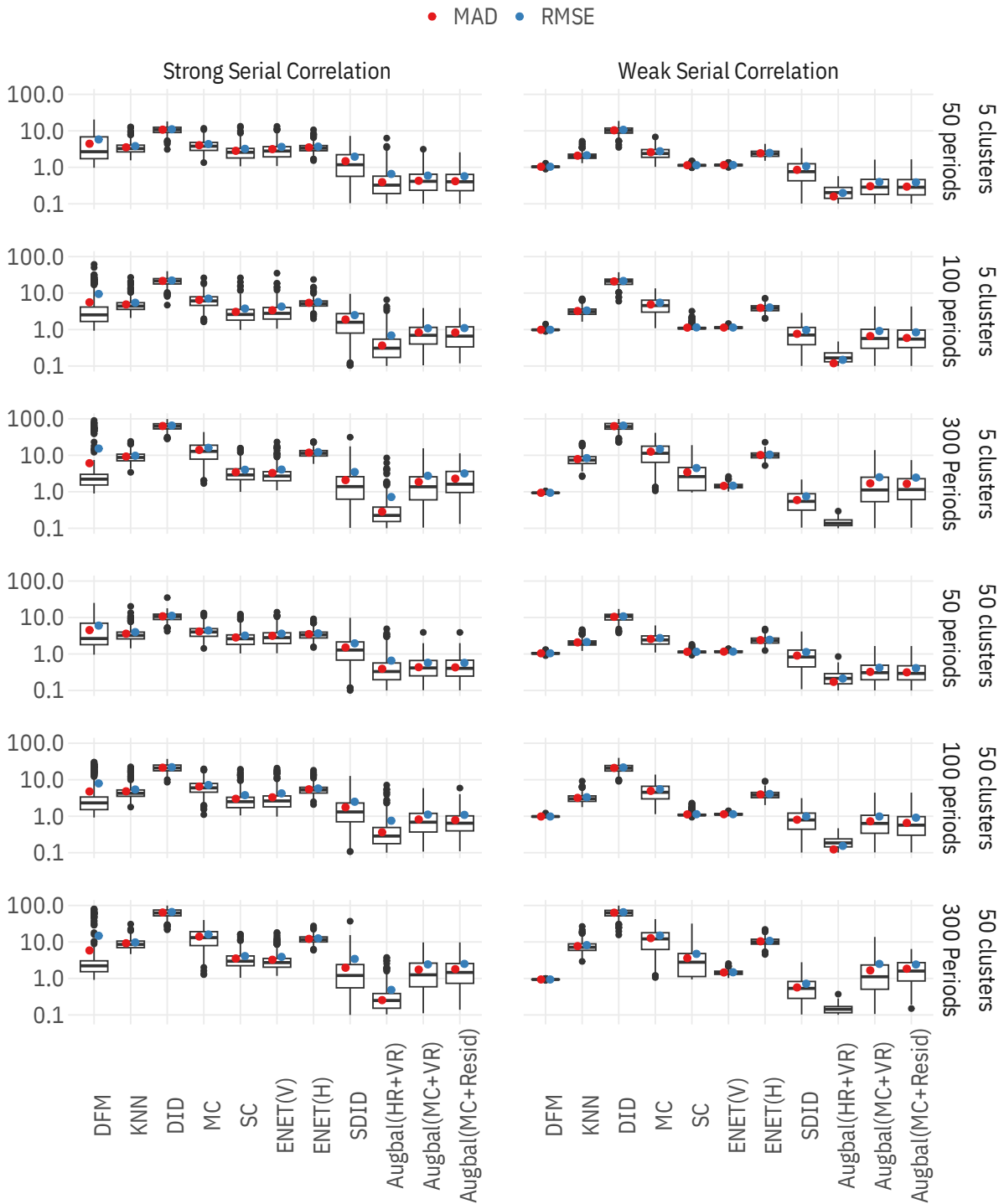


FIGURE A6. 25 Treated Units Simulation Results



A.3.3.  $N = 100$  simulations.

FIGURE A7. 100 units: 1 Treated Unit Simulation Results

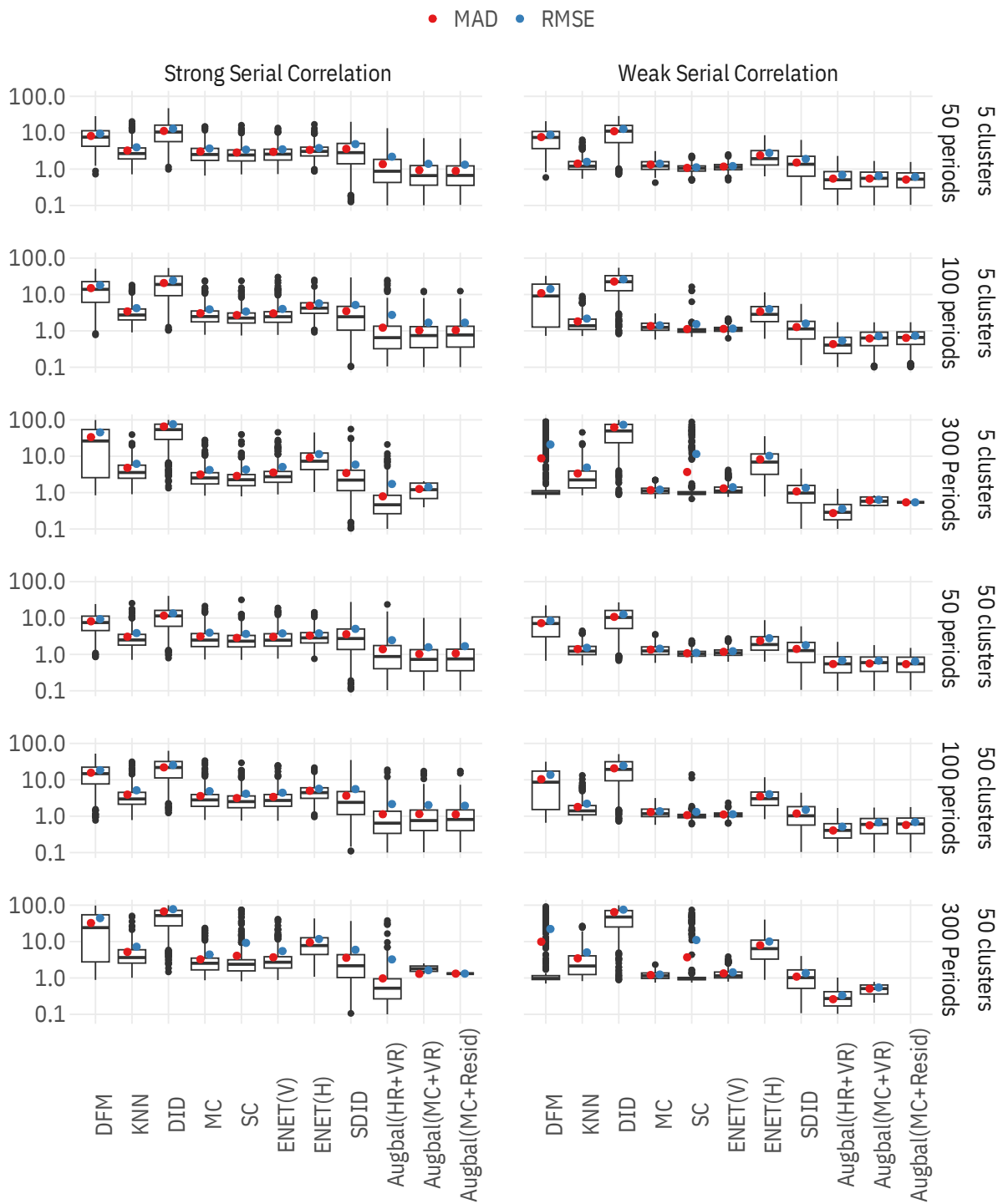




FIGURE A8. 100 Units: 10 Treated Units Simulation Results

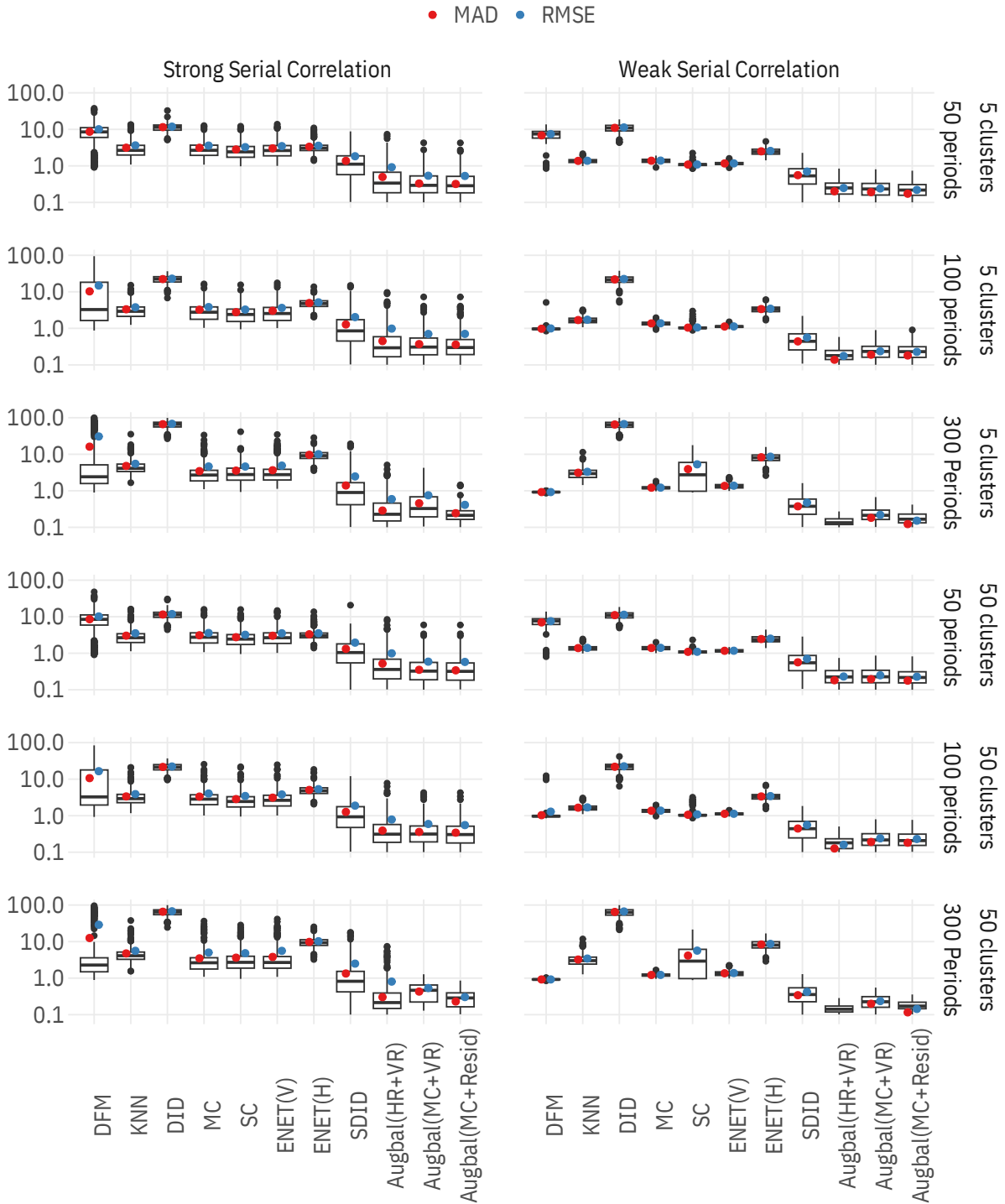
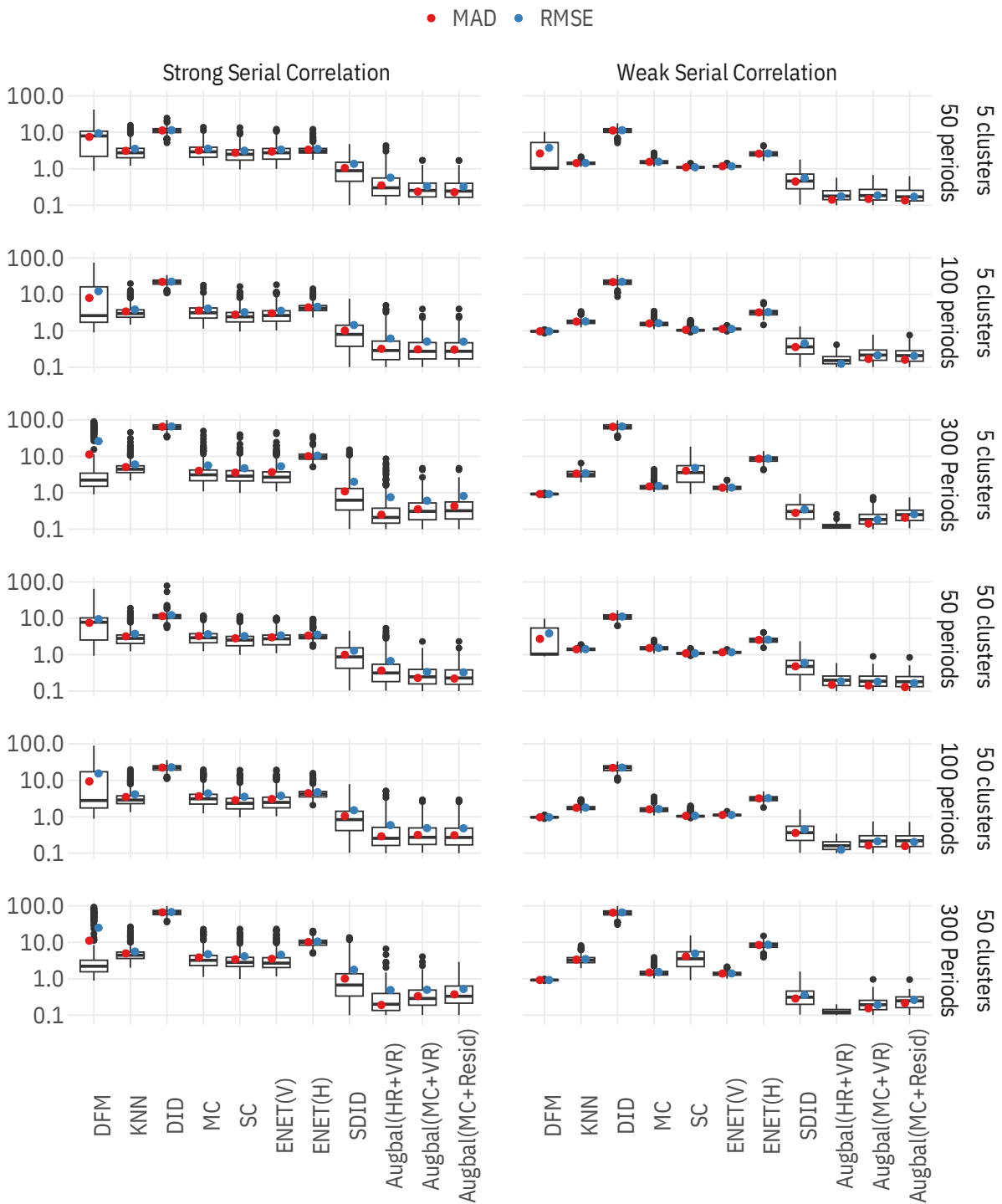


FIGURE A9. 100 Units: 25 Treated Units Simulation Results



## Appendix B. Computation

Bach et al. (2021), Wickham (2010), Dowle et al. (2018), Bergé (2018), Athey, Tibshirani, and Wager (2019), and Fu, Narasimhan, and Boyd (2020)

### References

- Athey, Susan, Guido W Imbens, and Stefan Wager (Sept. 2018). “Approximate residual balancing: debiased inference of average treatment effects in high dimensions”. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 80.4, pp. 597–623 (cit. on p. 37).
- Athey, Susan, Julie Tibshirani, and Stefan Wager (Apr. 2019). “Generalized random forests”. *Annals of statistics* 47.2, pp. 1148–1178 (cit. on p. 51).
- Bach, Philipp, Victor Chernozhukov, Malte S Kurz, and Martin Spindler (Mar. 2021). “DoubleML – An Object-Oriented Implementation of Double Machine Learning in R”. arXiv: [2103.09603 \[stat.ML\]](https://arxiv.org/abs/2103.09603) (cit. on p. 51).
- Bergé, Laurent (2018). “Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm”. *CREA Discussion Papers* 13 (cit. on p. 51).
- Dorie, Vincent, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone (2019). “Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition”. *Statistical Science* (cit. on p. 37).
- Dowle, Matt, Arun Srinivasan, Jan Gorecki, Michael Chirico, Pasha Stetsenko, Tom Short, Steve Lianoglou, Eduard Antonyan, Markus Bonsch, and Hugh Parsonage (2018). “Package ‘Data. Table’” (cit. on p. 51).
- Fu, Anqi, Balasubramanian Narasimhan, and Stephen Boyd (2020). “CVXR: An R Package for Disciplined Convex Optimization”. *Journal of Statistical Software* 94, pp. 1–34 (cit. on p. 51).
- Wickham, Hadley (2010). “Ggplot2: Elegant Graphics for Data Analysis”. *J Stat Softw* 35.1, pp. 65–88 (cit. on p. 51).
- Xu, Yiqing and Eddie Yang (2022). “Hierarchically Regularized Entropy Balancing”. *Political Analysis*, pp. 1–8 (cit. on p. 37).