Hierarchically Regularized Entropy Balancing

Yiqing Xu[1] and Eddie Yang[2]

[1] *Department of Political Science, Stanford University. Email: yiqingxu@stanford.edu*
[2] *Department of Political Science, UC San Diego. Email: z5yang@ucsd.edu*

Abstract

We introduce hierarchically regularized entropy balancing as an extension to entropy balancing, a reweighting method that adjusts weights for control group units to achieve covariate balance in observational studies with binary treatments. Our proposed extension expands the feature space by including higher-order terms (such as squared and cubic terms and interactions) of covariates and then achieves approximate balance on the expanded features using ridge penalties with a hierarchical structure. Compared with entropy balancing, this extension relaxes model dependency and improves the robustness of causal estimates while avoiding optimization failure or highly concentrated weights. It prevents specification searches by minimizing user discretion in selecting features to balance on and is also computationally more efficient than kernel balancing, a kernel-based covariate balancing method. We demonstrate its performance through simulations and an empirical example. We develop an open-source `R` package, `hbal`, to facilitate implementation.

Keywords: causal inference, statistical learning, covariate balance, weighting, entropy balancing.

1 Introduction

Entropy balancing (*ebal*) is a popular reweighting method that aims at estimating the average treatment on the treated (ATT) using nonexperimental data with binary treatments (Hainmueller 2012). It adjusts the weights for the control units to achieve exact covariate balance by solving the following constrained maximization problem:

$$\begin{aligned} \max_{w} H(w) = -\sum_{i\in C} w_i log(w_i/q_i); \\ s.t. \qquad \sum_{i\in C} w_i G_{ij} = m_j \text{ for } j \in 1,\ldots,J; \\ \sum_{i\in C} w_i = 1 \text{ and } w \geq 0 \text{ for all } i \in C; \end{aligned}$$

in which $w = \{w_i\}_{i\in C}$ is set of solution weights for units in the control group (denoted as C); $q_i > 0$ is the base weight for unit i (and $\sum_{i\in C} q_i = 1$); $H(\cdot)$ is the Kullback-Leibler divergence between the distributions of the solution weights and base weights; and $\sum_{i\in C} w_i G_{ij} = m_j$ specifies a set of J balance constraints, where $G \in \mathbb{R}^J$ includes J pretreatment covariates and m_j is the mean of the jth covariate of the treatment group.

Despite its appealing properties, such as exact balance, computational efficiency, and double robustness, *ebal* has two main drawbacks. First, it requires researchers to specify the moments of the covariates to be balanced on, which leaves room for specification searching and selective reporting. Second, when the number of control units is small relative to the number of available covariates, the algorithm either does not converge or generates highly concentrated weights. As a result, researchers face a dilemma that balancing on too few terms leads to biases while balancing on too many terms may be infeasible or induce high variance due to extreme weights. To address these problems, we propose hierarchically regularized entropy balancing (*hbal*) as an extension to *ebal*. *hbal* achieves approximate balance on reasonably flexible functions of the covariates through a ridge-regularized entropy balancing framework.

Political Analysis (2022)

DOI: 10.1017/pan.xxxx.xx

Corresponding author
Yiqing Xu

Edited by
Jeff Gill

2 Approximate Balancing with Hierarchical Regularization

Hainmueller (2012) proves that the global solution for each unit i's weight exists and is given by $w_i^{ebal} = \frac{q_i \exp(-G_i'Z)}{\sum_{D_i=0} q_i \exp(-G_i'Z)}$, where $Z = \{\lambda_1, \lambda_2 \dots \lambda_J\}'$ is a set of Lagrange multipliers for the balance and normalizing constraints and $D_i = \{0, 1\}$ is the binary treatment indicator. Using the Lagrangian multipliers and the solution weights w_i^{ebal}, the constrained optimization problem can be rewritten as the following dual problem:

$$\min_Z L^d = \log\left(\sum_{i \in C} q_i \exp\left(-\sum_{j=1}^{J} \lambda_j G_{ij}\right)\right) + \sum_{j=1}^{J} \lambda_j m_j. \tag{1}$$

After obtaining w_i^{ebal}, one can use a difference in means (DIM) approach to estimate the ATT:

$$\hat{\tau}_{ebal} = \frac{1}{n_1} \sum_{D_i=1} Y_i - \sum_{D_i=0} w_i^{ebal} Y_i.$$

in which n_1 is the number of treated units. Zhao and Percival (2016) show that Problem (1) is an M-estimator for the propensity score with a logistic link using G as predictors, and the solution weights w_i^{ebal} belong to a class of inverse probability weights. They also show that under strong ignorability and positivity, $\hat{\tau}_{ebal}$ is a doubly robust estimator for the ATT: when either the untreated potential outcome $Y_i(0)$ or treatment assignment is linear in G, $\hat{\tau}_{ebal}$ is consistent—see Supporting Materials (SM) for details.

In practice, the linearity assumption may be unrealistic. To make this assumption more plausible, we can conduct a series expansion of G, e.g., by including higher-order terms and various kinds of interactions, obtaining $X \in \mathbb{R}^T$, $(T \gg J)$. However, exact balancing on high-dimensional X is often infeasible; even when it is, the large number of balancing constraints may cause the solution weights to be heavily concentrated on a few control units, resulting in high variance of the ATT estimates. *hbal* addresses this problem by modifying the objective function in (1), i.e., adding an ℓ^2 penalty with a hierarchical structure to the Lagrangian multipliers Z:

$$\min_{Z^+} L^d = \log\left(\sum_{D_i=0} q_i \exp\left(-\sum_{t=1}^{T} \lambda_t X_{it}\right)\right) + \sum_{t=1}^{T} \lambda_t m_t + \sum_{k=1}^{K} \alpha_k r_k \tag{2}$$

where $Z^+ = \{\lambda_1 \dots \lambda_T\}'$ is a vector of Lagrangian multipliers corresponding to T moment conditions. $\sum_{k=1}^{K} \alpha_k r_k$ is the newly added penalty term, in which α_k is a scalar tuning parameter that adjusts the strength of penalty for the k^{th} group, for $k = 1, 2, \dots, K$; $r_k = \sum_{t \in P_k} \lambda_t^2$ is the squared ℓ^2 norm of the Lagrangian multipliers (λ_t) for moment conditions in the k^{th} group, in which P_k is the set of their indices. We choose ℓ^2 penalty over ℓ^1 penalty mainly because the former is twice differentiable, making computation much more efficient. This grouped structure allows differential strengths of regularization to be applied to different groups of balance constraints and prioritizes feature groups that have heavy influence on the overall covariate balance between the treatment and control groups. For example, it is possible that two-way interactions are more important to the overall balance in an application than the squared terms of the covariates (see the SM for a performance comparison of hierarchical and nonhierarchical regularization). This optimization problem gives $w_i^{hbal} = \frac{q_i \exp(-X'Z^+)}{\sum_{D_i=0} q_i \exp(-X'Z^+)}$ $(i \in C)$ and

$$\hat{\tau}_{hbal} = \frac{1}{n_1} \sum_{D_i=1} Y_i - \sum_{D_i=0} w_i^{hbal} Y_i.$$

Implementation details. Implementing *hbal* involves several technical details, such as grouping moment conditions, selecting tuning parameters, and prescreening covariates. Due to space

limitations, we only provide a sketch here and offer more details in the SM.

In terms of grouping, we put all the level terms of G in the first group ($k = 1$); two-way interactions ($k = 2$), squared terms ($k = 3$), three-way interactions ($k = 4$), interactions between square and level terms ($k = 5$), and cubic terms ($k = 6$) each represents a separate group. Because the Lagrangian multipliers can be interpreted as covariate coefficients in a logistic regression for propensity scores, shrinking the Lagrangian multipliers differentially enables us to prioritize groups of features in the expanded covariate space that are the more predictive of propensity scores. By default, we impose a hierarchical structure by setting $\alpha_1 = 0$, i.e., *hbal* seeks exact balance on the level terms just like *ebal*, and only regularizing higher moment constraints. When $\alpha_2 = \alpha_3 = \cdots = \alpha_K = 0$, *hbal* is reduced to *ebal* applied to the full series expansion of the covariates. To select the tuning parameters, we combine a trust-region optimization method (Powell 1994) with a V-fold cross-validation procedure that minimizes mean absolute error (MAE) of expanded covariate balance between the held-out subsample of control units and the treated units. This procedure encapsulates the intuition that the optimal Lagrangian multipliers, based on which the solution weights are constructed, should generalize to randomly selected held-out data and result in approximate covariate balance. While the proposed approach is data-driven, we also allow researcher to incorporate their prior knowledge when applying *hbal*, e.g. by imposing custom covariate groupings and by specifying the parameter space of the tuning parameters.

Combining *hbal* with an outcome model. When the number of control units is small relative to the number of moment conditions, *hbal* only achieves approximate balance. To remove bias caused by the remaining imbalance, we suggest combining *hbal* with an outcome model that includes the same set of covariates in the moment conditions, which we label as *hbal*+. Because *hbal* gives higher weights to units that have similar propensity scores to those of the treated units, this strategy leads to efficiency gain for a regression-based double selection approach. When combined with an outcome model, an ATT estimator is given by

$$\hat{\tau}_{hbal+} = \frac{1}{n_1} \sum_{D_i=1} (Y_i - \hat{g}_0(G_i)) - \sum_{D_i=0} w^{hbal} (Y_i - \hat{g}_0(G_i)),$$

where $\hat{g}_0(G_i) = X_i' \hat{\beta}$ is based on a linear regression on the expanded features. Zhao and Percival (2016) show that $\hat{\tau}_{hbal+}$ is consistent for the ATT when either g_0 is linear in X_i or w^{hbal} converges to the logit of the true propensity scores. With non-zero tuning parameters, *hbal* achieves exact balancing on G and only approximate balance on $X \backslash G$. Hence, if the true propensity scores depend on $X \backslash G$, w^{hbal} does not converge to the logit of the true propensity scores, in which case a correctly specified outcome model g_0 ensures the consistency of $\hat{\tau}_{hbal+}$.

Related work. *hbal* builds on a class of preprocessing methods that explicitly seek to achieve approximate covariate balance for causal inference purposes (e.g., Imai and Ratkovic 2014; Zubizarreta 2015; Athey, Imbens, and Wager 2018; Hazlett 2020; Ning, Sida, and Imai 2020). These methods are shown to estimate propensity scores with loss functions targeting good covariate balance (Zhao and Percival 2016; Wang and Zubizarreta 2019; Ben-Michael, Feller, and Rothstein 2021). *hbal* extends this line of research in that it aims at achieving approximate balance in a large covariate space. Hence, *hbal*'s solution weights can be interpreted as penalized propensity scores with a special loss function. Moreover, the balancing approach is closely connected to the survey literature on calibrated weighting, or raking (e.g., Deming and Stephan 1940). The key component of *hbal*, hierarchical ridge regularization, shares similarity with recent research in survey methods that deal with high dimensionality of crosstabs of respondent characteristics (Caughey and Hartman 2017; Tan 2020; Ben-Michael, Feller, and Hartman 2021).

3 Monte Carlo Evidence

To evaluate the performance of *hbal*, we conduct Monte Carlo simulations to compare *hbal* and *hbal+* with five commonly used matching and weighting methods, including inverse propensity score weighting (PSW) (e.g., Hirano, Imbens, and Ridder 2003), covariate balancing propensity score (CBPS) (Imai and Ratkovic 2014), coarsened exact matching (CEM) (Iacus, King, and Porro 2012), entropy balancing (*ebal*) (Hainmueller 2012), and kernel balancing (*kbal*) (Hazlett 2020). To illustrate the advantage of hierarchical regularization, we also report the results from using *ebal* to balance on the serially expanded covariate set (*ebal**). The naive DIM (Raw) estimator is also included as a benchmark.

Design. We use six covariates $G = \{G_1, G_2, G_3, G_4, G_5, G_6\}$, in which $G_1, \ldots, G_4$ are drawn independently from a multivariate normal distribution with mean 0, variance 1, and covariances of 0.2; G_5 and G_6 are independently drawn from Bernoulli distributions with probability of success 0.3 and 0.2 respectively. To simulate a complex treatment assignment process, we use all level terms and a random subset of higher-order terms to be relevant for treatment assignment and generate their individual effect from a normal distribution. Specifically, the treatment assignment indicator is given by $D = 1\{f(G) - 2 + \epsilon > 0\}$, where $f(G)$ is a linear combination of the subset of the serial expansion of G and $\epsilon \overset{i.i.d.}{\sim} N(0, \sqrt{8})$.[1] To capture the empirical variability in the outcome-generating process, we consider three outcome designs with increasing degree of nonlinearity: (1) linear: $Y = G_1 + G_2 + G_3 - G_4 + G_5 + G_6 + u$; (2) nonlinear $Y = G_1 + G_1^2 - G_4 G_6 + u$; and (3) trigonometric: $Y = 2 \times \cos(G_1) - \sin(\pi \times G_2) + u$, with $u \overset{i.i.d.}{\sim} N(0, 1)$ and the true treatment effect fixed at 0 for all units. For PSW, CBPS, CEM, *ebal*, and *kbal*, we include only the original variables G. Note that *kbal* expands the covariate space through Gaussian kernels; for *ebal**, *hbal*, and *hbal+*, we include all 69 covariates in the third-degree series expansion of G.

Results. Figure 1 presents the simulation result with sample size $N = 900$ and control to treatment ratio 5:1. We report additional results with varying outcome designs, sample sizes, and control to treatment ratios in the SM. The comparative performance of *hbal* (or *hbal+*) is similar across different simulation setups.

For the linear outcome design 1, most methods substantially reduce bias as compared to the naive DIM estimator. One exception is *ebal**, whose poor performance is caused by nonconvergence of the *ebal* algorithm with many moment constraints. In comparison, *hbal*'s ability to discriminate balance among covariate moments leads to superior performance in both bias and variance reduction. As the outcome-generating process becomes more complex, methods that rely only on G to estimate propensity scores or weights perform poorly. Compared to *ebal*, *ebal**, and *kbal*, *hbal* and *hbal+* yield estimates with substantially less bias and smaller variance in designs 2 and 3. These results demonstrate that, when the data-generating process is not too far off from a polynomial expansion of the covariates, *hbal*'s hierarchical structure is able to tailor the strengths of regularization to the importance of balance constraints, thus reducing bias while maintaining a relatively small variance of the estimates. In the SM, we provide additional evidence that hierarchical regularization leads to higher correlation between the solution weights and the true propensity scores.

Moreover, *hbal* is also much more computationally efficient and scalable than *kbal*. The bottom right panel of Figure 1 shows the average running time across 500 samples for varying sample sizes. Across sample sizes, *hbal* finds the solution weights using a fraction of *kbal*'s time and *hbal*'s advantage of scalability becomes more evident as the sample size increases.

1. The selected covariates are G_1, G_2, G_3, G_4, G_5,G_6, G_1G_2, G_1G_5, G_2G_3, G_2G_4, G_2G_6, G_3G_6, G_1^2, G_2^2, G_3^2, G_4^2. The selection of higher-order terms is a random draw from square and two-way interaction terms of the covariates. We also include an intercept of -2 in $f(G)$ so that it is centered close to 0. Replication data and code are available at https://doi.org/10.7910/DVN/QI2WP9.

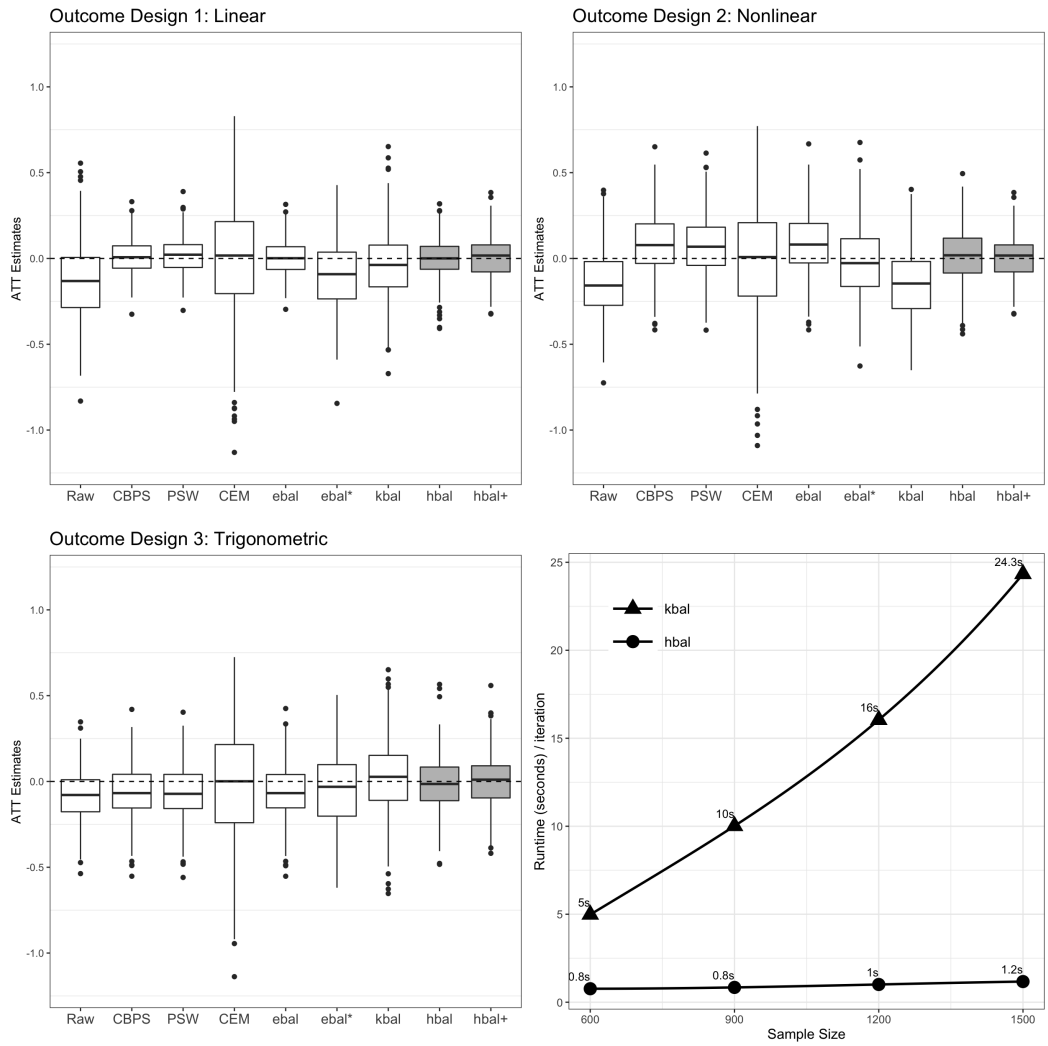

Figure 1. Monte Carlo evidence. The top-left, top-right, and bottom-left panels correspond to results from outcome designs 1, 2, and 3, respectively. All three designs share the same treatment assignment mechanism, i.e., $D = 1\{f(G) - 2 + \epsilon > 0\}$. The bottom-right panel compares the speed between *kbal* and *hbal*.

4 Promotion Prospect and Circuit Court Judge Behavior

To illustrate how *hbal* works in empirical settings, we replicate a recent article by Black and Owens (2016), who study the effect of promotion prospect to the Supreme Court on the behavior of circuit court judges. Using CEM, the authors show that judges who are on the president's shortlist to fill Supreme Court vacancies are more likely to vote in line with the president's position during the vacancy period as compared to the non-vacancy period; they find no such effect among non-contending judges who stand little chance to be nominated to the Supreme Court. We focus on whether circuit court judges ruled in line with the president as the outcome of interest. The binary treatment variable is vacancy period (vs. non-vacancy period). To address potential confounding, the authors use CEM to match cases on seven covariates that might influence a judge's behavior and the treatment, such as the judge's Judicial Common Space (JCS) score, the judge's ideological alignment with the president, and whether the case decision was reversed by the circuit court.

In Figure 2, we compare the results from mean balancing on the level terms of the covariates using *ebal*+ (shown in solid circle) and from balancing on a set of serially expanded covariate using *hbal*+ (shown in solid triangle). To assess whether the *ebal*+ estimate is sensitive to different model specifications, we also include an additional 500 models in which random higher moments of the

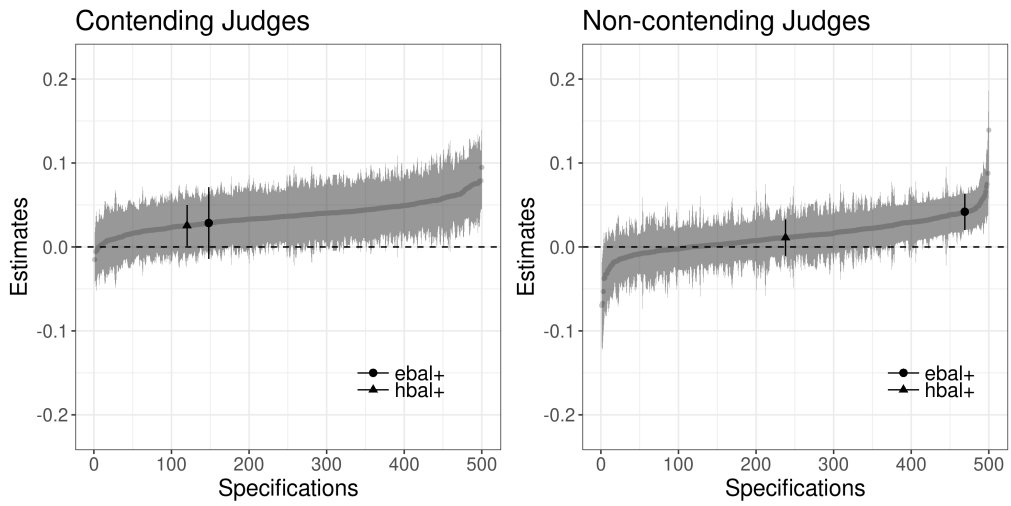

Figure 2. The effect of vacancy period on presidential ideological vote. The solid circle and triangle represent the ATT estimates from *ebal*+ using the original seven covariates and *hbal*+ using covariates after series expansion, respectively, with 95% confidence intervals. The gray bars represent the 95% confidence intervals of 500 *ebal*+ estimates using various combination of the covariates and their higher-order terms.

covariates are included (shown in gray). For both methods, we use the solution weights to estimate a weighted linear regression and report 95% confidence intervals based on standard errors clustered at the individual judge level. For contending judges, estimates from both methods indicate judges are more likely to rule in line with the president during vacancy periods than non-vacancy periods, although the estimate from *ebal*+ using the level terms only is not statistically significant at the 5% level. Because *hbal*+'s specification includes higher-order terms that can explain additional variation in the outcome and treatment assignment, we obtain a more precise and likely more reliable estimate than *ebal*+'s. For non-contending judges, *ebal*+'s estimate suggests that non-contending judges tend to be more likely to rule in line with the president during a vacancy period, while *hbal*+'s estimate shows no significant difference between the vacancy and non-vacancy periods. In short, *hbal*+'s results are broadly in line with Black and Owens (2016)' original findings whereas the estimates from *ebal*+ for both contending and non-contending judges vary widely depending on specifications, ranging from negative to positive effects.

5 Conclusion

In this letter, we extend *ebal* to *hbal* by introducing hierarchical regularization on the Lagrangian multiplier in the transformed objective function. It achieves approximate balance on a potentially large covariate space. Through simulations and an empirical study, we demonstrate *hbal*'s desirable properties in comparison to *ebal* and other commonly used preprocessing methods. We also show that *ebal* is computationally more efficient than *kbal*, another popular covariate balancing method. *hbal* thus can serve as a building block for methods that seek approximate covariate balance. To facilitate implementation, we develop an open source routine, `hbal`, in `R`.

In the SM, we provide more details about the identifying assumptions and theoretical guarantee of *hbal*, its algorithm, implementation procedure, and inferential method, as well as additional information on the simulation results and the empirical example. We also apply *hbal* to the famous Lalonde data and find reassuring results, which are provided in the SM.

Acknowledgements

We thank Ryan Black, Avi Feller, Justin Grimmer, Jens Hainmueller, Chad Hazlett, Molly Offer-Westort, Xiang Zhou, seminar participants at Stanford and UCSD, the Editor of this article Jeff Gill, as well as four anonymous reviewers, for extremely helpful comments.

Data Availability Agreement

Replication data and code for this article has been published at Harvard Dataverse at https://doi.org/10.7910/DVN/QI2WP9.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.xxxx.xx.

References

Athey, S., G. W. Imbens, and S. Wager. 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): 597–623.

Ben-Michael, E., A. Feller, and E. Hartman. 2021. "Multilevel calibration weighting for survey data." *arXiv preprint arXiv:2102.09052.*

Ben-Michael, E., A. Feller, and J. Rothstein. 2021. "The augmented synthetic control method." *Journal of the American Statistical Association* 116 (536): 1789–1803.

Black, R. C., and R. J. Owens. 2016. "Courting the president: how circuit court judges alter their behavior for promotion to the Supreme Court." *American Journal of Political Science* 60 (1): 30–43.

Caughey, D., and E. Hartman. 2017. "Target selection as variable selection: using the Lasso to select auxiliary vectors for the construction of survey weights." *Available at SSRN 3494436.*

Deming, W. E., and F. F. Stephan. 1940. "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known." *The Annals of Mathematical Statistics* 11 (4): 427–444.

Hainmueller, J. 2012. "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies." *Political Analysis* 20 (1): 25–46.

Hazlett, C. 2020. "Kernal balancing: a flexible non-parametric weighting procedure for estimating causal effects." *Statistica Sinica* 30 (3): 1155–1189.

Hirano, K., G. W. Imbens, and G. Ridder. 2003. "Efficient estimation of average treatment effects using the estimated propensity score." *Econometrica* 71 (4): 1161–1189.

Iacus, S. M., G. King, and G. Porro. 2012. "Causal inference without balance checking: Coarsened exact matching." *Political analysis* 20 (1): 1–24.

Imai, K., and M. Ratkovic. 2014. "Covariate balancing propensity score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (1): 243–263.

Ning, Y., P. Sida, and K. Imai. 2020. "Robust estimation of causal effects via a high-dimensional covariate balancing propensity score" [in en]. *Biometrika* 107, no. 3 (June): 533–554.

Powell, M. J. 1994. "A direct search optimization method that models the objective and constraint functions by linear interpolation." In *Advances in optimization and numerical analysis,* 51–67. Springer.

Tan, Z. 2020. "Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data." *Biometrika* 107, no. 1 (March): 137–158.

Wang, Y., and J. R. Zubizarreta. 2019. “Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations” [in en]. *Biometrika* 107, no. 1 (October): 93–105.

Zhao, Q., and D. Percival. 2016. “Entropy balancing is doubly robust.” *Journal of Causal Inference* 5 (1).

Zubizarreta, J. R. 2015. “Stable weights that balance covariates for estimation With incomplete outcome data.” *Journal of the American Statistical Association* 110 (511): 910–922.