Locally Equivalent Weights for Multilevel Regression and Poststratification*

Ryan Giordano, UC Berkeley
Alice Cima, UC Berkeley
Jared Murray, UT Austin
Erin Hartman, UC Berkeley
Avi Feller, UC Berkeley

May 11, 2026

Abstract

Multilevel regression and poststratification (MrP) has become a workhorse method for estimating population quantities from non-probability surveys, and is the primary model-based alternative to traditional survey calibration weighting methods, such as raking. For simple linear regression models, MrP methods admit "equivalent weights", allowing for direct comparisons between MrP and traditional calibration weighting. Such weights, however, have been unavailable for the most widely used MrP models, such as logistic regression. In this paper, we develop a natural generalization, "MrP locally equivalent weights" (MrPlew), which represent MrP as a weighting-style estimator that is locally equivalent to calibration weights near the observed responses. This enables a suite of standard weighting diagnostics, including frequentist sampling variability, covariate balance, and subgroup contribution. We formally justify the use of MrPlew in these cases: we prove the MrPlew-based variance estimator is asymptotically equivalent to the infinitesimal jackknife for common exponential family models, and we introduce a novel class of model checks based on invariance to data perturbations that generalize covariate balance and subgroup contribution to nonlinear models. We further show that MrPlew can be computed easily using existing MCMC samples and provide open-source software to compute MrPlew using the output of standard software. We illustrate our approach for several canonical studies that use MrP, including via a logistic regression outcome model, showing that implied covariate balance can sometimes be worse for MrP than for raking. Given the ease of computing, we recommend making MrPlew a standard part of the MrP model interrogation workflow.

1 Introduction

Multilevel regression and poststratification (MrP; Gelman and Little, 1997) has become a workhorse method for estimating population quantities from non-probability surveys, and is the primary model-based alternative to traditional survey calibration weighting procedures such as raking. MrP adjusts for survey nonresponse and nonprobability sampling by modeling the relationship between the survey response and observed covariates ("multilevel regression") applied to a specific target ("poststratification"). Estimates are typically obtained from approximate posterior draws from a nonlinear model, computed via Markov chain Monte Carlo (MCMC).

Calibration weighting (CW) instead constructs estimates as weighted averages of survey responses, where the weights are chosen to exactly or approximately balance observed covariates subject to some user-defined dispersion penalty (Deville and Särndal, 1992; Haziza and Beaumont, 2017). A key advantage of CW

*Corresponding author: rgiordano@berkeley.edu .

methods relative to MrP methods is their interpretability: visual and quantitative inspection of the weights gives direct insight into the CW procedure itself. For example, the variability of CW weights directly determines frequentist sampling variance. Practitioners can use CW weights to check "covariate balance" to assess whether the weights correct for differences in observable quantities between survey and target populations. CW weights also directly measure the contribution of particular subgroups, such as states or demographic groups, to the final estimate. In short, the weights themselves serve as a rich set of diagnostic tools.

In certain special cases, MrP estimates can be re-written as CW procedures with *equivalent weights*, or as a weighting representation of the MrP procedure; examples include simple linear regression models, Gaussian process (GP) estimation with fixed kernels, and regression trees with fixed structure (Gelman, 2007; Park and Fuller, 2009; Ben-Michael, Feller, and Hartman, 2024). Such weights enable the whole suite of diagnostics available for calibration weighting as well as apples-to-apples comparisons for understanding differences between MrP and CW estimates. Unfortunately, globally valid equivalent weights are unavailable for the most widely used MrP models, such as logistic regression and hierarchical models with estimated random effect variances (Lopez-Martin, Phillips, and Gelman, 2026). For these models, the MrP estimate is essentially a computational black box.

In this paper, we propose a natural generalization, "MrP locally equivalent weights" (MrPlew), which represent MrP as a weighting-style estimator similar to CW, but only for response vectors near the original set of responses, in a sense we make precise. While MrPlew is not (and cannot be in general) a globally equivalent weighting representation, we formally justify the use of MrPlew *as if they were globally equivalent weights* for use in common CW diagnostics such as frequentist sampling variability, comparisons of the weighted sample to the target population (which we refer to as covariate balance), and subgroup contribution. We further show that MrPlew can be computed easily using existing MCMC samples and provide open-source software to compute MrPlew using the output of standard MCMC software like `brms` (Bürkner, 2017).

To develop the theoretical framework for MrPlew, we make two main technical contributions. First, we formally prove the asymptotic equivalence of the MrPlew-based variance estimator with the infinitesimal jackknife variance estimator (Giordano and Broderick, 2024). This result justifies the use of MrPlew for assessing frequentist sampling variability, enabling an apples-to-apples comparison with the sampling variance of CW estimators. Second, we define a novel class of model checks based on *invariance to data perturbations*; in the linear case, these reduce to familiar diagnostics. For the nonlinear case, we formally prove that the MrPlew weights can be used to assess this invariance locally in an asymptotic regime, uniformly over a large class of potential perturbations. This result justifies the use of MrPlew for assessing local covariate balance and subgroup contribution. We discuss the gap between our local theoretical results and practically interesting larger perturbations, and recommend a method for checking the validity of our diagnostics in practice.

Finally, we apply MrPlew-based model diagnostics to several real-world MrP analyses. First, we consider an analysis that extrapolates a Twitter survey of name changes after marriage to the entire US population (Alexander, 2019; Cohen, 2019). Second, we consider a textbook example analyzing the 2020 US presidential election (Alexander, 2023). Third, we consider an example of correcting sampling bias in a national survey on support for same-sex marriage (Lax and Phillips, 2009; Kastellec, Lax, and Phillips, 2010). Across the three

examples, we directly compare the weights themselves, the frequentist variability, and the implied covariate balance, finding meaningful divergences, most strikingly in covariate balance on unmodeled interactions.

1.1 Introductory Example: Name Change

To preview our results, we replicate an MrP analysis from Alexander (2019) of the Marital Name Change Survey (MNCS; Cohen, 2019). The original survey is a convenience sample from Twitter respondents, and the goal is to estimate the corresponding rate in the overall US population. Following Alexander (2019), we limit the survey data to (self-reported) women married to men ($N_S = 4{,}364$), and take our response of interest y to be a binary indicator of whether the woman retains her surname when marrying; the survey average is $\bar{y} = 0.46$.

For the "multilevel regression" part of MrP, we estimate a hierarchical logistic regression with age group, education level, state of residence, and decade married:

$$y \sim \texttt{logit((1 | age\_group) + (1 | educ\_group) + (1 | state) + (1 | decade\_married))}.$$

We fit the model using `brms::brm` with the default priors. For the "poststratification" part of MrP, we match the overall US population using statistics from the 2017–2022 ACS survey (Ruggles et al., 2024). Finally, we estimate corresponding calibration weights via *raking*, with some coarsening of the categories. Section 5 gives full details of the analysis and further results.

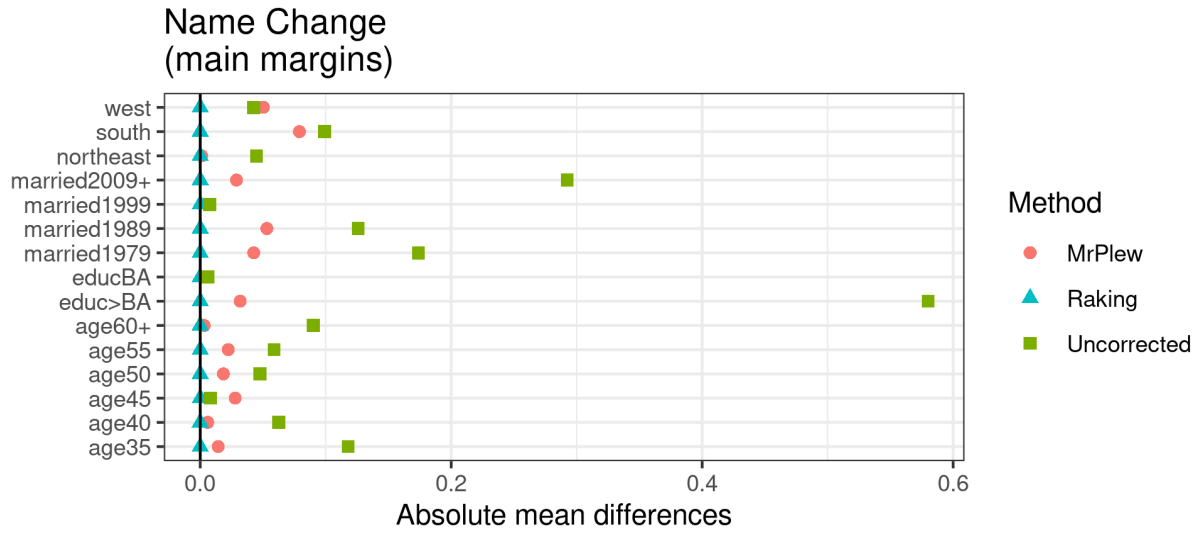


Figure 1: Balance

Figure 1 shows that there are substantial differences between the covariate distribution in the survey and in the target population. Both MrP and raking then make substantial adjustments, shifting the point estimate from $\bar{y} = 0.46$ to $\hat{\mu}^{\text{MrP}} = 0.29$ and $\hat{\mu}^{\text{CW}} = 0.26$. This is where comparisons between CW and MrP would typically stop; the Bayesian analyst would instead proceed with standard model checks as part of the

Bayesian workflow (see, for example Kennedy, Vehtari, and Gelman, 2023; Kuh et al., 2024; Lopez-Martin, Phillips, and Gelman, 2026).

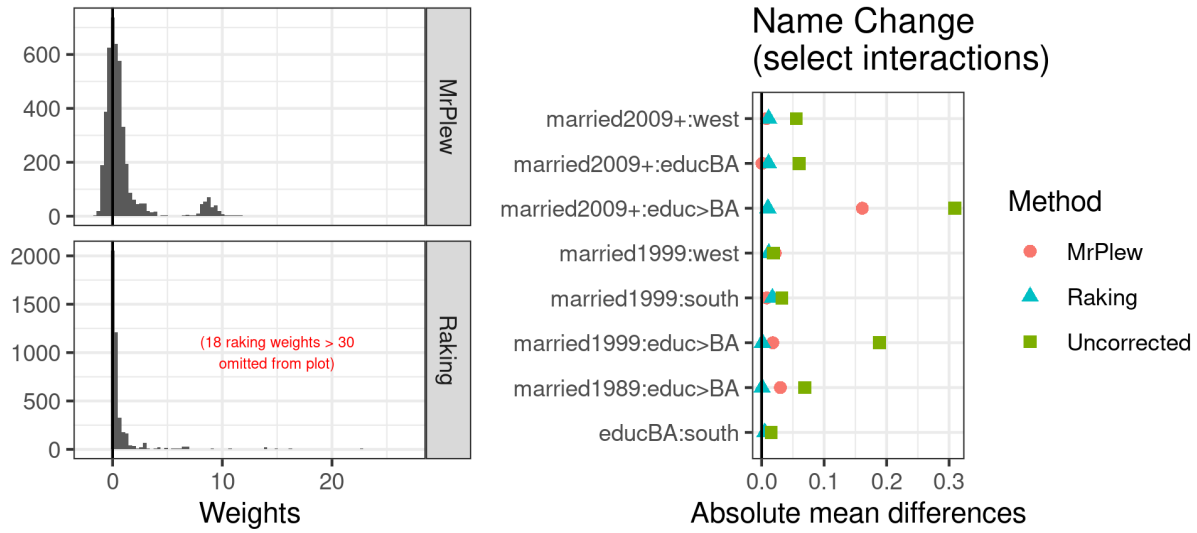


Figure 2: Preview of MrP Diagnostics made possible by MrPlew for the Name Change analysis

Our goal is to enable a suite of additional diagnostics based on MrP locally equivalent weights. First, Figure 1 shows that the implied covariate balance, measured by comparing the MrPlew weighted covariate means to the population means, for MrP is excellent for the main margins; raking balances these exactly by construction. However, Figure 2 shows substantial imbalance on a key interaction term between decade married and education level. Even though raking and MrP only explicitly adjust for the margins of these groups, raking nonetheless balances this interaction while MrP does not. Section 5 includes additional analyses to explore the impact of this imbalanced interaction.

Figure 2 also compares the raking weights and the locally equivalent MrP weights themselves. The latter weights are clustered much more around zero, with a substantial proportion of negative weights. Many raking weights, however, are extreme ($w_i^{\text{CW}} > 30$), which appears to drive up the overall variability of raking versus MrP. Using our results, we can assess this directly: after scaling by $\sqrt{N_S}$, the estimated frequentist sampling standard deviation is 1.39 for raking and 1.06 for the MCMC MrP procedure.

To the best of our knowledge, such direct comparisons of the strengths and weaknesses of MrP and calibration weighting methods on the same dataset were not previously possible.

1.2 Literature review

Survey calibration and soft calibration. Calibration weighting adjusts survey estimates by finding weights that balance observed covariates between sample and population (Deville, Särndal, and Sautory, 1993; Fuller, 2011; Deville and Särndal, 1992; Haziza and Beaumont, 2017). Wu and Sitter (2001) extend this to *model calibration*, where the calibration targets are derived from a fitted model rather than from the sampling design. A key recent development is "soft" calibration, which relaxes exact balance to allow

approximate balance on higher-order interactions (Park and Fuller, 2009; Guggemos and Tillé, 2010; Wang and Zubizarreta, 2020; Ben-Michael, Feller, and Hartman, 2024). In particular, Gao, Yang, and Kim (2023) make the connection to random-effects models precise: under a shared random-effects structure, the optimal calibration weights impose exact calibration on fixed effects and approximate calibration on random effects.

Equivalent weights. Survey researchers have long known that regression adjustment has an equivalent weighting form (e.g., Park and Fuller, 2009). In a foundational paper, Gelman (2007) applies this insight to a Bayesian Normal-Normal outcome model, showing that the MrP estimate can be re-written as a calibration weighting estimator with globally valid "equivalent weights." Ben-Michael, Feller, and Hartman (2024) extend this, showing that multilevel calibration weights are equivalent to the MAP estimate of a multilevel outcome model under certain conditions; see Chattopadhyay and Zubizarreta (2023) and Bruns-Smith et al. (2025) for parallel results from the causal inference literature.

Calibrated Bayes and Bayesian survey inference. Little (2004) frames the central tension in survey inference as "to model or not to model?" and proposes *calibrated Bayes* (Little, 2006; Little, 2012) as a resolution: use Bayesian models, but choose them so that the resulting procedures have good frequentist properties. A growing body of work pursues this vision, incorporating design information into Bayesian models through weighted pseudo-posteriors (Savitsky and Toth, 2016; Wang, Kim, and Yang, 2018), model-based survey weights (Si, Trangucci, et al., 2020; Si, Pillai, and Gelman, 2015), and Bayesian analogs of raking (Si and Zhou, 2021). Most recently, Gelman, Si, and West (2024) propose a framework to incorporate design weights into a Bayesian MrP model by jointly modeling the outcome y and the sampling weight w given covariates x, then poststratifying on (x, w). This is the complement to our approach, incorporating design weights into a broader Bayesian model, instead of finding locally equivalent weights of the original Bayesian model.

Extensions and diagnostics for MrP. There has been an explosion of interest in MrP; see the recent textbook from Lopez-Martin, Phillips, and Gelman (2026). Extensions include deep interactions (Ghitza and Gelman, 2013), tree-based methods (Montgomery and Olivella, 2018; Bisbee, 2019), and doubly robust-style combinations of weighting and outcome modeling (Chen, Li, and Wu, 2020; Ben-Michael, Feller, and Hartman, 2024). Despite this progress, model diagnostics for MrP remain limited. Kennedy, Vehtari, and Gelman (2023) and Kuh et al. (2024) propose cross-validation-based diagnostics, but these assess predictive accuracy rather than how the survey data are being used by the estimator. Meng (2018) introduces the data defect index, measuring the correlation between sample inclusion and the outcome. Our contribution can be understood as expanding the suite of tools for assessing model stability as part of veridical data science for the Bayesian workflow (Yu and Kumbier, 2020; Gelman, Vehtari, et al., 2020).

Local robustness. Our contributions follow in the tradition of the Bayesian local robustness literature, which studies the effect of infinitesimal perturbations to Bayesian posteriors (Basu, Jammalamadaka, and Liu, 1996; Gustafson, 1996; Gustafson, 2000; Giordano, Broderick, and Jordan, 2018; Giordano, Liu, et al., 2023b; Cabral, Bolin, and Rue, 2025; Di Noia, Ruggeri, and Mira, 2025), as well as related local sensitivity ideas in the frequentist case influence literature (Belsley, Kuh, and Welsch, 2005; Cook, 1977; Cook, 1986; Kass, Tierney, and Kadane, 1989; Zhu et al., 2007; Koh and Liang, 2017; Giordano, Stephenson,

et al., 2019; Thomas, MacEachern, and Peruggia, 2018). Our core technical results build on Giordano and Broderick (2024). Specifically, we prove frequentist consistency using the proof of consistency of the infinitesimal jackknife (IJ) estimator for Bayesian posterior expectations found in Giordano and Broderick (2024, Theorem 2), and our result for covariate balance follows from an extension of the series expansions of posterior expectations of Giordano and Broderick (2024, Theorem 1) to hold uniformly over a set of posteriors.

2 Methods

2.1 Problem setup

We frame our problem of interest in terms of estimating a population quantity from a non-representative sample, though the same formal problem arises in observational causal inference and in domain adaptation for regression more broadly. We observe scalar survey responses, denoted y, and vector-valued regressors, denoted $\boldsymbol{x}$; in the context of the Marital Name Change Survey in Section 1.1, y_i is a binary indicator of whether a married woman retains her surname and $\boldsymbol{x}_i$ collects demographic regressors like age and education level. We additionally observe regressors $\boldsymbol{x}$ from a target population, for which the responses are unobserved; in the running example, the target is the corresponding population of the United States. The problem is to infer the expected value of the response in the target population, under the assumption that the conditional distribution of the response given observed regressors is the same in the survey and target populations.

Notation. Let $\{(\boldsymbol{x}_i, y_i) : i \in [N_S]\}$ denote the survey data, where $[N_S] = \{1, \ldots, N_S\}$, y_i is a scalar survey response, $\boldsymbol{x}_i \in \mathbb{R}^P$ is a vector of regressors, and N_S is the number of survey observations. Let $\{\boldsymbol{x}_j : j \in [N_T]\}$ denote regressors observed for N_T units drawn from the target population; the corresponding responses are not observed. We write $\boldsymbol{Y} = (y_1, \ldots, y_{N_S})^\intercal$, $\boldsymbol{X}$ for the $N_S \times P$ matrix of survey regressors, and $\boldsymbol{X}_T$ for the $N_T \times P$ matrix of target regressors. The responses y_i may be continuous or discrete, though we are particularly interested in the binary case.

With some abuse of notation, the symbol $\boldsymbol{x}$ (without index) will denote a generic random variable in both the survey and target distributions. To avoid ambiguity, we adopt the following convention for expectations: for a random variable $\boldsymbol{z}$ with distribution $\mathcal{P}(\boldsymbol{z})$ and measurable function ϕ,

$$\mathbb{E}_{\mathcal{P}(\boldsymbol{z})}\left[\phi(\boldsymbol{z})\right] := \int \phi(\boldsymbol{z})\, \mathcal{P}(d\boldsymbol{z}),$$

with all other quantities taken as fixed. For example, $\mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}\left[\boldsymbol{x}y\right]$ is a function of $\boldsymbol{x}$ but not of y. Covariances follow the analogous convention.

Formal setup. The surveys literature has several distinct traditions for formalizing adjustment procedures; see Elliott and Valliant (2017). Following the MrP literature (Lopez-Martin, Phillips, and Gelman, 2026), we anchor our discussion in the super-population approach (Chang and Kott, 2008), though we expect our results to extend to the quasi-randomization approach (Kott and Chang, 2010). Our contributions are best understood as *diagnostics*: the assumptions below motivate the estimators we study, but the diagnostics

themselves do not rely on them.

We assume $\boldsymbol{x}$ has distribution $\mathcal{P}_S(\boldsymbol{x})$ in the survey and $\mathcal{P}_T(\boldsymbol{x})$ in the target, with these two distributions allowed to differ. The CW and MrP estimators we study are motivated by three assumptions.

Assumption 2.1 (Invariance)**.** The conditional distribution $\mathcal{P}(y|\boldsymbol{x})$ is the same in the survey and target populations. □

This assumption rules out any unmeasured factors that shift the response beyond what the observed regressors $\boldsymbol{x}$ capture. In the Name Change application, invariance is plausible only to the extent that age, education level, state, and decade of marriage fully capture the decision to change names.

Assumption 2.2 (Overlap)**.** $\mathcal{P}_T(\boldsymbol{x})$ is absolutely continuous with respect to $\mathcal{P}_S(\boldsymbol{x})$. □

This assumption requires that any $\boldsymbol{x}$ that occurs with positive probability in the target also occurs with positive probability in the survey. Overlap can be difficult to justify in convenience samples, where large strata of the target may be thinly represented or entirely absent in the survey.

Assumption 2.3 (IID sampling)**.** $(\boldsymbol{x}_i, y_i) \overset{iid}{\sim} \mathcal{P}_S(\boldsymbol{x})\mathcal{P}(y|\boldsymbol{x})$ for $i \in [N_S]$ and $\boldsymbol{x}_j \overset{iid}{\sim} \mathcal{P}_T(\boldsymbol{x})$ for $j \in [N_T]$. □

We state our asymptotic results under IID survey sampling for simplicity; the IID assumption on the target is not essential.

Our goal is to estimate

$$\mu := \mathbb{E}_{\mathcal{P}_T(\boldsymbol{x},y)}\left[\pi(\boldsymbol{x})y\right]$$

for some known weighting function $\pi(\boldsymbol{x})$. In the simplest case $\pi(\boldsymbol{x}) \equiv 1$ and μ is the target-population mean of y. This is the estimand in our Name Change example, corresponding to the overall U.S. rate of women retaining their surname. More generally, $\pi(\boldsymbol{x})$ accommodates subgroup means (e.g., rates within a state) and contrasts between subgroups, both of direct substantive interest. For compactness, we write $\pi_j := \pi(\boldsymbol{x}_j)$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{N_T})^\intercal$. If we observed y_j for $j \in [N_T]$, a natural estimator would be

$$\tilde{\mu} := \frac{1}{N_T}\sum_{j\in[N_T]} \pi_j y_j \qquad \text{(Infeasible)},$$

but y_j is unobserved outside the survey. We now turn to feasible estimators of this quantity under the assumptions above.

2.2 Survey adjustment: Calibration weighting

2.2.1 Overview

The classical approach to survey adjustment is *calibration weighting* (Deville and Särndal, 1992), with estimators of the form

$$\hat{\mu}^{\mathrm{CW}} := \frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\mathrm{CW}} y_i \quad \text{for some } \boldsymbol{W}^{\mathrm{CW}} := (w_1^{\mathrm{CW}}, \ldots, w_{N_S}^{\mathrm{CW}})^\intercal. \tag{1}$$

There is a rich literature on estimators of this form, reviewed in Section 1.2; see Haziza and Beaumont (2017) for a modern overview.

To build intuition, consider the infeasible case in which we observe target responses y_j. Then the weights should satisfy:

$$\frac{1}{N_T}\sum_{j\in[N_T]}\pi_j y_j - \frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\mathrm{CW}} y_i \overset{\text{check}}{\approx} 0, \qquad \text{(Infeasible)} \tag{2}$$

where $\overset{\text{check}}{\approx}$ denotes an approximate equality treated as a check on the model.

We cannot compute eq. (2) because we do not observe target responses y_j. However, we can write an analogous equation for the regressors. Let $r(\boldsymbol{x})$ denote some measurable function of $\boldsymbol{x}$, with $r_i = r(\boldsymbol{x}_i)$ and $r_j = r(\boldsymbol{x}_j)$ in the survey and target, respectively. Ideally, $r(\cdot)$ is predictive of either the outcome y or selection into the survey; see Särndal and Lundström (2005) or Ben-Michael, Feller, Hirshberg, et al. (2021) for discussion. Then *covariate balance* for $r(\cdot)$ measures the difference

$$\text{Imbalance}(r, \boldsymbol{W}^{\mathrm{CW}}) := \frac{1}{N_T}\sum_{j\in[N_T]}\pi_j r_j - \frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\mathrm{CW}} r_i \overset{\text{check}}{\approx} 0. \tag{3}$$

We use the term covariance balance to refer to the comparison of the weighted sample distribution of covariates to the target population, however it is also sometimes referred to as external consistency (Haziza and Beaumont, 2017). *Raking* (Deming and Stephan, 1940; Deville and Särndal, 1992) finds the weights $\boldsymbol{W}^{\mathrm{CW}}$ that minimize a dispersion criterion (e.g., entropy or variance) subject to $\text{Imbalance}(r, \boldsymbol{W}^{\mathrm{CW}}) = 0$ for a user-chosen set of covariates r. In the Name Change application, we fit raking weights using the `survey::calibrate` function with `calfun = "raking"` (Lumley, 2024), with entropy as the dispersion criterion and ACS population counts as targets. Alternative calibration estimators consider different dispersion functions and imbalance constraints. An important example we return to below is *soft calibration* (Gao, Yang, and Kim, 2023; Ben-Michael, Feller, and Hartman, 2024), which bounds $\text{Imbalance}(r, \boldsymbol{W}^{\mathrm{CW}})$ for some r rather than requiring exact balance.

A key feature of calibration weighting estimators is that they are *design-based*: the weights are estimated without using the survey responses y, which enter only through the weighted average in Equation (1). We assume throughout that the calibration weights are design-based in this sense.

2.2.2 Diagnostics for calibration weighting

Since CW estimators are design-based and linear in the responses y, practitioners can directly leverage the weights to understand how the estimator is constructed and to diagnose potential problems. We briefly review three common diagnostics for CW estimators, which we will later adapt to MrP.

Direct inspection of the weights and subgroup contribution. The weights themselves offer the most important initial diagnostic, for example by identifying observations with extreme weights. A natural diagnostic looks at this same influence but aggregated up to different groups; for example, how much do

survey respondents from each state contribute to the overall estimate? Formally, for some partition $\mathcal{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ of the regressor space (e.g. into US states), we can compute $\frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\text{CW}}\mathbb{I}(\boldsymbol{x}_i \in \mathcal{A}_k)$, where $\mathbb{I}(\cdot)$ is the indicator function. We provide concrete examples of this in Section 5.5.

Covariate balance. As outlined above, the covariate balance metric in Equation (3) is a key diagnostic for CW estimators. Raking explicitly sets $\text{Imbalance}(r, \boldsymbol{W}^{\text{CW}}) = 0$ for the covariates used in calibration, such as the margins, but the same metric lets us check imbalance on regressors *outside* the raking set — e.g., interactions or nonlinear transformations of the original regressors.

Frequentist variability. The weights also directly determine the frequentist variability of the CW estimator $\hat{\mu}^{\text{CW}}$. Conditional on $\boldsymbol{W}^{\text{CW}}$, the variance is:

$$\text{Var}_{\mathcal{P}(\boldsymbol{Y}|\boldsymbol{W}^{\text{CW}},\boldsymbol{X})}\left(\hat{\mu}^{\text{CW}}\right) = \frac{1}{N_S^2}\sum_{i\in[N_S]}\left(w_i^{\text{CW}}\right)^2 \text{Var}_{\mathcal{P}(y|\boldsymbol{x}_i)}(y_i)\ . \tag{4}$$

Ignoring structure in the conditional response variances, the conditional variance of $\hat{\mu}^{\text{CW}}$ is minimized when the weights are equal and grows as they become more dispersed.

Our goal is to develop analogs of these diagnostics for Bayesian survey adjustment methods, which we turn to next.

2.3 Survey adjustment: Multilevel regression and poststratification

2.3.1 Overview

Unlike CW, MrP explicitly models the outcome y as a function of $\boldsymbol{x}$, targeting the conditional mean rather than the density ratio (Gelman, 1997; Lopez-Martin, Phillips, and Gelman, 2026). Given a posited outcome model $\mathcal{P}(y|\boldsymbol{x})$ and a corresponding estimate $\hat{y}(\boldsymbol{x}) \approx \mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[y]$, the analyst sets $\hat{y}_j = \hat{y}(\boldsymbol{x}_j)$ for each target $\boldsymbol{x}_j$ and forms

$$\hat{\mu}^{\text{Generic MrP}} = \frac{1}{N_T}\sum_{j\in[N_T]}\pi_j\hat{y}_j. \tag{5}$$

To motivate Equation (5), note that if our estimates $\hat{y}_j$ are accurate in the sense that

$$\mathbb{E}_{\mathcal{P}_T(\boldsymbol{x}_j)}[\pi_j\hat{y}_j] \approx \mathbb{E}_{\mathcal{P}_T(\boldsymbol{x}_j)}\left[\pi(\boldsymbol{x}_j)\mathbb{E}_{\mathcal{P}(y|\boldsymbol{x}_j)}[y]\right] = \mathbb{E}_{\mathcal{P}_T(\boldsymbol{x},y)}[\pi(\boldsymbol{x})y] = \mu,$$

then $\hat{\mu}^{\text{Generic MrP}}$ is (nearly) unbiased under sampling from the target distribution.

The *multilevel regression* step estimates the function $\hat{y}(\cdot)$ from the survey data, typically via a multilevel model; the *poststratification* step averages $\hat{y}_j$ over the target $\boldsymbol{x}_j$ in Equation (5). We focus on generalized linear models of the form $\mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[y] = m(\boldsymbol{\beta}^\intercal\boldsymbol{x})$ for some inverse link $m(\cdot)$, where $\boldsymbol{x}$ may contain non-linear transformations of the observed regressors. For OLS, $m(\cdot)$ is the identity and $\hat{y}_j^{\text{OLS}} = \hat{\boldsymbol{\beta}}^\intercal\boldsymbol{x}_j$; for logistic regression, $m^{\text{logit}}(\boldsymbol{z}) := 1/(1+\exp(-\boldsymbol{z}))$ is the logistic link and $\hat{y}_j = m^{\text{logit}}(\hat{\boldsymbol{\beta}}^\intercal\boldsymbol{x}_j)$.

2.3.2 Bayesian computation

We focus on *Bayesian hierarchical* estimates of $\hat{y}_j$. A Bayesian estimator posits a likelihood $\mathcal{P}(y|\boldsymbol{x}, \boldsymbol{\beta})$, in this case a parametric model for the full conditional distribution of y given $\boldsymbol{x}$; for any $\boldsymbol{\beta}$, this gives a corresponding estimate $\mathbb{E}_{\mathcal{P}(y|\boldsymbol{x},\boldsymbol{\beta})}[y]$. We specify a possibly hierarchical prior $\mathcal{P}(\boldsymbol{\beta})$, expressed marginally over any hyperparameters throughout. While not central to our discussion, we note that hierarchical priors can induce non-linearity in posterior estimates as a function of $\boldsymbol{Y}$ via the implicit estimation of variance parameters. Finally, we leave the posterior's dependence on $\boldsymbol{X}$ implicit, since its distribution is ancillary.

Writing $\ell(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) := \log \mathcal{P}(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})$, let $\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})$ denote the posterior of $\boldsymbol{\beta}$ given $\boldsymbol{Y}$ via Bayes' rule,

$$\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y}) := \frac{\mathcal{P}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta})\mathcal{P}(\boldsymbol{\beta})}{\int \mathcal{P}(\boldsymbol{Y}|\tilde{\boldsymbol{\beta}}, \boldsymbol{X})\mathcal{P}(\tilde{\boldsymbol{\beta}})d\tilde{\boldsymbol{\beta}}} = \frac{\exp\left(\sum_{i\in[N_S]} \ell(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})\right)\mathcal{P}(\boldsymbol{\beta})}{\mathcal{P}(\boldsymbol{Y}|\boldsymbol{X})}. \tag{6}$$

We then form

$$\hat{y}_j^{\text{Bayes}} = \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}[m(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)] \quad \text{and} \quad \hat{\mu}^{\text{MrP}} = \frac{1}{N_T}\sum_{j\in[N_T]} \pi_j \hat{y}_j^{\text{Bayes}}. \tag{7}$$

The MrP family also includes optimization-based and other approaches, but we focus on this Bayesian estimator throughout.

In practice, $\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})$ is not available in closed form, and Equation (7) is estimated by Markov chain Monte Carlo (MCMC). Given posterior draws $\boldsymbol{\beta}_k$, $k \in [M]$, we construct MCMC estimates

$$\hat{y}_j^{\text{MCMC}} = \frac{1}{M}\sum_{k\in[M]} m(\boldsymbol{\beta}_k^\intercal \boldsymbol{x}_j) \approx \hat{y}_j^{\text{Bayes}},$$

and plug these into Equation (5) in place of $\hat{y}_j^{\text{Bayes}}$. We treat $\hat{\mu}^{\text{MrP}}$ as if computed from the true posterior, flagging MCMC issues where relevant.

Returning to the Name Change application, we fit the hierarchical logistic model of Section 1.1 via `brms::brm` (Bürkner, 2017) with default priors, running four MCMC chains of 2,000 iterations each (500 warmup). Posterior means $\hat{y}_j^{\text{MCMC}}$ are then averaged over the ACS target population to form $\hat{\mu}^{\text{MrP}}$.

2.3.3 Globally equivalent weights for linear outcome models

In general, the MrP estimator $\hat{\mu}^{\text{MrP}}$ depends on the survey responses y_i through the posterior $\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})$ and through the inverse link function $m(\cdot)$. As a consequence, the mapping $\boldsymbol{Y} \mapsto \hat{\mu}^{\text{MrP}}(\boldsymbol{Y})$ can be highly nonlinear. As a result, we cannot in general directly apply the diagnostics described in Section 2.2.2, which rely on the linearity of $\hat{\mu}^{\text{CW}}$ in y_i.

However, an important special case arises when the outcome model is linear. In a foundational paper, Gelman (2007) shows that the MrP estimator with a linear outcome model can be expressed in closed form as a CW estimator with specific weights, known as *equivalent weights*. For a simple OLS outcome model

(see Park and Fuller, 2009), the equivalent weights are given by:

$$\hat{\mu}^{\text{OLS}} = \frac{1}{N_T} \sum_{j \in [N_T]} \pi_j \hat{y}_j^{\text{OLS}} \tag{8}$$

$$= \frac{1}{N_S} \sum_{i \in [N_S]} w_i^{\text{OLS}} y_i \quad \text{where} \quad w_i^{\text{OLS}} := \frac{N_S}{N_T} \boldsymbol{\pi}^\intercal \boldsymbol{X}_T (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{x}_i. \tag{9}$$

It follows that the map $\boldsymbol{Y} \mapsto \hat{\mu}^{\text{OLS}}(\boldsymbol{Y})$ is linear, and that w_i^{OLS} can be used for all the diagnostics described in Section 2.2.2. Moreover, $\hat{\mu}^{\text{OLS}}$ is still linear under ridge penalization and therefore when $\hat{\boldsymbol{\beta}}$ is the posterior mean of a Bayesian linear model with multivariate normal prior $\mathcal{P}(\boldsymbol{\beta})$; see Example 3.1 below for more details.

3 MrP Locally Equivalent Weights

We now turn to defining locally equivalent weights for MrP, beginning with the general case and then providing closed-form examples. In the following sections, we show how these weights can be used to construct diagnostics for MrP analogous to those in Section 2.2.2.

3.1 Motivation and definition

To motivate our approach to locally equivalent weights, suppose we want equivalent weights for OLS, w^{OLS} as in Section 2.3.3, but do not have access to the closed-form expression. A practical instance of such a case might be black-box software that computes some linear function of the data, such as a regression tree, but without providing access to the internal parameters of the model. Importantly, suppose we can nevertheless repeatedly call this software to construct the estimate $\hat{\mu}^{\text{OLS}}(\tilde{\boldsymbol{Y}})$ for any $\tilde{\boldsymbol{Y}}$ in a small neighborhood of $\boldsymbol{Y}$. We can then use these black box evaluations to compute w_i^{OLS} via the relation

$$w_i^{\text{OLS}} = N_S \left. \frac{\partial \hat{\mu}^{\text{OLS}}(\tilde{\boldsymbol{Y}})}{\partial \tilde{y}_i} \right|_{\tilde{\boldsymbol{Y}} = \boldsymbol{Y}}.$$

This immediately recovers the OLS weights in Equation (9), noting that the mapping $\boldsymbol{Y} \mapsto \hat{\mu}^{\text{OLS}}(\boldsymbol{Y})$ is linear, and thus the Taylor series expansion of $\hat{\mu}^{\text{OLS}}(\tilde{\boldsymbol{Y}})$ around $\boldsymbol{Y}$ is exact.

Our key idea is to apply this same approach to MrP, even though the mapping $\boldsymbol{Y} \mapsto \hat{\mu}^{\text{MrP}}(\boldsymbol{Y})$ is not linear, arguing that this gives the appropriate notion of a Taylor series approximation to $\hat{\mu}^{\text{MrP}}(\tilde{\boldsymbol{Y}})$ around $\boldsymbol{Y}$.

Definition 3.1 (MrP locally equivalent weights)**.** The *MrP locally equivalent weight* for observation i is

$$w_i^{\text{MrP}} := N_S \frac{\partial \hat{\mu}^{\text{MrP}}(\boldsymbol{Y})}{\partial y_i} = N_S \text{Cov}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})} \left(\nabla_y \ell(y_i | \boldsymbol{x}_i, \boldsymbol{\beta}), g(\boldsymbol{\beta}) \right), \tag{10}$$

where $g(\boldsymbol{\beta}) := \frac{1}{N_T} \sum_{j \in [N_T]} \pi_j m(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)$ is the model's implied target-population mean at a fixed $\boldsymbol{\beta}$, and $\nabla_y \ell(y_i | \boldsymbol{x}_i, \boldsymbol{\beta}) := \partial \ell(y_i | \boldsymbol{x}_i, \boldsymbol{\beta}) / \partial y_i$ denotes the derivative of the log-likelihood contribution with respect to y_i.

The formula for the derivative in Equation (10) follows immediately from well-known results in Bayesian local robustness (e.g. Theorem 1 of Giordano, Broderick, and Jordan, 2018). Importantly, the weights w_i^{MrP} can be easily estimated using MCMC samples from the posterior $\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})$ with minimal additional computation. Further, as we highlight below, for generalized linear models such as logistic regression, this partial derivative simplifies to $\boldsymbol{x}_i^\intercal \boldsymbol{\beta}$, which is also easy to compute from standard model output.

3.2 Justification and interpretation

While the locally equivalent weights are straightforward to define and compute, justifying their use and interpreting their meaning requires more care; technical results follow in the next sections. In short, the weights $\boldsymbol{W}^{\mathrm{MrP}}$ are justified as the unique first-order representation of $\hat{\mu}^{\mathrm{MrP}}$ near $\boldsymbol{Y}$, and this local equivalence is what licenses the CW-style diagnostics developed in Sections 4.1 and 4.2.

First, we must make "local" precise by extending the definition of $\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})$ in Equation (6) to accommodate $\tilde{\boldsymbol{Y}}$ in a Euclidean neighborhood of $\boldsymbol{Y}$. Such $\tilde{\boldsymbol{Y}}$ may not be valid inputs for the log likelihood $\ell(y|\boldsymbol{x},\boldsymbol{\beta})$, but as long as both $\mathcal{P}(\tilde{\boldsymbol{Y}})$ and $\mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\tilde{\boldsymbol{Y}})}\left[m(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)\right]$ are finite for all $\boldsymbol{x}_i$, the mapping $\tilde{\boldsymbol{Y}} \mapsto \hat{\mu}^{\mathrm{MrP}}(\tilde{\boldsymbol{Y}})$ remains well-defined. We assume here that $\hat{\mu}^{\mathrm{MrP}}(\tilde{\boldsymbol{Y}})$ is defined and smooth in a neighborhood of $\boldsymbol{Y}$; we give precise conditions in Sections 4.1 and 4.2 and discuss the binary-response case in Section 6.2.

In general, $\hat{\mu}^{\mathrm{MrP}}$ is not globally linear in $\boldsymbol{Y}$, so we cannot hope that $\hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}) \approx \frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\mathrm{MrP}} y_i$. What we do have is a *locally affine* approximation, the first-order Taylor expansion of $\hat{\mu}^{\mathrm{MrP}}$ around $\boldsymbol{Y}$:

$$\hat{\mu}^{\mathrm{MrPlew}}(\tilde{\boldsymbol{Y}}) := \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}) + \frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\mathrm{MrP}}(\tilde{y}_i - y_i) \tag{11}$$

$$\hat{\mu}^{\mathrm{MrP}}(\tilde{\boldsymbol{Y}}) = \hat{\mu}^{\mathrm{MrPlew}}(\tilde{\boldsymbol{Y}}) + \mathcal{E}(\tilde{\boldsymbol{Y}}, \boldsymbol{Y}), \tag{12}$$

where $\mathcal{E}(\tilde{\boldsymbol{Y}}, \boldsymbol{Y})$ is the appropriate residual. For a globally linear estimator such as $\hat{\mu}^{\mathrm{OLS}}$, the residual vanishes: $\mathcal{E}(\tilde{\boldsymbol{Y}}, \boldsymbol{Y}) = 0$ and $\hat{\mu}^{\mathrm{MrPlew}}(\tilde{\boldsymbol{Y}}) = \hat{\mu}^{\mathrm{OLS}}(\tilde{\boldsymbol{Y}})$, recovering the OLS weights of Equation (9). While $\mathcal{E}(\tilde{\boldsymbol{Y}}, \boldsymbol{Y}) \neq 0$ in general, under regularity conditions established below, we show that $\mathcal{E}(\tilde{\boldsymbol{Y}}, \boldsymbol{Y})$ is of order $\left\|\tilde{\boldsymbol{Y}} - \boldsymbol{Y}\right\|_2^2$. In other words, when $\tilde{\boldsymbol{Y}}$ is close to $\boldsymbol{Y}$, the residuals are small and $\hat{\mu}^{\mathrm{MrPlew}}$ is a locally affine approximation to $\hat{\mu}^{\mathrm{MrP}}$.

Under such an approximation, we define below several senses in which we can justify using the weights $\boldsymbol{W}^{\mathrm{MrP}}$ for diagnostics *as if* $\hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y})$ were a CW estimator with weights $\boldsymbol{W}^{\mathrm{MrP}}$. In Section 4.1, we justify using the weights to compute frequentist variance, leveraging recent results for the Bayesian infinitesimal jackknife. In Section 4.2, we justify using the weights to compute covariate balance: we show that judicious choices of $\tilde{\boldsymbol{Y}}$ produce *nonlinear generalizations* of more standard covariate balance checks and give conditions under which the corresponding nonlinearity is appropriately small.

3.3 Illustration: Closed-form examples

To build intuition, we now derive locally equivalent weights for two closed-form Bayesian examples: a conjugate linear model, and logistic regression in the asymptotic limit. We revisit these examples in the context of covariate balance in Section 4.2.

While illustrative, these closed-form analyses are only possible due to exact (Example 3.1) or asymptotic (Example 3.2) linearity. For more complex models, such as those in our applications in Section 5, no closed form is available, and we instead compute $\boldsymbol{W}^{\mathrm{MrP}}$ directly from Equation (10) using MCMC samples.

Example 3.1. First, consider the conjugate linear model

$$\mathcal{P}(y|\boldsymbol{x},\boldsymbol{\beta}) = \mathcal{N}\left(\boldsymbol{\beta}^\intercal \boldsymbol{x}, \sigma^2\right) \quad \text{and} \quad \mathcal{P}(\boldsymbol{\beta}) = \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right),$$

for known $\sigma \neq 0$ and invertible $\boldsymbol{\Sigma}$. In this case, the posterior has a closed form and $\hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y})$ is linear in $\boldsymbol{Y}$. Building on Gelman (2007), we show in Section A.2 that

$$\boldsymbol{W}^{\mathrm{MrP}} = \frac{1}{N_T}\boldsymbol{X}\left(\frac{1}{N_S}\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\sigma^2}{N_S}\boldsymbol{\Sigma}^{-1}\right)^{-1}\boldsymbol{X}_T^\intercal\boldsymbol{\pi}. \tag{13}$$

For a tight prior, $\boldsymbol{\Sigma} \to \mathbf{0}$, the weights $\boldsymbol{W}^{\mathrm{MrP}}$ shrink and incorporate less of the covariance $\frac{1}{N_S}\boldsymbol{X}^\intercal\boldsymbol{X}$. Conversely, in the flat-prior limit, $\boldsymbol{\Sigma}^{-1} \to \mathbf{0}$ with $\boldsymbol{X}_T = \boldsymbol{X}$, $\boldsymbol{W}^{\mathrm{MrP}}$ reduces to the projection of $(N_S/N_T)\boldsymbol{\pi}$ onto the column space of $\boldsymbol{X}$. □

Example 3.2. Next, consider logistic regression in the asymptotic regime. Suppose that y is binary, with

$$\mathbb{E}_{\mathcal{P}(y|\boldsymbol{x},\boldsymbol{\beta})}[y] = m^{\mathrm{logit}}(\boldsymbol{\beta}^\intercal\boldsymbol{x}) \quad \text{where} \quad m^{\mathrm{logit}}(z) = 1/(1+\exp(-z)).$$

We assume the prior $\mathcal{P}(\boldsymbol{\beta})$ is smooth and N_S is large, so that $\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})$ is well-approximated by Bernstein-von Mises (Van der Vaart, 2000, Ch. 10.2). Write $\hat{\boldsymbol{\beta}}$ for the MLE, $\hat{y}_i := m^{\mathrm{logit}}(\hat{\boldsymbol{\beta}}^\intercal\boldsymbol{x}_i)$ and $\hat{v}_i := \hat{y}_i(1-\hat{y}_i)$ for the approximate mean and variance, with analogous definitions for the target $(\hat{y}_j, \hat{v}_j)$. Let $\boldsymbol{V}$ and $\boldsymbol{V}_T$ denote diagonal matrices with survey and target variance estimates on the diagonal, respectively. As we show in Section A.3, in this case

$$\boldsymbol{W}^{\mathrm{MrP}} \approx \frac{1}{N_T}\boldsymbol{X}\left(\frac{1}{N_S}\boldsymbol{X}^\intercal\boldsymbol{V}\boldsymbol{X}\right)^{-1}\boldsymbol{X}_T^\intercal\boldsymbol{V}_T\boldsymbol{\pi}, \tag{14}$$

where $\approx$ reflects the delta method and Bernstein-von Mises approximations.

Equation (14) mirrors the flat-prior limit of Equation (13) ($\boldsymbol{\Sigma}^{-1} = \mathbf{0}$), with extra factors of $\boldsymbol{V}$ and $\boldsymbol{V}_T$. The absence of the prior covariance reflects the fact that Equation (14) is computed in the asymptotic limit and thus any prior influence vanishes. The variance factors $\boldsymbol{V}$ and $\boldsymbol{V}_T$ reflect the fact that $\hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y})$ is nonlinear in $\boldsymbol{Y}$, since different $\boldsymbol{Y}$ gives rise to different $\boldsymbol{V}$ and so different $\boldsymbol{W}^{\mathrm{MrP}}$. □

4 Formal results

We now formally justify the use of MrPlew weights for estimating frequentist variability and for assessing covariate balance.

4.1 Variance estimation

We begin with frequentist variability. As we will see, this is conceptually distinct from Bayesian posterior uncertainty, which captures uncertainty about the parameter $\boldsymbol{\beta}$ given the data. The frequentist variance, by contrast, captures the variability of the estimator across repeated samples.

Our proof will rely on some technical assumptions required for consistency of the Bayesian infinitesimal jackknife (Giordano and Broderick, 2024). To state these conditions, we need to introduce some notation. Let $\nabla_\eta^k A$ denote the k-th derivative of the log partition function $\mathcal{A}(\cdot)$ of an exponential family, and let $r^{\otimes k}$ denote the $D_r \times \ldots D_r$ array of products of r. Let $\|\cdot\|_2^2$ of a multidimensional array denote the squared Euclidean norm of the stacked array, i.e., the sum of the squares of the array entries. Finally, let $\overset{dist}{\to}$ denote convergence in distribution and $\overset{prob}{\to}$ convergence in probability, both with respect to IID samples from $\mathcal{P}_S(\boldsymbol{x}, y)$.

Assumption 4.1 (Canonical exponential family)**.** Assume that the likelihood is given by a one-parameter natural exponential family with sufficient statistic y and natural parameter $\eta = \boldsymbol{\beta}^\intercal \boldsymbol{x}$. (This model may be misspecified.) Specifically, the log likelihood for $(\boldsymbol{x}_i, y_i)$ is given by

$$\ell(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = y_i\eta_i - A(\eta_i) \text{ for } \eta_i := \boldsymbol{\beta}^\intercal \boldsymbol{x}_i,$$

where $A(\cdot)$ is the log partition function and the density is assumed to be relative to some fixed base measure on y. □

Popular models satisfying this restriction are generalized linear models with canonical link functions. We primarily focus on logistic regression; other examples include Poisson regression and Normal regression with known residual variance.

Next, we assume that the limit of the maximum likelihood estimator of $\boldsymbol{\beta}$ exists and is identifiable as $N_S \to \infty$.

Assumption 4.2 (Identifiability of the MLE)**.** Assume that $\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\mathcal{A}(\boldsymbol{\beta}^\intercal \boldsymbol{x})\right]$ is finite for all finite $\boldsymbol{\beta}$, and define $\ell(\boldsymbol{\beta}) := \mathbb{E}_{\mathcal{P}_S(y,\boldsymbol{x})}\left[\ell(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})\right]$. Assume that $\boldsymbol{\beta}^* := \text{argmax}_{\boldsymbol{\beta}\in\mathbb{R}^{D_{\boldsymbol{\beta}}}} \ell(\boldsymbol{\beta})$ exists, is unique, and that $\mathcal{I} := -\left.\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\intercal}\right|_{\boldsymbol{\beta}^*}$ is positive definite. □

If we can exchange integration and differentiation in the definition of $\ell(\boldsymbol{\beta})$ (sufficient conditions are given in Assumption B.1 and Lemma B.1 in Appendix B), then the information matrix from Assumption 4.2 takes the form

$$\mathcal{I} = \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\nabla_\eta^2 A(\boldsymbol{\beta}^{*\intercal}\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\intercal\right].$$

Since $\nabla_\eta^2 A(\boldsymbol{\beta}^{*\intercal}\boldsymbol{x}) > 0$ by standard properties of exponential families, positive definiteness of $\mathcal{I}$ reduces to a mild moment condition on $\boldsymbol{x}$. This positive-definiteness condition fails if $\boldsymbol{x}$ lies almost surely in a proper linear subspace of $\mathbb{R}^{D_r}$, but holds whenever the support of $\boldsymbol{x}$ spans $\mathbb{R}^{D_r}$.

We impose a mild set of conditions on the survey data-generating distribution $\mathcal{P}_S(\boldsymbol{x}, y)$ and the prior $\mathcal{P}(\boldsymbol{\beta})$.

Assumption 4.3 (Regularity for the infinitesimal jackknife)**.** Under Assumptions 4.1 and 4.2, assume that as $N_S \to \infty$ (with observations IID from $\mathcal{P}_S(\boldsymbol{x}, y)$), the following stay fixed:

- The dimension of $\boldsymbol{\beta}$,
- The target observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_T}\}$, and
- The weighting function $\pi(\cdot)$.

Additionally assume that:

- With probability one under $\mathcal{P}_S(\boldsymbol{x})$, $\boldsymbol{x}$ is bounded.
- The prior $\mathcal{P}(\boldsymbol{\beta})$ has bounded support.
- Marginally, $\mathbb{E}_{\mathcal{P}_S(y)}\left[y^2\right] < \infty$.
- The prior $\mathcal{P}(\boldsymbol{\beta})$ is proper and has a density that is nonzero and four times continuously differentiable in a neighborhood of $\boldsymbol{\beta}^*$.

The boundedness assumptions on $\boldsymbol{x}$ and $\boldsymbol{\beta}$ can be relaxed to weaker moment conditions; see Assumption B.1 in Appendix B. □

We now state the main result that a sample-variance estimator $\hat{V}$, analogous to the CW variance formula in Equation (4), is consistent for the frequentist variance of $\hat{\mu}^{\mathrm{MrP}}$.

Theorem 4.1 (Infinitesimal jackknife-based variance for MrPlew)**.** *Let Assumptions 4.1 to 4.3 hold. For $i \in [N_S]$, let $\hat{y}_i := \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left[m(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)\right]$ and $\varepsilon_i := (y_i - \hat{y}_i)$. Define*

$$\hat{V} := \frac{1}{N_S} \sum_{i \in [N_S]} \left(N_S w_i^{\mathrm{MrP}} \varepsilon_i - N_S \overline{w^{\mathrm{MrP}} \varepsilon} \right)^2 \quad \textit{where} \quad \overline{w^{\mathrm{MrP}} \varepsilon} := \frac{1}{N_S} \sum_{i \in [N_S]} w_i^{\mathrm{MrP}} \varepsilon_i, \tag{15}$$

and where $\hat{V}$ is the sample variance of $N_S w_i^{\mathrm{MrP}} \varepsilon_i$. Then, as $N_S \to \infty$,

$$\sqrt{N_S}\left(\hat{\mu}^{\mathrm{MrP}} - \hat{\mu}^{\infty}\right) \overset{dist}{\to} \mathcal{N}(0, V) \quad \textit{and} \quad \hat{V} \overset{prob}{\to} V$$

for some variance $V \geq 0$ and $\hat{\mu}^{\infty} = \frac{1}{N_T} \sum_{j \in [N_T]} \pi_j m(\boldsymbol{\beta}^{*\intercal} \boldsymbol{x}_j)$.

See Appendix B for the proof.

The proof of Theorem 4.1 operates by showing that $\hat{V}$ is asymptotically equivalent to the infinitesimal jackknife covariance; Giordano and Broderick (2024, Theorem 2) show this is consistent. The bulk of the proof verifies that the model satisfies the conditions of the IJ consistency theorem under Assumptions 4.1 to 4.3.

Remark 4.1 (Failure of variance estimation when the canonical exponential family does not hold)**.** Assumption 4.1 is essential for Theorem 4.1, though not for Theorem 4.2 below: if the model's sufficient statistic is not y alone, then $\hat{V}$ is generally an *inconsistent* estimator of the frequentist variance. For example, if

we model $\mathcal{P}(y|\boldsymbol{x},\boldsymbol{\beta},\sigma) = \mathcal{N}\left(\boldsymbol{x}^\intercal\boldsymbol{\beta},\sigma^2\right)$ with unknown variance σ^2, the sufficient statistic is (y, y^2), Assumption 4.1 is not satisfied, and $\hat{V}$ is inconsistent; see Example A.1 in Appendix A for further discussion of this counterexample. This is not a problem for variance estimation itself, since one can use the infinitesimal jackknife covariance instead. However, this counterexample emphasizes that $\boldsymbol{W}^{\text{MrP}}$ does not automatically inherit CW-style interpretations for variance.

4.2 Covariate balance

4.2.1 Motivation and intuition

We now turn to covariate balance. The key idea is to reframe the standard balance check as a sensitivity diagnostic that assesses whether the weighting procedure would have been able to detect a small, directed perturbation of the response variable. This reframing admits a local interpretation to which we can apply the Taylor series approximation in Equation (12), justifying the use of $\boldsymbol{W}^{\text{MrP}}$ in a local analogue of the standard CW balance check.

To build intuition, consider a perturbation of a continuous response y of the form:

$$\tilde{y} = y + \delta r(\boldsymbol{x}) \tag{16}$$

for some measurable function of the covariates $r = r(\boldsymbol{x})$ and small δ. For continuous y, the additive construction in eq. (16) can produce valid observations. For binary y or other bounded responses, the additive construction of eq. (16) no longer produces a valid response in general. We will *define and analyze* $\tilde{y}$ of eq. (16) through the generalized posterior below, deferring the question of which binary processes approximately satisfy eq. (21) to Section 4.2.4.

The corresponding change in the (unknown) target mean is:

$$\frac{1}{N_T}\sum_{j\in[N_T]}\pi_j\tilde{y}_j - \frac{1}{N_T}\sum_{j\in[N_T]}\pi_j y_j = \underbrace{\delta\frac{1}{N_T}\sum_{j\in[N_T]}\pi_j r_j}_{\text{actual change}}.$$

For linear calibration weights, the corresponding change in the estimator is:

$$\frac{1}{N_S}\sum_{i\in[N_S]}w_i^{\text{CW}}\tilde{y}_i - \frac{1}{N_S}\sum_{i\in[N_S]}w_i^{\text{CW}}y_i = \underbrace{\delta\frac{1}{N_S}\sum_{i\in[N_S]}w_i^{\text{CW}}r_i}_{\text{inferred change}}. \tag{17}$$

Subtracting the inferred change from the actual change and dividing by δ immediately recovers the standard covariate balance check in Equation (3):

$$\frac{1}{\delta}\text{Imbalance}(r,\boldsymbol{W}^{\text{CW}}) = \frac{1}{N_T}\sum_{j\in[N_T]}\pi_j r_j - \frac{1}{N_S}\sum_{i\in[N_S]}w_i^{\text{CW}}r_i \overset{\text{check}}{\approx} 0. \tag{18}$$

The main advantage of this reframing is that we can immediately extend this argument to nonlinear MrP

via its Taylor expansion. Specifically, as long as we can sensibly define $\hat{\mu}^{\mathrm{MrP}}(\tilde{\boldsymbol{Y}})$ and the Taylor expansion in Equation (12) is valid, we have an analogous *local* change:

$$\begin{aligned}\hat{\mu}^{\mathrm{MrP}}(\tilde{\boldsymbol{Y}}) - \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}) &= \frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\mathrm{MrP}}(\tilde{y}_i - y_i) + \mathcal{E}(\tilde{\boldsymbol{Y}}, \boldsymbol{Y}) \\ &= \underbrace{\delta\frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\mathrm{MrP}} r_i}_{\text{inferred change}} + \mathcal{E}(\tilde{\boldsymbol{Y}}, \boldsymbol{Y}). \end{aligned} \tag{19}$$

Theorem 4.2 below shows that, under regularity conditions, the residual term $\mathcal{E}(\tilde{\boldsymbol{Y}}, \boldsymbol{Y})$ is of order δ^2, uniformly over a Donsker class of functions r. After dividing by δ and dropping the residual term, we can therefore form a *local* balance check that replaces w^{CW} with w^{MrP} but otherwise mirrors the linear version.

4.2.2 Main result

Definition 4.1 (Perturbed likelihood and generalized posterior)**.** Let $\delta_+ > 0$ denote an upper bound on the perturbation. Under Assumption 4.1, for all $\delta \in [0, \delta_+]$, we formally define the perturbed likelihood

$$\begin{aligned}\ell(y|\boldsymbol{x}, \boldsymbol{\beta}; \delta r) &= (y + \delta r(\boldsymbol{x}))\boldsymbol{x}^\intercal\boldsymbol{\beta} - A(\boldsymbol{\beta}^\intercal\boldsymbol{x}) \\ &= \ell(y|\boldsymbol{x}, \boldsymbol{\beta}) + \delta r(\boldsymbol{x})\boldsymbol{x}^\intercal\boldsymbol{\beta} \\ \mathcal{P}(\boldsymbol{Y}|\boldsymbol{\beta}; \delta r) &:= \exp\left(\sum_{i\in[N_S]} \ell(y_i + \delta r_i|\boldsymbol{x}_i, \boldsymbol{\beta})\right).\end{aligned}$$

We then define the *generalized posterior*

$$\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y}; \delta r) := \frac{\mathcal{P}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}; \delta r)\mathcal{P}(\boldsymbol{\beta})}{\int \mathcal{P}(\boldsymbol{Y}|\boldsymbol{X}, \tilde{\boldsymbol{\beta}}; \delta r)\mathcal{P}(\tilde{\boldsymbol{\beta}})d\tilde{\boldsymbol{\beta}}}$$

when the denominator is finite, and let $\hat{\mu}^{\mathrm{MrP}}(\tilde{\boldsymbol{Y}}) := \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\delta r)}[g(\boldsymbol{\beta})]$ ($g(\boldsymbol{\beta})$ is defined in eq. (10)).

Assumption 4.4 (Donsker class)**.** Let $\mathcal{R}$ denote a Donsker class of $\mathcal{P}_S(\boldsymbol{x})$-measurable functions for which $\sup_{r\in\mathcal{R}} \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\|\boldsymbol{x}r(\boldsymbol{x})\|_2^2\right] < \mathcal{R}_{\max}^2 < \infty$. □

Donsker classes are classes of functions restricted enough to obey uniform laws of large numbers (Van der Vaart, 2000, Chapter 19). Readers less familiar with Donsker classes can instead imagine that $\mathcal{R}$ is a finite set of functions without losing any essential understanding.

Theorem 4.2. *Take $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + \delta\boldsymbol{R}$, where $\boldsymbol{R} = (r(\boldsymbol{x}_1), \ldots, r(\boldsymbol{x}_{N_S}))^\intercal$. Under Assumptions 4.1 to 4.4, with probability approaching one as $N_S \to \infty$, $\hat{\mu}^{\mathrm{MrP}}(\tilde{\boldsymbol{Y}})$ exists, and satisfies*

$$\sup_{r\in\mathcal{R}} \frac{1}{\delta}\left|\hat{\mu}^{\mathrm{MrP}}(\tilde{\boldsymbol{Y}}) - \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}) - \frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\mathrm{MrP}} r_i\right| = O(\delta) \quad \text{as } \delta \to 0.$$

See Appendix C for a proof. Theorem 4.2 shows that, to leading order in δ, the change in the perturbed MrP estimator per unit δ is $\frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\text{MrP}} r_i$. The resulting MrP balance check Imbalance$(r, \boldsymbol{W}^{\text{MrP}})$ is thus the natural analogue of Equation (18). The fact that Theorem 4.2 holds uniformly over a wide class of functions $r(\cdot)$ justifies searching over a set of prespecified covariates to check for imbalance, as is commonly done in practice.

Unlike Theorem 4.1, the assumption in Assumption 4.1 that y is a sufficient statistic of the model is not essential for Theorem 4.2, and an inspection of the proof of Lemma C.3 in Appendix C will show that versions of Theorem 4.2 should hold for a much broader class of models, including generic multivariate exponential families. As with Theorem 4.1, the boundedness conditions in Assumption 4.3 can be relaxed to a set of technical moment conditions (see Assumption B.1 in Appendix B).

4.2.3 Closed-form examples

We next apply the balance check of Theorem 4.2 to the two closed-form Bayesian examples from Section 4.2.3.

Example 4.1 (Conjugate normal models)**.** We continue Example 3.1, the conjugate normal model, focusing on the consequences of prior shrinkage for covariate balance; Section A.2 gives full details. Applying Theorem 4.2 and taking $r(\boldsymbol{x})$ to be the components of $\boldsymbol{x}$, we have the following imbalance:

$$\frac{1}{\delta}\text{Imbalance}(\boldsymbol{x}, \boldsymbol{W}^{\text{MrP}}) = \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T \left(\boldsymbol{I}_P - \left(\frac{1}{N_S}\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\sigma^2}{N_S}\boldsymbol{\Sigma}^{-1} \right)^{-1} \left(\frac{1}{N_S}\boldsymbol{X}^\intercal\boldsymbol{X} \right) \right) \tag{20}$$

When $\boldsymbol{\Sigma}^{-1} \neq \boldsymbol{0}$, $\hat{\mu}^{\text{MrP}}$ only *approximately* balances $\boldsymbol{x}$. As $\boldsymbol{\Sigma}^{-1} \to \boldsymbol{0}$, however, the matrix product inside eq. (20) converges to $\boldsymbol{I}_P$, so the imbalance converges to zero. This reflects existing results on the implied imbalance of ridge regression; see, for example, Bruns-Smith et al. (2025).

This setup is formally equivalent to *soft calibration* (Gao, Yang, and Kim, 2022), where $\boldsymbol{\beta}$ is treated as a random effect with distribution $\mathcal{P}(\boldsymbol{\beta})$. Indeed, as with soft calibration, we show in section A.2 that $\boldsymbol{W}^{\text{MrP}}$ minimizes the expected mean squared error of $\frac{1}{N_T}\boldsymbol{\pi}^\intercal\boldsymbol{Y}_T - \frac{1}{N_S}\boldsymbol{W}^\intercal\boldsymbol{Y}$ when $\boldsymbol{Y}$ and $\boldsymbol{Y}_T$ are generated from a shared draw of $\boldsymbol{\beta} \sim \mathcal{P}(\boldsymbol{\beta})$. □

Example 4.2 (Asymptotic logistic regression)**.** We continue Example 3.2, the asymptotic logistic regression setting; Section A.3 gives full details. Importantly, we show that logistic regression generally balances the *variance-weighted* covariates, but not the covariates themselves. This mirrors the classical result for model-assisted generalized regression estimation under a logistic link (Firth and Bennett, 1998). Consider variance-weighted regressors $r(\boldsymbol{x}) = v(\boldsymbol{x})\boldsymbol{x}$, where $v(\boldsymbol{x}) = \text{Var}_{\mathcal{P}(y|\boldsymbol{x})}(y)$ is the conditional variance of y, estimated by $\hat{v}_i$ at $\boldsymbol{x}_i$. Then:

$$\frac{1}{\delta}\text{Imbalance}(v(\boldsymbol{x})\boldsymbol{x}, \boldsymbol{W}^{\text{MrP}}) = \boldsymbol{0}.$$

By contrast, logistic regression does not balance the covariates $\boldsymbol{x}$ themselves:

$$\frac{1}{\delta}\text{Imbalance}(\boldsymbol{x}, \boldsymbol{W}^{\text{MrP}}) = \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T \left(\boldsymbol{I}_P - \left(\frac{1}{N_S}\boldsymbol{X}^\intercal\boldsymbol{V}\boldsymbol{X} \right)^{-1} \frac{1}{N_S}\boldsymbol{X}^\intercal\boldsymbol{X} \right) \neq \boldsymbol{0}. \text{ (in general)}$$

Unlike the conjugate normal model above, this imbalance is not due to prior influence, which vanishes in the asymptotic regime. Rather, it arises from a mismatch in the scale of the model and the perturbation: Equation (21) defines perturbations on the y scale, whereas logistic regression operates on the log-odds scale.

Somewhat surprisingly, although it is the presence of $\boldsymbol{V}$ that causes the imbalance, this imbalance is not fundamentally driven by nonlinearity of the map $\hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y})$ as a function of $\boldsymbol{Y}$. In Section A.4, we consider the special case in which the density ratio is linear in $\boldsymbol{x}$ and show that the logistic regression MrP estimator is then approximately linear in $\boldsymbol{Y}$ for large N_S. Even in this case, however, the imbalance of $\boldsymbol{x}$ remains nonzero. □

4.2.4 Using a parametric bootstrap to produce restricted responses

As we discuss above, when the response y is binary or otherwise bounded, the additive construction in eq. (16) cannot be valid in general, and yet it forms the basis for our theoretical results in Theorem 4.2. In this section, we bridge the gap and discuss how to generate valid response vectors that *approximately reproduce* imbalance identified for the continuously perturbed $\tilde{y}$.

The key idea is that the perturbation eq. (16) can be a good approximation to a restricted random variable $\breve{y}$ with *conditional expectation* shifted relative to y:

$$\mathbb{E}_{\mathcal{P}(\breve{y}|\boldsymbol{x})}[\breve{y}] = \mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[y] + \delta r(\boldsymbol{x}). \tag{21}$$

Let $\breve{\boldsymbol{Y}}$ denote the corresponding vector of perturbed responses. Equation (21) makes sense since, for any model satisfying Assumption 4.1, the posterior $\mathcal{P}(\boldsymbol{\beta}|\breve{\boldsymbol{Y}})$ depends on $\breve{\boldsymbol{Y}}$ only through

$$\boldsymbol{\beta} \mapsto \boldsymbol{\beta}^\intercal \frac{1}{N_S} \sum_{i \in [N_S]} y_i \boldsymbol{x}_i \approx \boldsymbol{\beta}^\intercal \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[y]\ \boldsymbol{x}\right] \text{ for large } N_S.$$

So if $\breve{y}$ satisfies eq. (21) then we can expect $\mathcal{P}(\boldsymbol{\beta}|\breve{\boldsymbol{Y}}) \approx \mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y}; \tilde{\boldsymbol{Y}})$, and we can thus think of Theorem 4.2 as approximating the behavior of $\mathcal{P}(\boldsymbol{\beta}|\breve{\boldsymbol{Y}})$.

We briefly describe a simple procedure based on a perturbed parametric bootstrap to generate $\breve{\boldsymbol{Y}}$ satisfying eq. (21), while still being reasonably close to $\boldsymbol{Y}$; Appendix D gives the full details. We begin with an estimate $\hat{m}(\boldsymbol{x}_i)$ of $\mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[y]$, as one would for performing the parametric bootstrap (Efron and Tibshirani, 1994). We then draw $\breve{y}$ with mean $\hat{m}(\boldsymbol{x}_i) + \delta r(\boldsymbol{x}_i)$, which requires restricting δ small enough that $\hat{m}(\boldsymbol{x}_i) + \delta r(\boldsymbol{x}_i) \in [0, 1]$ for all i. The final step is to correlate $\breve{y}_i$ with y_i while maintaining the desired marginals, in order to keep $\breve{\boldsymbol{Y}} - \boldsymbol{Y}$ as small as possible; a procedure for doing so is described in Appendix D. In section 6.2 we apply this technique to our experiments and find that the resulting binary vectors match the corresponding predicted results closely.

4.2.5 Subgroup contribution

An important special case of the general perturbation results is for measuring the contribution of particular subgroups to the final estimate. Consider the Name Change application from Section 1.1, where we group respondents by education level: less than a college degree, a college degree, or more than a college degree.

If we index these groups from $s = 1, \ldots, S$ (here, $S = 3$), and let $\mathcal{I}_s$ denote the set of indices i that are from subgroup s, then we can rewrite any CW estimator as

$$\hat{\mu}^{\mathrm{CW}}(\boldsymbol{Y}) = \frac{1}{N_S} \sum_{s=1}^{S} \sum_{i \in \mathcal{I}_s} w_i^{\mathrm{CW}} y_i.$$

Here, $w_s^{\mathrm{CW}} := \sum_{i \in \mathcal{I}_s} w_i^{\mathrm{CW}}$ is the total weight given to subgroup s, and measures how much subgroup s contributes to the overall estimate. Comparing how w_s varies from subgroup to subgroup can provide intuition to the analyst about how the data is being used.

For MrPlew weights, the quantity $w_s^{\mathrm{MrP}} := \sum_{i \in \mathcal{I}_s} w_i^{\mathrm{MrP}}$ is precisely the left-hand side of the covariate balance check for the indicator $\boldsymbol{z}_{is} = \mathbb{I}(\boldsymbol{x}_i \text{ is in subgroup } s)$. Here, $\boldsymbol{z}_{is}$ takes value 1 if the survey observation i is from subgroup s, and 0 otherwise. Thus, w_s^{MrP} admits an interpretation similar to that of covariate balance: if the responses in subgroup s were all to increase in expectation by a small δ, we would expect $\hat{\mu}^{\mathrm{MrP}}$ to increase by $\delta w_s^{\mathrm{MrP}}/N_S$. This provides an intuitively meaningful measure of the "importance" of subgroup s in the estimator $\hat{\mu}^{\mathrm{MrP}}$. This interpretation is supported by Theorem 4.2 without modification, simply by taking $r(\boldsymbol{x})$ to be the indicator of the subgroup categories.

4.2.6 Why not perturb the log odds?

A key feature of our generalized balance checks is that the perturbation to the response in Equations (16) and (21) is defined in the space of responses, y, rather than in the space of log-odds, $m^{\mathrm{logit}^{-1}}(\mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[y])$. Before proceeding, we briefly motivate and discuss this decision.

As a motivating question, in light of the failure of logistic regression to balance the regressors as shown in Example 4.2, one might wonder why we do not use the following perturbation rather than eq. (21) to define our perturbations to the data:

$$\mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[\tilde{y}] = m^{\mathrm{logit}}\left(m^{\mathrm{logit}^{-1}}(\mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[y]) + \delta' r(\boldsymbol{x})\right). \tag{22}$$

Since

$$\frac{\partial}{\partial \delta'} \mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[\tilde{y}] = v(\boldsymbol{x}) r(\boldsymbol{x}),$$

adding a small $r(\boldsymbol{x})$ to the log odds is equivalent to adding a small $v(\boldsymbol{x})r(\boldsymbol{x})$ to the expectation. Since the logistic regression models the log odds as a linear combination of $\boldsymbol{x}$, it is perfectly able to capture perturbations of the log odds in the direction of $\boldsymbol{x}$, and such directions correspond to perturbations of the form $v(\boldsymbol{x})\boldsymbol{x}$ in the space of expectations.

We might argue that eq. (21) is easier for a practitioner to think about intuitively. But a more fundamental reason is that the perturbation eq. (22) is *defined in terms of a particular model*, a fact which makes it difficult to meaningfully compute and compare perturbations in a model-agnostic way. For one, in order to actually induce the perturbation eq. (22), one needs to have an estimate of $v(\boldsymbol{x})$ that can only be provided by one of the very models we are trying to interrogate. Different logistic regression models, with different priors, will define different perturbations to the data. Also, the perturbation eq. (22) privileges the (fairly arbitary) log-odds parameterization when defining the data perturbation. Other link functions, such as the normal distribution

function for probit regression will exhibit unfavorable performance under the perturbation eq. (22). The perturbation eq. (22) appears to preclude direct comparison between methods that are not guaranteed to produce valid log-odds estimates, such as OLS and raking. Thus, we believe that the failure of logistic regression to balance $\boldsymbol{x}$ asymptotically should be taken as a warning about logistic regression, not a failure of our balance checks.

5 Applying MrPlew

We demonstrate how to use MrPlew for three existing MrP analyses: (1) an analysis that extrapolates a Twitter survey of name changes after marriage to the entire US population (Alexander, 2019; Cohen, 2019); (2) an example of correcting sampling bias in a national survey on support for same-sex marriage (Lax and Phillips, 2009; Kastellec, Lax, and Phillips, 2010); and (3) a textbook example analyzing the 2020 US presidential election (Alexander, 2023).

In each case, we compare the original MrP analysis to raking on marginals of coarsened versions of the same regressors used in the MrP analysis (`survey::calibrate` from Lumley (2024)). Following DeBell and Krosnick (2009), the raking covariates were coarsened so that no marginal category contains fewer than 5% of the survey observations. For example, when MrP regressions included US state as a regressor, state indicators were coarsened to geographic region (west, south, northeast, and midwest) for raking. Finally, for simplicity we removed the small number of observations with missing regressor values.[1]

5.1 Application descriptions

We first describe the two additional datasets and analyses that we reproduce; Section 1.1 above gives additional details for the "Name Change" analysis.

"Same-Sex Marriage" analysis. Our next analysis is based on the classic MrP primer from Kastellec, Lax, and Phillips (2010), which provides code and data that is amenable to re-analysis with MCMC (Kastellec, 2024). Kastellec, Lax, and Phillips (2010) analyze five consistently-coded national polls from 2004 that surveyed support for same-sex marriage; the authors call these polls a "megapoll." The response variable y is a binary indicator where 1 encodes support for same-sex marriage, and 0 encodes either opposition or no expressed opinion; see the appendix of Lax and Phillips (2009) for more details.

Following Kastellec, Lax, and Phillips (2010), we then use MrP to calibrate the megapoll responses to individual states using a 5% Public Use Microdata Sample (PUMS) from the 2000 US census. We fit an MCMC version of the following model, which the original paper fit using marginal maximum likelihood:

$$y \sim \texttt{logit}\Big(\texttt{(1 | race.female) + (1 | age.cat) + (1 | edu.cat) + (1 | age.edu.cat) +}$$
$$\texttt{(1 | state) + (1 | region) + (1 | poll) + p.relig.full + p.kerry.full}\Big)$$

See Kastellec, Lax, and Phillips (2010), as well as the original research paper from Lax and Phillips (2009)

[1] MrPlew weights remain computable under Bayesian data imputation, provided the posterior covariance in Equation (10) incorporates the additional variability due to imputation in the $\nabla_y \ell(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})$ term.

	Name Change	Same-Sex Marriage	Election Forecasting
N_S	$4,364$	$6,341$	$4,803$
$\overline{y}$	0.462	0.333	0.539
$\hat{\mu}^{\text{MrP}}$	0.288	0.457	0.522
$\hat{\mu}^{\text{CW}}$	0.263	0.345	0.522

Table 1: High-level descriptions of the three applications.

for additional context for the surveys and regressor definitions. Following the "MrP Primer" chapter of Mastny (2018), which re-analyzes the same data using `brms`, we set standard Gaussian priors for each of the fixed effects and the standard deviations.[2]

Our main illustration uses MrP to predict support for same-sex marriage in California, which is an example of using MrP for small area estimation.[3] In Section 5.5, we also predict support in Missouri, as a contrast to the estimate for California. To estimate raking weights for California, we eliminated or coarsened interactions that occurred less than 5% of the time in either the survey or target population. In the end, we used the following for raking: gender, education level, age category, race (white, black, or other), white / non-white interacted with gender, and age category interacted with secondary / no secondary education.

"Election Forecasting" analysis. Our third illustration is based on an analysis of the 2020 US presidential election taken from Alexander (2023, Ch. 6, 8, and 16). The survey dataset is from the Nationscape project (Tausanovitch and Vavreck, 2021), which combines a large number of surveys conducted between July 2019 and January 2021. The response variable y is a binary indicator where 1 encodes support for Joe Biden in the 2020 US presidential election, and 0 encodes support for Donald Trump.

Our objective with MrP is to adjust for the fact that the original survey is a convenience sample that is potentially unrepresentative of the entire US population. The population dataset is taken from the 2019 American Community Survey (ACS) dataset, accessed through IPUMS (Ruggles et al., 2024), and is selected as a proxy for the demographic profile of the entire US population.

The regressors are gender (encoded as male or female), four age groups in roughly 15-year bins, three levels of education, and state. We fit the following hierarchical logistic regression:

$$y \sim \texttt{logit}\Big(\texttt{ gender + (1 | age\_group) + (1 | state) + (1 | education\_level) }\Big),$$

with normal priors for the scale and intercept terms. For raking, we coarsened the states to regions as in the Name Change analysis above.

5.2 Comparing MrPlew and raking weights

Table 1 gives a high-level summary of the three main applications, which exhibit different relationships between raking, MrP, and the uncorrected survey mean $\overline{y}$. In the Election Forecasting application, the survey mean, raking, and MrP are all similar; in the Name Change application, raking and MrP are similar

[2] In `brms`, which follows conventions from the `stan` software package, the prior is a truncated half-normal for the standard deviations.

[3] When forming predictions, we set the `poll` random effect to zero.

to one another but quite different from the survey mean; and in the Same-Sex Marriage application, raking and the survey mean are similar, but MrP is very different.

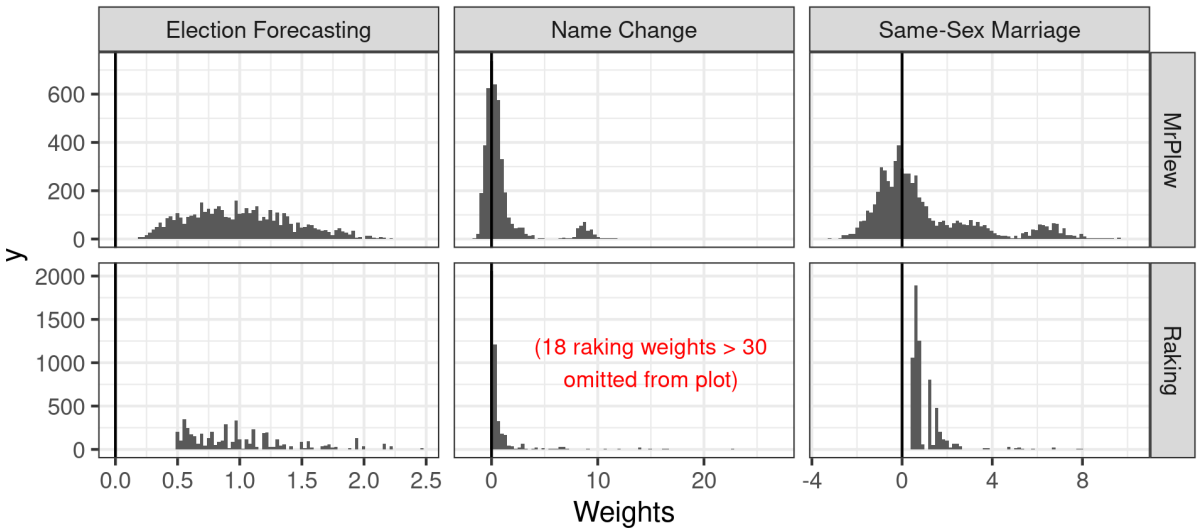


Figure 3: Comparison of weights

Figure 3 shows the corresponding MrPlew and raking weights. A qualitative inspection of the weights largely aligns with the comparisons across point estimates. In the Election Forecasting application, the MrP and raking weights are broadly similar. By contrast, the MrP and raking weights display very different patterns in the other two applications. Using the tools developed above, we will assess whether these differences are meaningful.

5.3 Frequentist variance

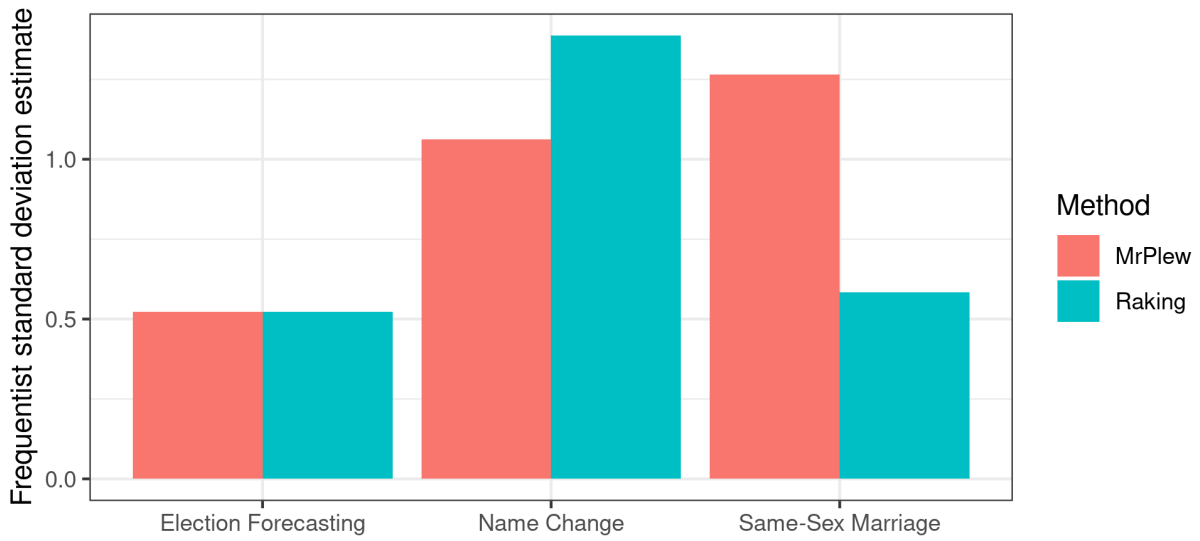


Figure 4: Estimates of the frequentist standard deviation of $\sqrt{N_S}\hat{\mu}^{\mathrm{MrP}}$ and $\sqrt{N_S}\hat{\mu}^{\mathrm{CW}}$

The weights in Figure 3 suggest that the frequentist variance of MrP and raking may be similar in the Election Forecasting example, and that the MrP variance may be higher than the raking variance in the Same-Sex Marriage example; for the Name Change application, this is harder to assess visually given the extreme outliers for raking.

Figure 4 shows the frequentist variance estimates for MrP from Theorem 4.1, compared to the corresponding raking-based variance estimate, which we compute as:

$$\hat{V}_{\mathrm{CW}} := \frac{1}{N_S} \sum_{i \in [N_S]} \left(N_S w_i^{\mathrm{CW}} \varepsilon_i - N_S \overline{w^{\mathrm{CW}} \varepsilon} \right)^2 \quad \text{where} \quad \overline{w^{\mathrm{CW}} \varepsilon} := \frac{1}{N_S} \sum_{i \in [N_S]} w_i^{\mathrm{CW}} \varepsilon_i,$$

for $\varepsilon_i = (y_i - \hat{y}_i)$ defined in Theorem 4.1 and used in our estimate of $\hat{V}$.

Importantly, the frequentist standard error estimates in Figure 4 enable an apples-to-apples comparison between MrP and raking, since these estimates capture uncertainty under repeated sampling for both estimators. More generally, frequentist variances and Bayesian posterior variances only coincide asymptotically under posterior concentration and correct specification; otherwise, these two quantities generally differ, especially in the presence of weakly estimated random effects (Kleijn and van der Vaart, 2012; Giordano and Broderick, 2024).

Figure 4 shows three distinct patterns. For the Election Forecasting application, the estimated variances are quite close between MrP and raking, consistent with the qualitative inspection of the weights in Figure 3. For the Name Change application, by contrast, the extreme raking weights lead to higher variance for raking than MrP, likely due to the fact that MrP uses more information than raking, and so is able to produce a greater variety of weights. Finally, for the Same-Sex Marriage application, we see the opposite pattern, with nearly triple the frequentist standard deviation for MrP versus raking.

Finally, Appendix E assesses the accuracy of the MrPlew variance estimates by bootstrapping the MCMC procedure (Huggins and Miller, 2023; Giordano and Broderick, 2024). We consider parametric and nonparametric bootstraps; both confirm that the standard deviation of $\hat{\mu}^{\text{MrP}}$ across bootstrap draws closely matches the $\hat{V}$ estimates.

5.4 Covariate balance

Next, we use the results in Section 4.2 to examine differences in the implied covariate balance between MrPlew and raking weights. For each application, we computed covariate balance for each regressor used in raking, which are all binary or discrete, as well as all two-way interactions of those regressors. Appendix E gives complete results.

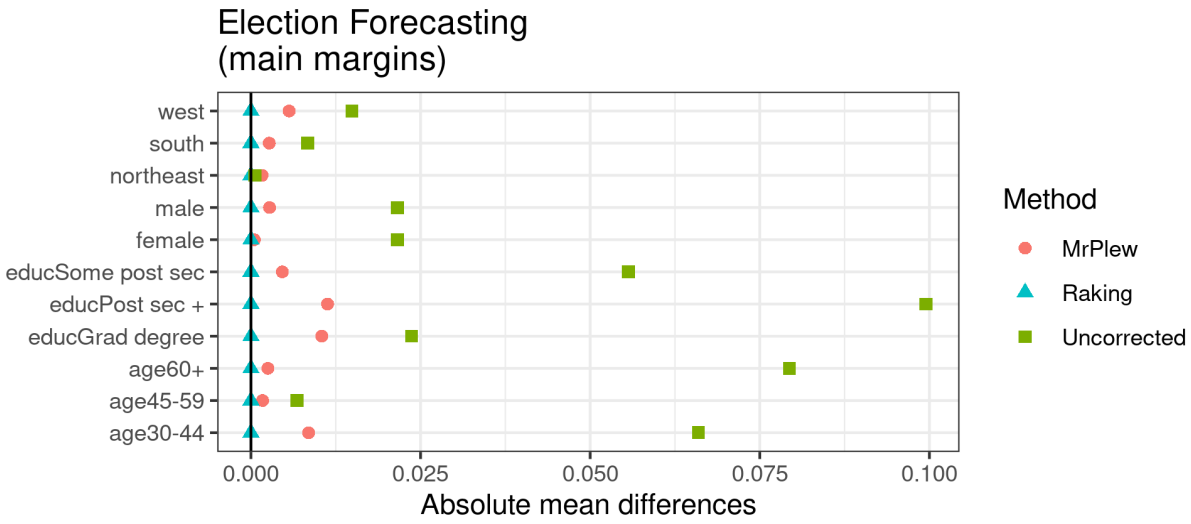
Figure 5 shows the implied covariate balance for the marginal regressors in the Same-Sex Marriage and Election Forecasting applications; we discuss the covariate balance for the Name Change application in Section 1.1. As expected, raking exactly balances the marginal covariates in both examples. For the Election Forecasting application, MrPlew also achieves excellent (if not exact) covariate balance on the marginals. For the Same-Sex Marriage application, however, the implied covariate balance for MrP appears substantially worse than the *uncorrected* covariate balance across several age and education levels. We assess whether these local imbalances translate to meaningful sensitivity in $\hat{\mu}^{\text{MrP}}$ in Section 6.2.

Appendix E includes additional balance plots reporting selected two-way interactions of the regressors used in raking.[4] For the Election Forecasting application, both raking and MrP yield good balance on the interactions as well. For the Same-Sex Marriage application, however, raking largely balances two-way interactions but MrP again fails to achieve good balance.

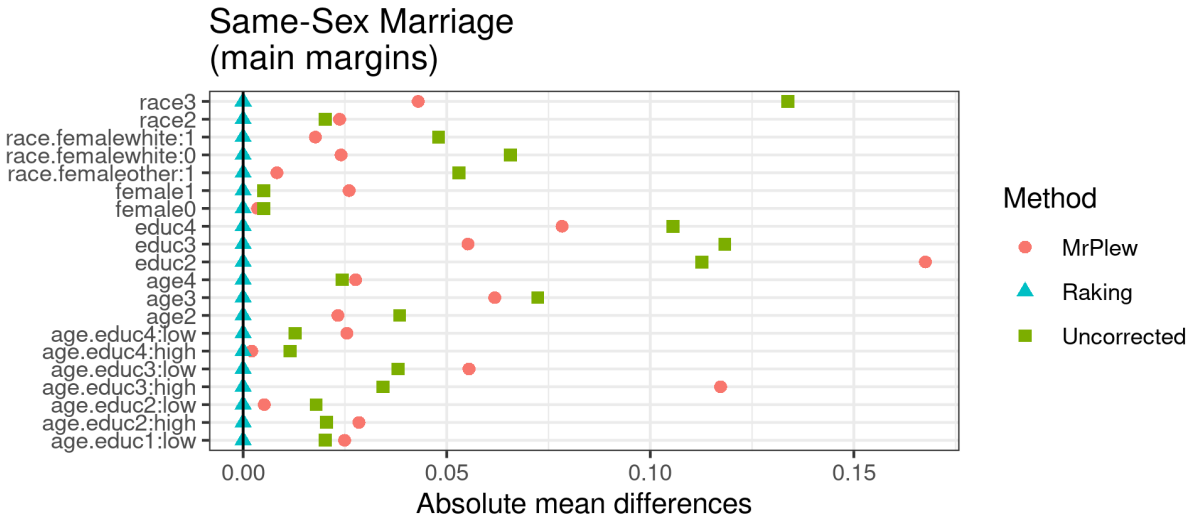
5.5 Subgroup contribution

Finally, we apply the results in Section 4.2.5 to assess the contribution of each subgroup to the final MrP estimate. We begin with the Name Change application from Section 1.1. As Figure 2 shows, there is substantial imbalance in the interaction between education level and decade married, and this interaction is plausibly associated with the outcome of interest. Figure 6 shows the subgroup contribution for this interaction, comparing the raking and MrP estimates. The right-hand side shows that the subgroup with negative weights is precisely the subgroup with the largest imbalance in Figure 2, suggesting that the imbalance in this interaction could contribute to the difference between the MrP and raking estimates for this application. We explore this interaction further in Section 5.6.

[4] Due to the large number of resulting balance checks, we assess covariate balance for interactions that occurred in at least 5% of both the survey and target populations. To preserve space, we only plot interactions in which either MrPlew or raking weights exhibited some minimal degree of imbalance.

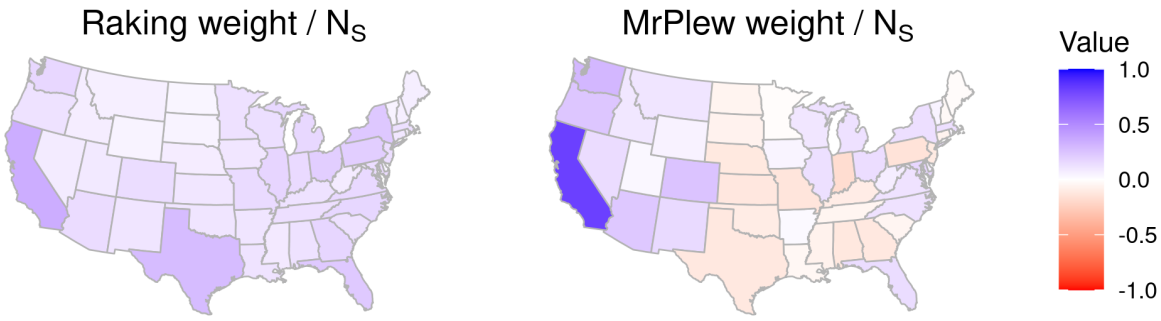


(a) Election Forecasting

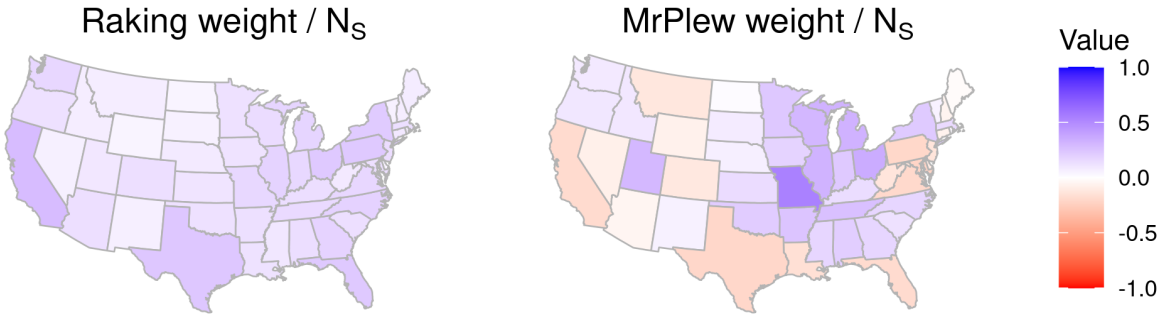


(b) Same-Sex Marriage

Figure 5: Implied covariate balance on raking marginals for the Election Forecasting and Same-Sex Marriage applications.



(a) Target: California



(b) Target: Missouri

Figure 7: Subgroup contribution for the Same-Sex Marriage application, with California and Missouri as targets.

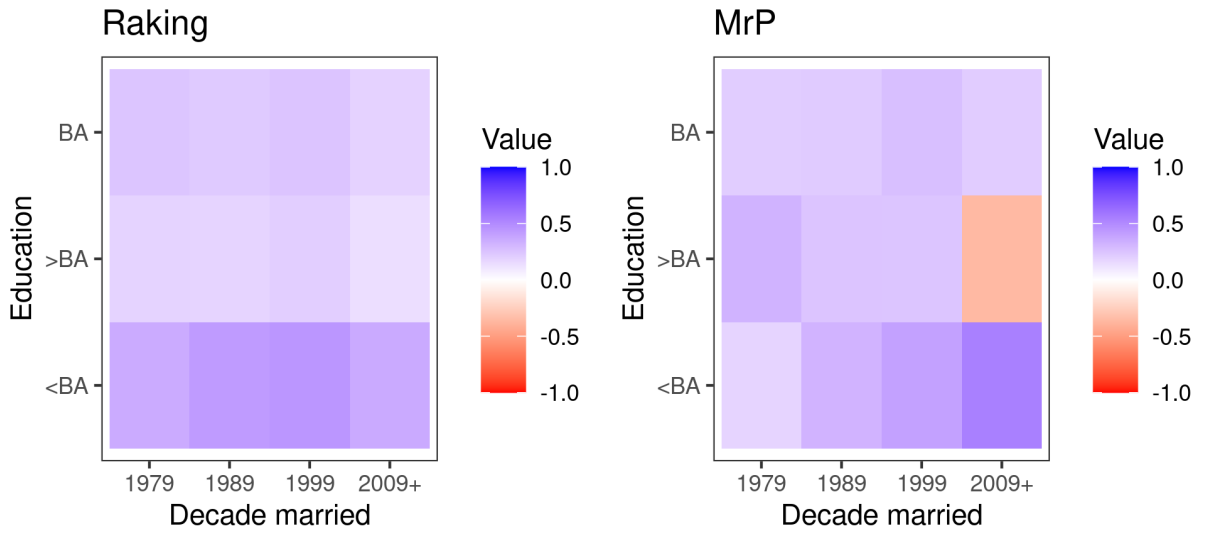


Figure 6: Subgroup contribution for Name Change

Next, we consider subgroup contributions for the Same-Sex Marriage application, where we focus on states as the subgroups of interest. To demonstrate this approach, we explore how state contributions change when the target geography changes from California to Missouri. Figure 7a shows the relative contribution of each state for the raking and MrP estimates of support for same-sex marriage in California. There are stark differences in the variation of state weights between the two methods, with substantially more variability for MrP than for raking.[5]

These differences partly reflect differences in the inputs for the two estimators: MrP includes state as a predictor in the model, while raking only includes non-geographic variables. They also reflect the fact that, as shown in Figure 3, locally equivalent weights for MrP can be negative, unlike raking weights. As a result, many states actually have a *negative* contribution on the California-specific estimate, with the largest negative weights for states in the Midwest. Finally, California itself receives much more weight under MrP than under raking; this reflects the fact that California is a large state with many survey observations and that MrP—but not raking—includes state as a predictor.

Figure 7b shows the same plot when we shift the target from California to Missouri. We see a similar overall pattern, with much more variability for MrPlew than for raking weights. However, the weights themselves are now quite different, with states like California and Pennsylvania showing negative weights for the MrP estimate for Missouri.

5.6 Name Change example: Expanding the outcome model

We fundamentally view MrPlew as an exploratory tool for model interrogation and argue that the results in this section give useful context for assessing the use of MrP in a given application. To illustrate one use of MrPlew in model development, we return to the Name Change example from Section 1.1. Recall that Figure 2 showed substantial covariate imbalance in the `decade_married_rk2009+:educ_group>BA` interaction, which was not modeled in the original MrP outcome model.

Although balance is intuitively desireable, imbalance does not itself diagnose model misspecification in general. For example, Example 4.2 demonstrates that correctly specified models can imbalance their own regressors, and Example 3.1 shows that calibration weights do not necessarily even include $\boldsymbol{Y}$ at all and so cannot possibly diagnose problems with $\mathcal{P}(y|\boldsymbol{x})$. In the presence of imbalance, we recommend first considering whether the imbalanced regressor plausibly aligns with the response, and then imformally checking whether there is evidence in the fitted residuals align with the imbalanced regressor. In this case, table 2 shows that, though the response does align with the imbalanced regressor, the residuals of the model do not, suggesting that the imbalance may not be affecting the MrP estimate.

Table 2: Mean response and residuals by interaction value for Name Change

`decade_married_rk2009+:educ_group>BA`	$\overline{y}$	$\overline{y - \hat{y}}$
0	0.412	-0.001
1	0.560	0.002

[5] In Figure 12 of Appendix E, we give additional evidence that the visual similarity between the raking subgroup contribution plots is not a bug. Raking produces very little state-to-state variation in weights between California and Missouri.

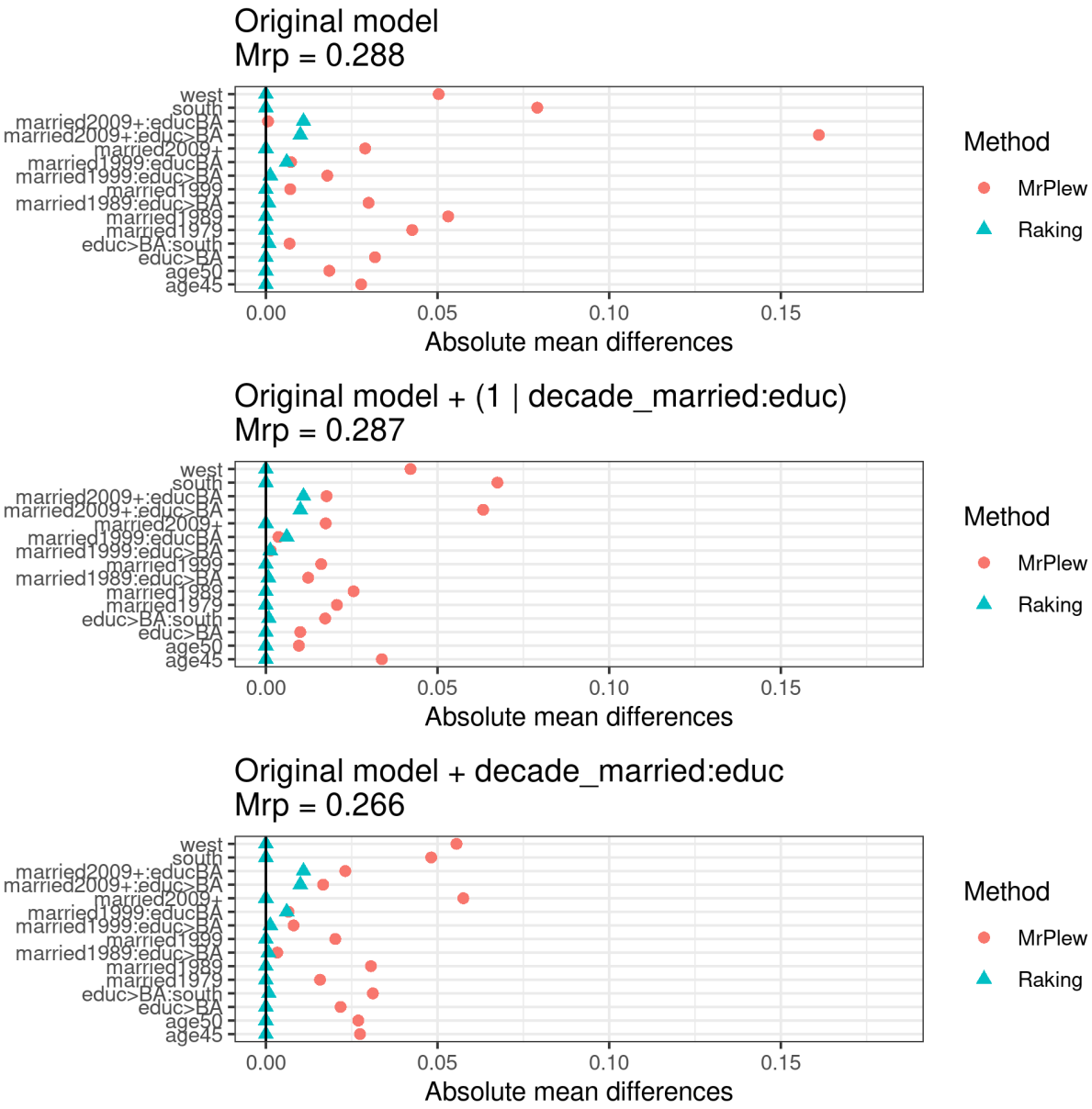


Figure 8: The effect of adding random and fixed effects for an imbalanced interaction in the Name Change analysis. Any regressor showing an imbalance in any of the models above a certain threshold is shown in each plot, so regressors that are not shown can be assumed to be reasonably balanced.

Nevertheless, suppose we want to modify the model to reduce the imbalance. A natural strategy to reduce imbalance is to expand the outcome model to include the interaction term.[6] Importantly, this interaction seems substantively plausible, so the choice to include this term is not driven by MrPlew alone. Figure 8 shows the effect on covariate balance of adding the imbalanced interaction to the model as a random effect (second

[6]Including the regressor directly does not generally yield exact balance: Example 4.2 shows that logistic regression balances $v(\boldsymbol{x})r(\boldsymbol{x})$, not $r(\boldsymbol{x})$. One could in principle include $\hat{v}(\boldsymbol{x})^{-1}r(\boldsymbol{x})$ for some plug-in $\hat{v}$, but this is suspect because $\hat{v}$ depends on $\boldsymbol{Y}$ and so is not properly a measurable function of $\boldsymbol{x}$.

panel) and as a fixed effect (third panel). Both lead to substantial improvement in covariate balance, even if the resulting point estimate is largely unchanged, as predicted by the residual means in table 2. Moreover, after including the interaction term, MrPlew covariate balance checks do not flag any additional imbalances of note.

6 Discussion and open questions

Our work raises a number of interesting questions which we hope will motivate further research.

6.1 Role of locally equivalent weights in model interrogation

We view MrPlew balance checks as diagnostics for the common setting in which the analyst trusts the model only approximately and wants to probe how it uses the data. Balance checks are not, however, tests of model specification. Examples 4.1 and 4.2 make this concrete. First, for MrP with an OLS outcome model (Section 2.3.3), the *globally* equivalent weights do not involve $\boldsymbol{Y}$ at all and so cannot detect misspecification of $\mathcal{P}(y|\boldsymbol{x}, \boldsymbol{\beta})$. Second, for MrP with a logistic regression model (Example 4.2), a correctly specified outcome model can still fail to achieve covariate balance.

Despite these limitations, we believe that the central idea of Section 4.2 can be modified to produce actual specification checks: design a perturbation of the data whose behavior is known under correct specification, and then design local robustness tests of whether the model exhibits the expected behavior. For example, if the response is correctly specified, then the MrP estimate should be approximately invariant to the distribution $\mathcal{P}_S(\boldsymbol{x})$. Pursuing this idea is the subject of ongoing work.

6.2 Assessing non-local robustness

Local robustness is best understood as a computationally efficient approximation to a non-local robustness question; see Giordano, Liu, et al. (2023a) and Giordano, Broderick, and Jordan (2018, Appendix C). In Section E.2, we assess how well our local approximation extrapolates in two applications, with mixed results: the imbalanced interaction in the Name Change example extrapolates well, while the imbalanced education level category in the Same-Sex Marriage example does not. The local results from Theorem 4.2 remain valid, but understanding when and why the approximation extrapolates is an important open question.

To assess non-local behavior, we first use MrPlew balance checks to identify covariate directions that, in principle, can lead to large changes in the estimate $\hat{\mu}^{\text{MrP}}$. Using the procedure described in section 4.2.4, we then repeatedly generate perturbed data sets $\breve{\boldsymbol{Y}}$ increasingly far from the original outcome and assess how well the predictions from MrPlew track the realized changes in $\hat{\mu}^{\text{MrP}}(\breve{\boldsymbol{Y}})$. Suppose we have selected a perturbation $\delta r(\boldsymbol{x})$ and that we have generated $\breve{\boldsymbol{Y}}$ satisfying Equation (21). Then we expect that

$$\underbrace{\frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\text{MrP}}\breve{y}_i - \frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\text{MrP}} y_i}_{\text{inferred change on binary vector}} \approx \underbrace{\delta\frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\text{MrP}} r_i}_{\substack{\text{inferred change on} \\ \text{continuous perturbation}}} . \tag{23}$$

Equation (23) is a combination of Equation (17) with the MrPlew weights, and Equation (21) with the law of large numbers as $N_S \to \infty$. Next, suppose we have identified a $\breve{\boldsymbol{Y}}$ satisfying Equation (23), and that our local approximation is good (e.g., see Equation (12)). Then we expect that

$$\underbrace{\hat{\mu}^{\mathrm{MrP}}(\breve{\boldsymbol{Y}}) - \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y})}_{\text{actual posterior change}} \approx \underbrace{\frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\mathrm{MrP}}\breve{y}_i - \frac{1}{N_S}\sum_{i\in[N_S]} w_i^{\mathrm{MrP}} y_i}_{\text{inferred change on binary vector}}. \tag{24}$$

Finally, if we have identified a perturbation that is imbalanced according to $\boldsymbol{W}^{\mathrm{MrP}}$, combining eqs. (23) and (24) we expect that

$$\underbrace{\hat{\mu}^{\mathrm{MrP}}(\breve{\boldsymbol{Y}}) - \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y})}_{\text{actual posterior change}} \quad \text{is far from} \quad \underbrace{\frac{1}{N_T}\sum_{j\in[N_T]} \pi_j r_j}_{\text{actual target change (eq. (18))}}. \tag{25}$$

That is, our local approximation *should* be able to produce binary datasets $\breve{\boldsymbol{Y}}$ whose posterior change does not match the true target population change.

In Section E.2, we ran experiments to check Equation (25) for the Name Change and Same-Sex Marriage analyses. For the Same-Sex Marriage example, we took $r(\boldsymbol{x})$ to be the imbalanced raking marginal `edu.cat2`; for the Name Change example, we took $r(\boldsymbol{x})$ to be the imbalanced interaction `decade_married_rk2009+:educ_group>BA`. As desired, the MrPlew prediction for the generated binary $\breve{\boldsymbol{Y}}$ indeed predicts a divergence with the truth: the generated $\breve{\boldsymbol{Y}}$ do in fact identify potentially problematic response vectors. For the Name Change example, this approach shows that the local approximation is quite good and extrapolates well. For the Same-Sex Marriage example, however, the extrapolation is quite poor, suggesting we should be wary of over-relying on the reported model checks.

It is known that linear approximations to posterior expectations may fail for large changes, especially in posteriors with large numbers of poorly-estimated random effects (Giordano and Broderick, 2024, Section 3). It is common in MrP problems to have a large number of random effects, and so local approximations should be taken with a grain of salt. However, precisely why the nonlinearity is so severe in the Same-Sex Marriage analysis but not in the Name Change analysis is an open question; see Appendix E for additional discussion. Exploring posterior nonlinearity in more depth is an important direction for future work.

6.3 Creating asymptotically globally linear MrP estimators with DrP

As a complement to assessing curvature as discussed in section 6.2, one might instead improve the usefulness of locally linear diagnostics by modifying MrP estimators that are approximately globally linear. Here, we briefly argue that a simple technique for doing so is provided by DrP (Ben-Michael, Feller, and Hartman, 2024). An alternative approach for creating globally linear MrP estimates is augmenting the MrP model with regressors that predict $\mathcal{P}_T(\boldsymbol{x})/\mathcal{P}_S(\boldsymbol{x})$, as discussed in Example A.2 of section A.4.

The DrP estimator is defined as follows. Let $\rho(\boldsymbol{x}) := \mathcal{P}_T(\boldsymbol{x})/\mathcal{P}_S(\boldsymbol{x})$ denote the Radon-Nikodym derivative of the target regressor distribution with respesct to the survey, and let $\hat{\rho}(\cdot)$ denote an estimate of $\rho(\cdot)$ that is independent of $\boldsymbol{Y}$ (e.g., formed with covariates only or by sample splitting). As usual, write $\hat{\rho}_i = \rho(\boldsymbol{x}_i)$.

For the present, take $\pi(\cdot) = 1$ for simplicity. Then the DrP estimator based on $\hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y})$ is (Ben-Michael, Feller, and Hartman, 2024, Equation (12))

$$\hat{\mu}^{\mathrm{DrP}}(\boldsymbol{Y}) = \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}) + \frac{1}{N_S} \sum_{i \in [N_S]} \hat{\rho}_i (y_i - \hat{m}_i).$$

Since $\hat{\mu}^{\mathrm{DrP}}(\boldsymbol{Y})$ is a function of $\boldsymbol{Y}$, we can form MrPlew weights for the DrP estimate, and a simple argument sketched in section A.5 shows that, if $\hat{\rho}(\cdot)$ is a consistent estimator of $\rho(\cdot)$, then

$$w_i^{\mathrm{DrP}} = \rho(\boldsymbol{x}_i) + o_p(1), \tag{26}$$

where $o_p(1)$ is a quantity that vanishes as both N_S and N_T go to infinity. Since $\rho(\boldsymbol{x}_i)$ does not depend on $\boldsymbol{Y}$, $\hat{\mu}^{\mathrm{DrP}}(\boldsymbol{Y})$ becomes linear as N_S and N_T go to infinity, and w_i^{DrP} converges to the importance ratio $\rho(\boldsymbol{x}_i)$. The authors are hopeful that modifications such as this can improve the shortcomings of local robustness demonstrated in section 6.2, though we leave rigorous theoretical and experimental analysis of this approach for future work.

7 Data and software availability

Code to produce all the experiments of our paper can be found at the git repo `https://github.com/rgiordan/MrPLocallyEquivalentWeightsPaper`. An open-source software implementation of MrPlew is available at `https://github.com/rgiordan/mrplew`.

References

Alexander, M. (2019). *Analyzing name changes after marriage using a non-representative survey*. URL: https://www.monicaalexander.com/posts/2019-08-07-mrp/.

Alexander, R. (2023). *Telling Stories with Data: With Applications in R*. Chapman and Hall/CRC.

Basu, S., S. Rao Jammalamadaka, and W. Liu (1996). "Local posterior robustness with parametric priors: Maximum and average sensitivity". In: *Maximum Entropy and Bayesian Methods*. Springer, pp. 97–106.

Belsley, D., E. Kuh, and R. Welsch (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley Classics Library. Originally published in 1980. Hoboken, New Jersey: John Wiley & Sons. ISBN: 978-0-471-69693-3.

Ben-Michael, Eli, Avi Feller, and Erin Hartman (2024). "Multilevel calibration weighting for survey data". In: *Political Analysis* 32.1, pp. 65–83.

Ben-Michael, Eli, Avi Feller, David A. Hirshberg, et al. (2021). "The Balancing Act in Causal Inference". arXiv: 2110.14831. URL: http://arxiv.org/abs/2110.14831.

Bisbee, James (2019). "BARP: Improving Mister P Using Bayesian Additive Regression Trees". In: *American Political Science Review* 113.4, pp. 1060–1065.

Bruns-Smith, David et al. (2025). "Augmented balancing weights as linear regression". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkaf019.

Bürkner, Paul-Christian (2017). "brms: An R Package for Bayesian Multilevel Models Using Stan". In: *Journal of Statistical Software* 80.1, pp. 1–28. DOI: 10.18637/jss.v080.i01.

Cabral, Rafael, David Bolin, and Håvard Rue (2025). "Robustness, model checking, and hierarchical models". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 87.3, pp. 632–652.

Chang, Ted and Phillip S Kott (2008). "Using calibration weighting to adjust for nonresponse under a plausible model". In: *Biometrika* 95.3, pp. 555–571.

Chattopadhyay, A. and J. Zubizarreta (2023). "On the implied weights of linear regression for causal inference". In: *Biometrika* 110.3, pp. 615–629.

Chen, Yilin, Pengfei Li, and Changbao Wu (2020). "Doubly robust inference with nonprobability survey samples". In: *Journal of the American Statistical Association* 115.532, pp. 2011–2021.

Cohen, P. (Apr. 2019). *Marital Name Change Survey*. DOI: 10.17605/OSF.IO/UZQDN. URL: osf.io/uzqdn.

Cook, D. (1977). "Detection of influential observation in linear regression". In: *Technometrics* 19.1, pp. 15–18.

— (1986). "Assessment of local influence". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 133–169.

DeBell, M. and J. Krosnick (2009). "Computing weights for American national election study survey data". In: *nes012427. Ann Arbor, MI, Palo Alto, CA: ANES Technical Report Series*.

Deming, W Edwards and Frederick F Stephan (1940). "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known". In: *The Annals of Mathematical Statistics* 11.4, pp. 427–444. ISSN: 0003-4851. DOI: 10.1214/aoms/1177731829. arXiv: arXiv:1306.3979v1.

Deville, J., C. Särndal, and O. Sautory (1993). "Generalized raking procedures in survey sampling". In: *Journal of the American statistical Association* 88.423, pp. 1013–1020.

Deville, Jean Claude and Carl Erik Särndal (1992). "Calibration estimators in survey sampling". In: *Journal of the American Statistical Association* 87.418, pp. 376–382. ISSN: 1537274X. DOI: 10.1080/01621459.1992.10475217.

Di Noia, Antonio, Fabrizio Ruggeri, and Antonietta Mira (2025). "Likelihood distortion and Bayesian local robustness". In: *Bayesian Analysis* 20.4, pp. 1261–1281.

Efron, B. and R. Tibshirani (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.

Elliott, Michael R and Richard Valliant (2017). "Inference for nonprobability samples". In: *Statistical Science*, pp. 249–264.

Firth, David and K. E. Bennett (1998). "Robust models in probability sampling". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 60.1, pp. 3–21. DOI: 10.1111/1467-9868.00105.

Fuller, W. (2011). *Sampling statistics*. John Wiley & Sons.

Gao, Chenyin, Shu Yang, and Jae Kwang Kim (2022). "Soft calibration for selection bias problems under mixed-effects models". In: *arXiv preprint arXiv:2206.01084*.

— (2023). "Soft calibration for selection bias problems under mixed-effects models". In: *Biometrika* 110.4, pp. 897–911.

Gelman, A. (1997). "Poststratification into many categories using hierarchical logistic regression". In: *Survey methodology* 23, p. 127.

Gelman, Andrew (2007). "Struggles with Survey Weighting and Regression Modeling". In: *Statistical Science* 22.2, pp. 153–164. DOI: 10.1214/088342306000000691.

Gelman, Andrew and Thomas C. Little (1997). "Poststratification Into Many Categories Using Hierarchical Logistic Regression". In: *Survey Methodology* 23.2, pp. 127–135.

Gelman, Andrew, Yajuan Si, and Brady T. West (2024). "MRPW: Regression, poststratification, and small-area estimation with weights". Working paper.

Gelman, Andrew, Aki Vehtari, et al. (2020). "Bayesian workflow". In: *arXiv preprint arXiv:2011.01808*.

Ghitza, Yair and Andrew Gelman (2013). "Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups". In: *American Journal of Political Science* 57.3, pp. 762–776. ISSN: 15405907. DOI: 10.1111/ajps.12004.

Giordano, R. and T. Broderick (2024). *The Bayesian Infinitesimal Jackknife for Variance*. arXiv: 2305.06466 [stat.ME]. URL: https://arxiv.org/abs/2305.06466.

Giordano, R., T. Broderick, and M. I. Jordan (2018). "Covariances, robustness and variational bayes". In: *Journal of machine learning research* 19.51.

Giordano, R., R. Liu, et al. (2023a). "Evaluating Sensitivity to the Stick-Breaking Prior in Bayesian Nonparametrics (Rejoinder)". In: *Bayesian Analysis* 18.1, pp. 287–366.

— (2023b). "Evaluating Sensitivity to the Stick-Breaking Prior in Bayesian Nonparametrics (with Discussion)". In: *Bayesian Analysis* 18.1, pp. 287–366.

Giordano, R., W. Stephenson, et al. (2019). "A Swiss army infinitesimal jackknife". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1139–1147.

Guggemos, Fabien and Yves Tillé (2010). "Penalized calibration in survey sampling: Design-based estimation assisted by mixed models". In: *Journal of Statistical Planning and Inference* 140.11, pp. 3199–3212. ISSN: 03783758. DOI: 10.1016/j.jspi.2010.04.010.

Gustafson, P. (1996). "Local sensitivity of posterior expectations". In: *The Annals of Statistics* 24.1, pp. 174–195.

— (2000). "Local robustness in Bayesian analysis". In: *Robust Bayesian Analysis*. Ed. by D. R. Insua and F. Ruggeri. Vol. 152. Springer Science & Business Media.

Haziza, David and Jean-François Beaumont (2017). "Construction of Weights in Surveys: A Review". In: *Statistical Science* 32.2, pp. 206–226. DOI: 10.1214/16-STS608. URL: https://doi.org/10.1214/16-STS608.

Henderson, H. and S. Searle (1981). "On deriving the inverse of a sum of matrices". In: *SIAM review* 23.1, pp. 53–60.

Huggins, J. and J. Miller (2023). "Reproducible model selection using bagged posteriors". In: *Bayesian analysis* 18.1, pp. 79–104.

Kass, R., L. Tierney, and J. Kadane (1989). "Approximate methods for assessing influence and sensitivity in Bayesian analysis". In: *Biometrika* 76.4, pp. 663–674.

Kastellec, J. (2024). *Publications | Jonathan P. Kastellec*. https://jkastellec.scholar.princeton.edu/publications. Accessed: July 17, 2025.

Kastellec, J., J. Lax, and J. Phillips (2010). "Estimating state public opinion with multi-level regression and poststratification using R". In: *Unpublished manuscript, Princeton University* 29.3.

Kennedy, Lauren, Aki Vehtari, and Andrew Gelman (2023). "Model validation for aggregate inferences in out-of-sample prediction". In: *arXiv preprint arXiv:2312.06334*.

Kleijn, B. and A. van der Vaart (2012). "The Bernstein-von-Mises theorem under misspecification". In: *Electronic Journal of Statistics* 6, pp. 354–381.

Koh, P. and P. Liang (2017). "Understanding black-box predictions via influence functions". In: *International conference on machine learning*. PMLR, pp. 1885–1894.

Kott, Phillip S and Ted Chang (2010). "Using calibration weighting to adjust for nonignorable unit nonresponse". In: *Journal of the American Statistical Association* 105.491, pp. 1265–1275.

Krantz, S. and H. Parks (2002). *The implicit function theorem: History, theory, and applications*. Vol. 202. 11. Springer.

Kuh, Swen et al. (2024). "Using leave-one-out cross validation (LOO) in a multilevel regression and poststratification (MRP) workflow: A cautionary tale". In: *Statistics in Medicine* 43.5, pp. 953–982.

Lax, J. and J. Phillips (2009). "Same-sex rights in the states: Public opinion and policy responsiveness". In: *American Political Science Review* 103.3, pp. 367–386.

Little, Roderick J.A. (2004). "To model or not to model? Competing modes of inference for finite population sampling". In: *Journal of the American Statistical Association* 99.466, pp. 546–556.

— (2006). "Calibrated Bayes: A Bayes/frequentist roadmap". In: *The American Statistician* 60.3, pp. 213–223.

— (2012). "Calibrated Bayes: An alternative inferential paradigm for official statistics". In: *Journal of Official Statistics* 28.3, pp. 309–372.

Lopez-Martin, J., J. Phillips, and A. Gelman (2026). *Multilevel regression and poststratification case studies*. URL: https://juanlopezmartin.github.io (visited on 03/26/2026).

Lumley, T. (2024). *survey: Analysis of complex survey samples*. R package version 4.4.

Mastny, T. (2018). *Study-of-Complex-Surveys*. Version master. Accessed: April 7, 2026. URL: https://github.com/tmastny/Study-of-Complex-Surveys.

Meng, Xiao-Li (2018). "Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election". In: *Annals of Applied Statistics* 12.2, pp. 685–726.

Montgomery, Jacob M. and Santiago Olivella (2018). "Tree-Based Models for Political Science Data". In: *American Journal of Political Science* 62.3, pp. 729–744. ISSN: 15405907. DOI: 10.1111/ajps.12361.

Park, Mingue and Wayne A. Fuller (2009). "The mixed model for survey regression estimation". In: *Journal of Statistical Planning and Inference* 139.4, pp. 1320–1331. ISSN: 03783758. DOI: 10.1016/j.jspi.2008.02.021.

Ruggles, S. et al. (2024). *IPUMS USA: Version 15.0 [dataset]*. DOI: 10.18128/D010.V15.0. URL: https://usa.ipums.org.

Särndal, Carl-Erik and Sixten Lundström (2005). *Estimation in surveys with nonresponse*. John Wiley & Sons.

Savitsky, Terrance D. and Daniell Toth (2016). "Bayesian estimation under informative sampling". In: *Electronic Journal of Statistics* 10.1, pp. 1677–1708.

Si, Yajuan, Natesh S. Pillai, and Andrew Gelman (2015). "Bayesian nonparametric weighted sampling inference". In: *Bayesian Analysis* 10.3, pp. 605–625.

Si, Yajuan, Rob Trangucci, et al. (2020). "Bayesian hierarchical weighting adjustment and survey inference". In: *Survey Methodology* 46.2, pp. 181–214.

Si, Yajuan and Peigen Zhou (2021). "Bayes-Raking: Bayesian finite population inference with known margins". In: *Journal of Survey Statistics and Methodology* 9.4, pp. 833–855.

Tausanovitch, Chris and Lynn Vavreck (2021). *Democracy Fund + UCLA Nationscape Project*. Accessed: August 1, 2025. URL: https://www.voterstudygroup.org/data/nationscape.

Thomas, Z., S. MacEachern, and M. Peruggia (2018). "Reconciling curvature and importance sampling based procedures for summarizing case influence in Bayesian models". In: *Journal of the American Statistical Association* 113.524, pp. 1669–1683.

Van der Vaart, A. (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.

Wang, Yixin and Jose R Zubizarreta (2020). "Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations". In: *Biometrika* 107.1, pp. 93–105. ISSN: 14643510. DOI: 10.1093/biomet/asz050. arXiv: 1705.00998.

Wang, Zhenhua, Jae Kwang Kim, and Shu Yang (2018). "An approximate Bayesian inference under informative sampling". In: *Biometrika* 105.1, pp. 91–102.

Wu, Changbao and Randy R. Sitter (2001). "A model-calibration approach to using complete auxiliary information from survey data". In: *Journal of the American Statistical Association* 96.453, pp. 185–193.

Yu, Bin and Karl Kumbier (2020). "Veridical data science". In: *Proceedings of the National Academy of Sciences* 117.8, pp. 3920–3929. DOI: 10.1073/pnas.1901326117.

Zhu, H. et al. (2007). "Perturbation selection and influence measures in local influence analysis". In: *Annals of Statistics* 35.6, pp. 2565–2588.

A Details of closed-form examples

A.1 Counterexample for Theorem 4.1

Example A.1. Suppose that $\mathcal{P}(y|\boldsymbol{x}, \boldsymbol{\beta}, \sigma)$ is normal with mean $\boldsymbol{\beta}^\intercal \boldsymbol{x}$ and variance σ^2. Then the sufficient statistics are (y, y^2), and the model does not satisfy Assumption 4.1. We will show that, in general, the estimator $\hat{V}$ in eq. (15) does not provide consistent estimates of the frequentist variance.

In this case, the log likelihood is given (up to a constant C not depending on y_i) by

$$\begin{aligned}
\ell(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma) &= -\frac{1}{2}\sigma^{-2}\left(y_i^2 - 2\boldsymbol{\beta}^\intercal \boldsymbol{x}_i y_i + \boldsymbol{\beta}^\intercal \boldsymbol{x}_i^\intercal \boldsymbol{x}_i \boldsymbol{\beta}\right) + C \Rightarrow \\
\nabla_y \ell(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma) &= -\sigma^{-2}\left(y_i - 2\boldsymbol{\beta}^\intercal \boldsymbol{x}_i y_i\right)
\end{aligned}$$

so w_i^{MrP} is given by (see eq. (10))

$$w_i^{\mathrm{MrP}} = N_S \mathrm{Cov}_{\mathcal{P}(\boldsymbol{\beta},\sigma|\boldsymbol{Y})}\left(-\sigma^{-2}\left(y_i - \boldsymbol{\beta}^\intercal \boldsymbol{x}_i\right), g(\boldsymbol{\beta})\right) .$$

In contrast, the empirical influence function of $\hat{\mu}^{\mathrm{MrP}}$ (Giordano and Broderick, 2024) is given by

$$\psi_i = \mathrm{Cov}_{\mathcal{P}(\boldsymbol{\beta},\sigma|\boldsymbol{Y})}\left(\ell(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma), g(\boldsymbol{\beta})\right) ,$$

and $N_S \psi_i \neq w_i^{\mathrm{MrP}} \varepsilon_i$ in general, even asymptotically. For example, the coefficient in front of the y_i^2 term is different by a factor of two. However, Theorem 2 of (Giordano and Broderick, 2024) shows that, under mild regularity conditions,

$$\frac{1}{N_S}\sum_{i \in [N_S]}\left(N\psi_i - (N\overline{\psi})\right)^2 \to \boldsymbol{V}.$$

Since $\hat{V}$ of Theorem 4.1 converges, in general, to a different limit than that of the IJ estimator, it cannot be consistent.

□

A.2 Details for Example 3.1

It will be convenient to define the following quantities:

$$\hat{\mathbf{M}}_{xx} := \frac{1}{N_S}\boldsymbol{X}^\intercal \boldsymbol{X} \qquad \widetilde{\mathbf{M}}_{xx} := \hat{\mathbf{M}}_{xx} + \frac{\sigma^2}{N_S}\boldsymbol{\Sigma}^{-1} \qquad \hat{\mathbf{M}}_{xy} := \frac{1}{N_S}\boldsymbol{X}^\intercal \boldsymbol{Y}.$$

For eq. (13), we have

$$\begin{aligned}
\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y}) &= \mathcal{N}\left(\widetilde{\mathbf{M}}_{xx}^{-1}\hat{\mathbf{M}}_{xy}, \frac{\sigma^2}{N_S}\widetilde{\mathbf{M}}_{xx}^{-1}\right) \Rightarrow \\
\hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}) &= \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left[\boldsymbol{\beta}\right] = \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T \widetilde{\mathbf{M}}_{xx}^{-1}\hat{\mathbf{M}}_{xy} \Rightarrow \\
\boldsymbol{W}^{\mathrm{MrP}} &= \frac{1}{N_T}\boldsymbol{X}\widetilde{\mathbf{M}}_{xx}^{-1}\boldsymbol{X}_T^\intercal \pi. \qquad (27)
\end{aligned}$$

We can derive eq. (20)

$$\begin{aligned}
\frac{1}{\delta}\mathrm{Imbalance}(\boldsymbol{x}, \boldsymbol{W}^{\mathrm{MrP}}) &= \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T - \frac{1}{N_S}\boldsymbol{W}^{\mathrm{MrP}\intercal}\boldsymbol{X} \\
&= \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T - \frac{1}{N_S N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T \widetilde{\mathbf{M}}_{xx}^{-1}\boldsymbol{X}^\intercal \boldsymbol{X} \\
&= \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T \left(\boldsymbol{I}_P - \widetilde{\mathbf{M}}_{xx}^{-1}\hat{\mathbf{M}}_{xx}\right)
\end{aligned}$$

Finally, we prove that $\boldsymbol{W}^{\mathrm{MrP}}$ minimizes the expected mean squared error marginally over $\boldsymbol{\beta}$. First, marginally we have

$$\mathrm{Cov}_{p(\boldsymbol{Y}|\boldsymbol{X})}\left(\boldsymbol{Y}\right) = \sigma^2 \boldsymbol{I}_{N_S} + \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\intercal \quad \text{(marginally over } \boldsymbol{\beta}\text{, correct specification).}$$

Then

$$\begin{aligned}
\mathcal{E}(\boldsymbol{W}) &:= \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{Y}_T - \frac{1}{N_S}\boldsymbol{W}^\intercal \boldsymbol{Y} \\
&= \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T\boldsymbol{\beta} + \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{\varepsilon}_T - \frac{1}{N_S}\boldsymbol{W}^\intercal \boldsymbol{X}\boldsymbol{\beta} - \frac{1}{N_S}\boldsymbol{W}^\intercal \boldsymbol{\varepsilon} \\
&= \left(\frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T - \frac{1}{N_S}\boldsymbol{W}^\intercal \boldsymbol{X}\right)\boldsymbol{\beta} + \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{\varepsilon}_T - \frac{1}{N_S}\boldsymbol{W}^\intercal \boldsymbol{\varepsilon} \Rightarrow \\
\mathbb{E}_{\mathcal{P}(\boldsymbol{Y},\boldsymbol{Y}_T|\boldsymbol{X},\boldsymbol{X}_T)}\left[\mathcal{E}(\boldsymbol{W})^2\right] &= \left(\frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T - \frac{1}{N_S}\boldsymbol{W}^\intercal \boldsymbol{X}\right)\boldsymbol{\Sigma}\left(\frac{1}{N_T}\boldsymbol{X}_T^\intercal \boldsymbol{\pi} - \frac{1}{N_S}\boldsymbol{X}^\intercal \boldsymbol{W}\right) + \\
&\quad \frac{1}{N_S^2}\boldsymbol{W}^\intercal \boldsymbol{W}\sigma^2 + \frac{1}{N_T^2}\boldsymbol{\pi}^\intercal \boldsymbol{\pi}\sigma^2 \Rightarrow \\
\frac{\partial}{\partial \boldsymbol{W}}\mathbb{E}_{\mathcal{P}(\boldsymbol{Y},\boldsymbol{Y}_T|\boldsymbol{X},\boldsymbol{X}_T)}\left[\mathcal{E}(\boldsymbol{W})^2\right] &= \frac{2}{N_S^2}\left(\sigma^2 \boldsymbol{I}_{N_S} + \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\intercal\right)\boldsymbol{W} - \frac{2}{N_S N_T}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}_T^\intercal \boldsymbol{\pi} \Rightarrow \\
\boldsymbol{W}^* &= \frac{N_S^2}{N_S N_T}\left(\sigma^2 \boldsymbol{I}_{N_S} + \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\intercal\right)^{-1}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}_T^\intercal \boldsymbol{\pi} \\
&= \frac{N_S}{N_T}\boldsymbol{X}\left(\sigma^2 \boldsymbol{I}_P + \boldsymbol{\Sigma}\boldsymbol{X}^\intercal \boldsymbol{X}\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{X}_T^\intercal \boldsymbol{\pi} \\
&= \frac{N_S}{N_T}\boldsymbol{X}\left(\sigma^2 \boldsymbol{\Sigma}^{-1} + \boldsymbol{X}^\intercal \boldsymbol{X}\right)^{-1}\boldsymbol{X}_T^\intercal \boldsymbol{\pi} \\
&= \frac{1}{N_T}\boldsymbol{X}\widetilde{\mathbf{M}}_{xx}^{-1}\boldsymbol{X}_T^\intercal \boldsymbol{\pi} = \boldsymbol{W}^{\mathrm{MrP}}.
\end{aligned}$$

In the preceding display, we used the push-through identity for matrix inverses (Henderson and Searle, 1981).

A.3 Details for Examples 3.2 and 4.2

Recall the setup in Example 3.2. We have

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} \frac{1}{N_S} \sum_{i \in [N_S]} \log \mathcal{P}(y|\boldsymbol{x}, \boldsymbol{\beta}),$$

and so

$$\nabla_{\boldsymbol{\beta}} \frac{1}{N_S} \sum_{i \in [N_S]} \log \mathcal{P}(y_i|\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}) = \frac{1}{N_S} \sum_{i \in [N_S]} (y_i - \hat{y}_i)\boldsymbol{x}_i \quad \text{and}$$
$$\nabla^2_{\boldsymbol{\beta}} \frac{1}{N_S} \sum_{i \in [N_S]} \log \mathcal{P}(y_i|\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}) = -\frac{1}{N_S} \sum_{i \in [N_S]} \hat{v}_i \boldsymbol{x}_i \boldsymbol{x}_i^\intercal = -\frac{1}{N_S} \boldsymbol{X}^\intercal \boldsymbol{V} \boldsymbol{X}.$$

Note that

$$\nabla_{\boldsymbol{\beta}} g(\hat{\boldsymbol{\beta}}) = \frac{1}{N_T} \sum_{j \in [N_T]} \pi_j \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})} \left[m^{\text{logit}}(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j) \right]$$

The Bernstein-von Mises theorem then states that, for large N_S,

$$\mathcal{P}\left(\sqrt{N_S}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) | \boldsymbol{Y} \right) \overset{\text{approx}}{\approx} \mathcal{N}\left(\mathbf{0}, \left(\frac{1}{N_S} \boldsymbol{X}^\intercal \boldsymbol{V} \boldsymbol{X} \right)^{-1} \right). \tag{28}$$

We can use the delta method and the approximation eq. (28) to approximate the MrPlew weights:

$$\begin{aligned} w_i^{\text{MrP}} &= N_S \text{Cov}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})} \left(g(\boldsymbol{\beta}), \frac{\partial}{\partial y_i} \log \mathcal{P}(y_i|\boldsymbol{x}_i, \beta) \right) \\ &= N_S \text{Cov}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})} \left(g(\boldsymbol{\beta}), \boldsymbol{\beta}^\intercal \boldsymbol{x}_i \right) \\ &\approx \left. \frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\intercal} \right|_{\hat{\boldsymbol{\beta}}} \left(\frac{1}{N_S} \boldsymbol{X}^\intercal \boldsymbol{V} \boldsymbol{X} \right)^{-1} \boldsymbol{x}_i. \end{aligned}$$

Next,

$$\left. \frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\hat{\boldsymbol{\beta}}} = \frac{1}{N_T} \sum_{j \in [N_T]} \pi_j \left. \frac{\partial m^{\text{logit}}(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)}{\partial \boldsymbol{\beta}^\intercal} \right|_{\hat{\boldsymbol{\beta}}} = \frac{1}{N_T} \sum_{j \in [N_T]} \pi_j \hat{v}_j \boldsymbol{x}_j = \frac{1}{N_T} \boldsymbol{X}_T^\intercal \boldsymbol{V}_T \boldsymbol{\pi}.$$

Combining gives eq. (14).

Next, we consider covariate balance, beginning with the variance-weighted function $r(\boldsymbol{x}) = v\boldsymbol{x}$, where we approximate $v(\boldsymbol{x}_i)$ with $\hat{v}_i$ and $v(\boldsymbol{x}_j)$ with $\hat{v}_j$.

$$\begin{aligned} &\frac{1}{\delta} \text{Imbalance}(v\boldsymbol{x}, \boldsymbol{W}^{\text{MrP}}) = \\ &\quad \frac{1}{N_T} \boldsymbol{\pi}^\intercal \boldsymbol{V}_T \boldsymbol{X}_T - \frac{1}{N_S} \boldsymbol{W}^{\text{MrP}\intercal} \boldsymbol{V} \boldsymbol{X} = \\ &\quad \frac{1}{N_T} \boldsymbol{\pi}^\intercal \boldsymbol{V}_T \boldsymbol{X}_T - \frac{1}{N_T} \boldsymbol{\pi}^\intercal \boldsymbol{V}_T \boldsymbol{X}_T \left(\frac{1}{N_S} \boldsymbol{X}^\intercal \boldsymbol{V} \boldsymbol{X} \right)^{-1} \frac{1}{N_S} \boldsymbol{X}^\intercal \boldsymbol{V} \boldsymbol{X} = \\ &\quad \frac{1}{N_T} \boldsymbol{\pi}^\intercal \boldsymbol{V}_T \boldsymbol{X}_T - \frac{1}{N_S} \boldsymbol{\pi}^\intercal \boldsymbol{V}_T \boldsymbol{X}_T = \mathbf{0}. \end{aligned}$$

In contrast, the imbalance for $r(\boldsymbol{x}) = \boldsymbol{x}$ is generally not zero:

$$\begin{aligned}
&\frac{1}{\delta}\text{Imbalance}(\boldsymbol{x}, \boldsymbol{W}^{\text{MrP}}) = \\
&\quad \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T - \frac{1}{N_S}\boldsymbol{W}^{\text{MrP}\intercal}\boldsymbol{X} = \\
&\quad \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T - \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{V}_T \boldsymbol{X}_T \left(\frac{1}{N_S}\boldsymbol{X}^\intercal \boldsymbol{V}\boldsymbol{X}\right)^{-1} \frac{1}{N_S}\boldsymbol{X}^\intercal \boldsymbol{X} = \\
&\quad \frac{1}{N_T}\boldsymbol{\pi}^\intercal \boldsymbol{X}_T \left(\boldsymbol{I}_P - \left(\frac{1}{N_S}\boldsymbol{X}^\intercal \boldsymbol{V}\boldsymbol{X}\right)^{-1} \frac{1}{N_S}\boldsymbol{X}^\intercal \boldsymbol{X}\right) \neq \boldsymbol{0}. \text{ (in general)}
\end{aligned}$$

A.4 Imbalance is not due to nonlinearity alone

Example A.2 (Imbalance is not due to nonlinearity alone)**.** One might wonder whether the failure of logistic regression to balance the covariates in Example 4.2 is due to the nonlinearity of the mapping $\boldsymbol{Y} \mapsto \hat{\mu}^{\text{MrP}}(\boldsymbol{Y})$. We show here that this is not the case, and that in fact certain MrP estimators with nonlinear link functions can be approximately linear in $\boldsymbol{Y}$ for large N_S.

We consider again the asymptotic regime of logistic regression in Examples 3.2 and 4.2. Suppose we make the following (improbable) assumption:

$$\text{Assume that there exists } \alpha \text{ such that } \frac{\pi(\boldsymbol{x})\mathcal{P}_T(\boldsymbol{x})}{\mathcal{P}_S(\boldsymbol{x})} = \boldsymbol{\alpha}^\intercal \boldsymbol{x}. \tag{29}$$

Equation (29) may be an unreasonable assumpton in general, but with appropriate restrictions on the domain of $\boldsymbol{x}$, one could certainly generate data according to eq. (29), and so it is not impossible that eq. (29) can be satisfied. When eq. (29) holds, then we show below that, up to asymptotic approximations,

$$\hat{\mu}^{\text{MrP}}(\boldsymbol{Y}) \approx \frac{1}{N_S}\sum_{i\in[N_S]} \boldsymbol{\alpha}^\intercal \boldsymbol{x}_i y_i, \tag{30}$$

where the approximation becomes exact as $N_S \to \infty$.

Equation (30) is perhaps surprising. Note that, even if one found the assumption in eq. (29) plausible, it would be difficult to directly estimate $\boldsymbol{\alpha}$ because we do not directly observe $\mathcal{P}_T(\boldsymbol{x})$ and $\mathcal{P}_S(\boldsymbol{x})$. However, eq. (30) shows that w_i^{MrP} in fact acts as a consistent estimator of $\boldsymbol{\alpha}$. □

Derivation. Recall that $\hat{\boldsymbol{\beta}}$ is the solution to the estimating equation

$$\nabla_{\boldsymbol{\beta}} \frac{1}{N_S}\sum_{i\in[N_S]} \log \mathcal{P}(y_i|\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}) = \frac{1}{N_S}\sum_{i\in[N_S]} (y_i - \hat{y}_i)\boldsymbol{x}_i = \boldsymbol{0}. \tag{31}$$

We will refer to the Radon-Nikodym derivative of $\pi(\cdot)\mathcal{P}_T(\cdot)$ with respect to $\mathcal{P}_S(\boldsymbol{x})$ as the "importance ratio," since it is the appropriate weighting for a Horwitz-Thompson importance sampling estimator for converting a sample from $\mathcal{P}_S(\boldsymbol{x})$ into a sample from $\mathcal{P}_T(\boldsymbol{x})$. In a causal inference setting, the importance ratio is equivalent to the propensity score.

When eq. (29) holds, and when N_T and N_S are large enough to invoke a law of large numbers, we can write

$$\begin{aligned}
\hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}) &= \frac{1}{N_T}\sum_{j\in[N_T]}\pi_j\hat{y}_j && \text{(definition)}\\
&\approx \frac{1}{N_T}\sum_{j\in[N_T]}\pi_j\boldsymbol{x}_j^\intercal\hat{\boldsymbol{\beta}} && \text{(Bernstein von-Mises)}\\
&\approx \int \pi(\boldsymbol{x})\boldsymbol{x}^\intercal\mathcal{P}_T(\boldsymbol{x})d\boldsymbol{x}\hat{\boldsymbol{\beta}} && \text{(law of large numbers)}\\
&= \boldsymbol{\alpha}^\intercal\left(\int \boldsymbol{x}\boldsymbol{x}^\intercal\mathcal{P}_S(\boldsymbol{x})d\boldsymbol{x}\right)\hat{\boldsymbol{\beta}} && \text{(eq. (29))}\\
&\approx \boldsymbol{\alpha}^\intercal\left(\frac{1}{N_S}\sum_{i\in[N_S]}\boldsymbol{x}_i\boldsymbol{x}_i^\intercal\right)\hat{\boldsymbol{\beta}} && \text{(law of large numbers)}\\
&= \boldsymbol{\alpha}^\intercal\frac{1}{N_S}\sum_{i\in[N_S]}\boldsymbol{x}_i\hat{y}_i && \text{(definition)}\\
&= \frac{1}{N_S}\sum_{i\in[N_S]}\boldsymbol{\alpha}^\intercal\boldsymbol{x}_i y_i. && \text{eq. (31)}
\end{aligned}$$

Up to the approximations in the preceding display, we thus have that $\boldsymbol{Y}\mapsto\hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y})$ is linear, and

$$w_i^{\mathrm{MrP}} \approx \boldsymbol{\alpha}^\intercal\boldsymbol{x}_i = \frac{\pi(\boldsymbol{x}_i)\mathcal{P}_T(\boldsymbol{x}_i)}{\mathcal{P}_S(\boldsymbol{x}_i)}. \tag{32}$$

Despite the linearity of $\boldsymbol{Y}\mapsto\hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y})$, the conclusions of Example 4.2 still hold: logistic regression, even when linear, balances $v(\boldsymbol{x})\boldsymbol{x}$, not $\boldsymbol{x}$.

A.5 Global linearity of DrP

We continue the discussion in section 6.3 to support eq. (26). First, we can expand DrP as

$$\hat{\mu}^{\mathrm{DrP}}(\boldsymbol{Y}) = \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}) + \frac{1}{N_S}\sum_{i\in[N_S]}\hat{\rho}_i(y_i-\hat{m}_i) = \frac{1}{N_T}\sum_{j\in[N_T]}\hat{m}_j + \frac{1}{N_S}\sum_{i\in[N_S]}\hat{\rho}_i(y_i-\hat{m}_i).$$

The MrPlew weights for DrP are given by

$$w_{i'}^{\mathrm{DrP}} = \frac{N_S}{N_T}\sum_{j\in[N_T]}\frac{\partial\hat{m}_j}{\partial y_{i'}} - \frac{N_S}{N_T}\sum_{i\in[N_S]}\hat{\rho}_i\frac{\partial\hat{m}_j}{\partial y_{i'}} + \hat{\rho}_{i'},$$

where we have used the fact that $\hat{\rho}(\cdot)$ does not depend on $y_{i'}$.

Applying a LLN to both the survey and target populations gives

$$\begin{aligned}
w_{i'}^{\mathrm{DrP}} &\approx \int N_S\frac{\partial\hat{m}(\boldsymbol{x})}{\partial y_{i'}}\mathcal{P}_T(\boldsymbol{x}) - \int N_S\frac{\partial\hat{m}(\boldsymbol{x})}{\partial y_{i'}}\hat{\rho}(\boldsymbol{x})\mathcal{P}_S(\boldsymbol{x}) + \hat{\rho}_{i'}\\
&= \int N_S\frac{\partial\hat{m}(\boldsymbol{x})}{\partial y_{i'}}(\rho(\boldsymbol{x})-\hat{\rho}(\boldsymbol{x}))\mathcal{P}_S(\boldsymbol{x}) + \hat{\rho}_{i'}.
\end{aligned}$$

Under the assumption that $\mathrm{Var}_{\mathcal{P}_S(\boldsymbol{x})}\left(N_S \frac{\partial \hat{m}_i(\boldsymbol{x})}{\partial y_{i'}}\right) = O_p(1)$ (see Lemma B.4) and $\hat{\rho}(\cdot)$ is consistent in the sense that $\mathrm{Var}_{\mathcal{P}_S(\boldsymbol{x})}\left(\rho(\boldsymbol{x}) - \hat{\rho}(\boldsymbol{x})\right) = o_p(1)$, Cauchy-Schwartz gives

$$w_{i'}^{\mathrm{DrP}} = \hat{\rho}_{i'} + o_p(1). \tag{33}$$

An argument analogous to Theorem 4.2 can make this argument uniform in $'$, and justify the replacement of $\hat{\rho}_{i'}$ with $\rho(\boldsymbol{x}_i)$, though a careful rigorous treatment is beyond the scope of the current paper.

B Proof of Theorem 4.1

We first note that the boundedness of $\boldsymbol{x}$ in Assumption 4.3 implies Assumption B.1, which is more technical but closer to what we actually need in the proof of Lemma B.1.

Assumption B.1. Let Assumption 4.3 hold (recall that this includes Assumptions 4.1 and 4.2), but in place of the boundedness of $\boldsymbol{x}$, assume that

- $\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\|\boldsymbol{x}\|_2^2\right] < \infty$
- $\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[y^2\|\boldsymbol{x}\|_2^2\right] < \infty$
- $\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\left\|\nabla_\eta^k A(\boldsymbol{\beta}^{*\intercal}\boldsymbol{x})\boldsymbol{x}^{\otimes k}\right\|_2^2\right] < \infty$ for $k \in \{0,\dots,4\}$
- $\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\left\|\nabla_\eta^4 A(\boldsymbol{\beta}^\intercal\boldsymbol{x})\boldsymbol{x}^{\otimes 4}\right\|_2^2\right] < \infty$ for some $\Delta > 0$.

In place of the assumption that $\mathcal{P}(\boldsymbol{\beta})$ has bounded support, assume that

- $\mathbb{E}_{\mathcal{P}(\boldsymbol{\beta})}\left[g(\boldsymbol{\beta})^2\right] < \infty$.
- $\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\mathbb{E}_{\mathcal{P}(\boldsymbol{\beta})}\left[\ell(y|\boldsymbol{x},\boldsymbol{\beta})^2\right]\right] < \infty$.
- $\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\mathbb{E}_{\mathcal{P}(\boldsymbol{\beta})}\left[g(\boldsymbol{\beta})^2\ell(y|\boldsymbol{x},\boldsymbol{\beta})^2\right]\right] < \infty$.

□

For the remainder of the proof, let $\mathcal{B}_\Delta := \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \le \Delta\}$ to be the Δ–ball around $\boldsymbol{\beta}^*$. We first show a technical regularity lemma.

Lemma B.1. *Under Assumptions 4.1 to 4.3, there exists a $\Delta > 0$ such that, for all $\boldsymbol{\beta} \in \mathcal{B}_\Delta$, $\ell(\boldsymbol{\beta})$ is four times continuously differentiable and the exchange of partial differentiation with respect to $\boldsymbol{\beta}$ and integration with respect to $\mathcal{P}_S(y,\boldsymbol{x})$ is justified.*

Additionally, for $0 \le k \le 4$, there exists functions $M_k(\boldsymbol{x},y)$ such that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\left\|\nabla_{\boldsymbol{\beta}}^k \ell(\boldsymbol{\beta})\right\|_2^2 \le M_k(\boldsymbol{x},y) \quad \textit{and} \quad \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[M_k(\boldsymbol{x},y)\right] < \infty. \tag{34}$$

Proof. We first show that

$$\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\left\|\nabla_\eta^k A(\boldsymbol{\beta}^\intercal\boldsymbol{x})\boldsymbol{x}^{\otimes k}\right\|_2^2\right] \leq \infty \quad \text{for } 0\leq k\leq 4.$$

Note that $\nabla_\eta^k A(\boldsymbol{\beta})$ exists for all k by standard properties of exponential families. Let $\boldsymbol{\beta}_t := \boldsymbol{\beta}^* + t(\boldsymbol{\beta}-\boldsymbol{\beta}^*)$ for $t\in[0,1]$. Then for any $k\leq 3$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\left\|\nabla_\eta^k A(\boldsymbol{\beta}^\intercal\boldsymbol{x})\boldsymbol{x}^{\otimes k}\right\|_2^2\right] \\
&\quad= \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\left\|\left(\int_0^1 \frac{\partial}{\partial t}\nabla_\eta^k A(\boldsymbol{\beta}_t^\intercal\boldsymbol{x})dt + \nabla_\eta^k A(\boldsymbol{\beta}^{*\intercal}\boldsymbol{x})\right)\boldsymbol{x}^{\otimes k}\right\|_2^2\right] \\
&\quad= \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\left\|\left(\int_0^1 \nabla_\eta^{k+1} A(\boldsymbol{\beta}_t^\intercal\boldsymbol{x})dt(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\intercal\boldsymbol{x} + \nabla_\eta^k A(\boldsymbol{\beta}^{*\intercal}\boldsymbol{x})\right)\boldsymbol{x}^{\otimes k}\right\|_2^2\right] \\
&\quad\leq \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\left\|\left(\int_0^1 \nabla_\eta^{k+1} A(\boldsymbol{\beta}_t^\intercal\boldsymbol{x})dt(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\intercal\boldsymbol{x}\right)\boldsymbol{x}^{\otimes k}\right\|_2^2\right] + \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\left\|\nabla_\eta^k A(\boldsymbol{\beta}^{*\intercal}\boldsymbol{x})\boldsymbol{x}^{\otimes k}\right\|_2^2\right] \\
&\quad\leq \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\left\|\left(\int_0^1 \nabla_\eta^{k+1} A(\boldsymbol{\beta}_t^\intercal\boldsymbol{x})dt\right)\boldsymbol{x}^{\otimes(k+1)}\right\|_2^2\right]\Delta + \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\left\|\nabla_\eta^k A(\boldsymbol{\beta}^{*\intercal}\boldsymbol{x})\boldsymbol{x}^{\otimes k}\right\|_2^2\right] \\
&\quad\leq \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\left\|\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\nabla_\eta^{k+1} A(\boldsymbol{\beta}^\intercal\boldsymbol{x})\boldsymbol{x}^{\otimes(k+1)}\right\|_2^2\right]\Delta + \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\left\|\nabla_\eta^k A(\boldsymbol{\beta}^{*\intercal}\boldsymbol{x})\boldsymbol{x}^{\otimes k}\right\|_2^2\right]
\end{aligned}$$

The result follows by backward induction on k starting at $k=4$ using Assumption B.1.

We now turn to bounding the gradients of $\ell(\boldsymbol{\beta})$. Since $\nabla_{\boldsymbol{\beta}}^k\ell(\boldsymbol{\beta}) = \nabla_\eta^k A(\boldsymbol{\beta}^{*\intercal}\boldsymbol{x})\boldsymbol{x}^{\otimes k}$ for $k\geq 2$, we have already uniformly bounded the second derivatives and higher. For the first derivative, we can expand

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\|\nabla_{\boldsymbol{\beta}}\ell(y|\boldsymbol{x},\boldsymbol{\beta})\|_2^2\right] = \\
&\quad\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\left\|y\boldsymbol{x} + \nabla_\eta^1 A(\boldsymbol{\beta}^\intercal\boldsymbol{x})\boldsymbol{x}\right\|_2^2\right] \leq \\
&\quad\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[y^2\|\boldsymbol{x}\|_2^2\right] + \sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\left\|\nabla_\eta^1 A(\boldsymbol{\beta}^\intercal\boldsymbol{x})\boldsymbol{x}\right\|_2^2\right] + 2\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[y\nabla_\eta^1 A(\boldsymbol{\beta}^\intercal\boldsymbol{x})\boldsymbol{x}\right] \leq \\
&\quad\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[y^2\|\boldsymbol{x}\|_2^2\right] + \sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\left\|\nabla_\eta^1 A(\boldsymbol{\beta}^\intercal\boldsymbol{x})\boldsymbol{x}\right\|_2^2\right] + \\
&\qquad 2\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\left\|\nabla_\eta^1 A(\boldsymbol{\beta}^\intercal\boldsymbol{x})\boldsymbol{x}\right\|_2^2\right]\mathbb{E}_{\mathcal{P}_S(y)}\left[y^2\right]\right)^{1/2} < \infty,
\end{aligned}$$

by Cauchy-Schwartz, Appendix B, and Assumption B.1. Similarly,

$$\begin{aligned}
\ell(y|\boldsymbol{x},\boldsymbol{\beta})^2 &= \left(y\boldsymbol{x}^\intercal\boldsymbol{\beta} - \mathcal{A}(\boldsymbol{x}^\intercal\boldsymbol{\beta})\right)^2 \\
&= y^2\text{Trace}\left(\boldsymbol{x}\boldsymbol{x}^\intercal\boldsymbol{\beta}\boldsymbol{\beta}^\intercal\right) + \mathcal{A}(\boldsymbol{x}^\intercal\boldsymbol{\beta})^2 + 2\mathcal{A}(\boldsymbol{x}^\intercal\boldsymbol{\beta})y\boldsymbol{x}^\intercal\boldsymbol{\beta},
\end{aligned} \tag{35}$$

and so

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\ell(y|\boldsymbol{x},\boldsymbol{\beta})^2\right] \\
&\quad= \text{Trace}\left(\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[y^2\boldsymbol{x}\boldsymbol{x}^\intercal\right]\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\boldsymbol{\beta}\boldsymbol{\beta}^\intercal\right)+\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\mathcal{A}(\boldsymbol{x}^\intercal\boldsymbol{\beta})^2\right]+ \\
&\qquad 2\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\mathcal{A}(\boldsymbol{x}^\intercal\boldsymbol{\beta})y\boldsymbol{x}^\intercal\boldsymbol{\beta}\right] \\
&\quad\leq \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[y^2\left\|\boldsymbol{x}\right\|_2^2\right]\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\left\|\boldsymbol{\beta}\right\|_2^2+\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\mathcal{A}(\boldsymbol{x}^\intercal\boldsymbol{\beta})^2\right]+ \\
&\qquad 2\left(\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\mathcal{A}(\boldsymbol{x}^\intercal\boldsymbol{\beta})^2\right]\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[y^2\left\|\boldsymbol{x}\right\|_2^2\right]\right)^{1/2}\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\left\|\boldsymbol{\beta}\right\|_2<\infty,
\end{aligned}$$

again by Cauchy-Schwartz, Appendix B, and Assumption B.1.

The exchange of differentiation and integration now follows from Cauchy-Schwartz and the dominated convergence theorem, and for each $0\leq k\leq 4$ we can bound

$$\left\|\nabla^k_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})\right\|_2^2\leq\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta}\left\|\nabla^k_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})\right\|_2^2=:M_k(\boldsymbol{x},y),$$

where we have shown that $\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[M_k(\boldsymbol{x},y)\right]<\infty$.

□

Lemma B.2. *Let Assumption B.1 hold. With probability approaching one, there exists a $\gamma>0$ such that*

$$\sup_{\boldsymbol{\beta}\in\mathbb{R}^{D_{\boldsymbol{\beta}}}\setminus\mathcal{B}_\Delta}\left(\frac{1}{N_S}\sum_{i\in[N_S]}\ell(y_i|\boldsymbol{\beta}^*,\boldsymbol{x}_i)-\frac{1}{N_S}\sum_{i\in[N_S]}\ell(y_i|\boldsymbol{\beta},\boldsymbol{x}_i)\right)\geq\gamma>0.$$

That is, the empirical log likelihood is strictly bounded away from the value achieved at $\boldsymbol{\beta}^*$ outside of $\mathcal{B}_\Delta$.

Proof. Let $\partial\mathcal{B}_\Delta$ denote the boundary of the set $\mathcal{B}_\Delta$. By Assumption B.1, $\ell(\boldsymbol{\beta}^*)>\ell(\boldsymbol{\beta})$ for all $\boldsymbol{\beta}\in\partial\mathcal{B}_\Delta$ because $\boldsymbol{\beta}^*$ is a strict maximum. Define $\hat{\ell}_{N_S}(\boldsymbol{\beta}):=\frac{1}{N_S}\sum_{i\in[N_S]}\ell(y_i|\boldsymbol{\beta}^*,\boldsymbol{x}_i)$. By Lemma B.1, $\ell(y_i|\boldsymbol{\beta},\boldsymbol{x}_i)$ obeys a uniform law of large numbers (ULLN) in $\mathcal{B}_\Delta$, so that for sufficiently large N, with probability approaching one we have, for some $\gamma<0$,

$$\hat{\ell}_{N_S}(\boldsymbol{\beta}^*)-\sup_{\boldsymbol{\beta}\in\partial\mathcal{B}_\Delta}\hat{\ell}_{N_S}(\boldsymbol{\beta})>\gamma>0.$$

Recall that

$$\nabla^2_{\boldsymbol{\beta}}\hat{\ell}_{N_S}(\boldsymbol{\beta})=-\frac{1}{N_S}\sum_{i\in[N_S]}\nabla^2_\eta A(\boldsymbol{x}_i^\intercal\boldsymbol{\beta})\boldsymbol{x}_i^{\otimes 2}.$$

So, on $\mathbb{R}^{D_{\boldsymbol{\beta}}}$, $\frac{1}{N_S}\sum_{i\in[N_S]}\nabla^2_{\boldsymbol{\beta}}\hat{\ell}_{N_S}(\boldsymbol{\beta})$ is negative semi-definite, since $\nabla^2_\eta A(\boldsymbol{\beta}^\intercal\boldsymbol{x}_i)\succeq 0$ by standard properties of the given exponential family.

For any point $\boldsymbol{\beta}'' \in \mathbb{R}^{D_{\boldsymbol{\beta}}} \setminus \mathcal{B}_\Delta$, the line connecting $\boldsymbol{\beta}^*$ to $\boldsymbol{\beta}''$ must pass through some $\boldsymbol{\beta}' \in \partial\mathcal{B}_\Delta$. Since $\nabla^2_{\boldsymbol{\beta}}\hat{\ell}_{N_S}(\boldsymbol{\beta})$ is negative semidefinite, the slope along any line cannot decrease, and so we must have that

$$\hat{\ell}_{N_S}(\boldsymbol{\beta}^*) \geq \hat{\ell}_{N_S}(\boldsymbol{\beta}') + \gamma \geq \hat{\ell}_{N_S}(\boldsymbol{\beta}'') + \gamma,$$

from which the result follows.

□

Lemma B.3. *Under Assumption B.1, Lemmas B.1 and B.2 imply that our problem satisfies Assumptions 1 and 2 of Giordano and Broderick (2024) for the quantity of interest* $g(\boldsymbol{\beta}) := \frac{1}{N_T}\sum_{j\in[N_T]} \pi_j m(\boldsymbol{x}_j^\intercal \boldsymbol{\beta})$.

Additionally, the quantities $\mathcal{A}(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)$ *and* $\nabla^1_\eta A(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)$ *are third-order BCLT-okay as given by Definition 2 of Giordano and Broderick (2024)*

Proof. Each part of Assumption 1 of Giordano and Broderick (2024) is already satisfied directly in Assumption B.1 or by Lemmas B.1 and B.2. Note that we take Ω_θ to be the set $\{\theta : \mathcal{P}(\theta) > 0\}$.

For the remainder of the lemma, we must show that $g(\boldsymbol{\beta})$, $\ell(\boldsymbol{x}|\boldsymbol{\beta})$, $g(\boldsymbol{\beta})\ell(\boldsymbol{x}|\boldsymbol{\beta})$, $\mathcal{A}(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)$, and $\nabla^1_\eta A(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)$ are BCLT-okay, as given by the three items in Definition 2 of Giordano and Broderick (2024).

For item 1 (almost sure differentiability), the assumption follows from properties of exponential families.

For item 2 (order one average derivatives of prior expectations), we have assumed in Assumption B.1 that $\frac{1}{N_S}\sum_{i\in[N_S]} \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta})}\left[g(\boldsymbol{\beta})^2\right] < \infty$, and that we can apply the law of large numbers to the quantities

$$\frac{1}{N_S}\sum_{i\in[N_S]} \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta})}\left[g(\boldsymbol{\beta})^2 \ell(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})^2\right] \quad \text{and} \quad \frac{1}{N_S}\sum_{i\in[N_S]} \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta})}\left[\ell(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})^2\right].$$

Item 2 is satisfied for $\mathcal{A}(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)$, and $\nabla^1_\eta A(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)$ by the ULLN implied by Lemma B.1.

For item 3 (sample averages bounded in probability in $\mathcal{B}_\Delta$), the assumption follows from the ULLNs of Assumption B.1 and Lemma B.1, together with $\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta} g(\boldsymbol{\beta}) < \infty$ by continuity.

□

For the duration of this section, let $\hat{\boldsymbol{\beta}}$ denote the MAP $\text{argmax}_{\boldsymbol{\beta}}\, \mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})$. Note that earlier in the paper $\hat{\boldsymbol{\beta}}$ denoted the OLS coefficient, but we will overload that notation for the moment.

Lemma B.4. *Let Assumption 1 of Giordano and Broderick (2024) hold, and suppose that* $a_i(\boldsymbol{\beta})$, $b(\boldsymbol{\beta})$, *and* $a_i(\boldsymbol{\beta})b(\boldsymbol{\beta})$ *satisfy Assumption 2 of Giordano and Broderick (2024). Here,* $a_i(\boldsymbol{\beta})$ *may depend on datapoint* i *but* $b(\boldsymbol{\beta})$ *does not. Then*

$$N_S \text{Cov}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left(a_i(\boldsymbol{\beta}), b(\boldsymbol{\beta})\right) = \frac{1}{2}\nabla_{\boldsymbol{\beta}} a_i(\boldsymbol{\beta})\hat{\mathcal{I}}^{-1}\nabla_{\boldsymbol{\beta}} b(\boldsymbol{\beta}) + \tilde{O}(N_S^{-1})\mathcal{E}_i^{cov},$$

where $\frac{1}{N_S}\sum_{i\in[N_S]}\left(\mathcal{E}_i^{cov}\right)^2 = \tilde{\mathcal{O}}_p(1)$.

Proof. The proof simply states a general version of an argument from the proof of Giordano and Broderick (2024) Theorem 2.

Define

$$\begin{aligned}
\Delta_i^a &:= a_i(\hat{\boldsymbol{\beta}}) - \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left[a_i(\boldsymbol{\beta})\right] \\
\Delta^b &:= b(\hat{\boldsymbol{\beta}}) - \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left[b(\boldsymbol{\beta})\right] \\
\overline{a}_i(\boldsymbol{\beta}) &:= a_i(\boldsymbol{\beta}) - a_i(\hat{\boldsymbol{\beta}}) \\
\overline{b}(\boldsymbol{\beta}) &:= b(\boldsymbol{\beta}) - b(\hat{\boldsymbol{\beta}}) \\
\overline{ab}_i(\boldsymbol{\beta}) &:= \overline{a}_i(\boldsymbol{\beta})\overline{b}(\boldsymbol{\beta}).
\end{aligned}$$

We can then rewrite the covariance as

$$\begin{aligned}
\mathrm{Cov}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left(a_i(\boldsymbol{\beta}), b(\boldsymbol{\beta})\right) &= \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left[(\overline{a}_i(\boldsymbol{\beta}) + \Delta_i^a)(\overline{b}(\boldsymbol{\beta}) + \Delta^b)\right] \\
&= \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left[\overline{ab}_i(\boldsymbol{\beta})\right] + \Delta_i^a\Delta^b + \\
&\quad \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left[\overline{a}_i(\boldsymbol{\beta})\right]\Delta^b + \Delta^a\mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left[\overline{b}(\boldsymbol{\beta})\right] \\
&= \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left[\overline{ab}_i(\boldsymbol{\beta})\right] - \Delta_i^a\Delta^b.
\end{aligned}$$

By Theorem 1 of Giordano and Broderick (2024),

$$\Delta_i^a = \tilde{\mathcal{O}}\left(N_S^{-1}\right)\mathcal{E}_i^{a_i} \quad \text{and} \quad \Delta^b = \tilde{\mathcal{O}}\left(N_S^{-1}\right)\mathcal{E}^b,$$

where the residuals $\mathcal{E}_i^{a_i}$ and $\mathcal{E}^b$ combine leading-order and residuals terms, and satisfy $\frac{1}{N_S}\sum_{i\in[N_S]}\left(\mathcal{E}_i^a\right)^2 = \tilde{\mathcal{O}}_p\left(1\right)$ by the BCLT okay assumption and Giordano and Broderick (2024) Theorem 1.

Then, noting that $\nabla_{\boldsymbol{\beta}}\overline{ab}_i(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, another application of Giordano and Broderick (2024) Theorem 1 also gives that

$$N_S\mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left[\overline{ab}_i(\boldsymbol{\beta})\right] = \frac{1}{2}\nabla_{\boldsymbol{\beta}}a_i(\hat{\boldsymbol{\beta}})^\intercal\hat{\mathcal{I}}^{-1}\nabla_{\boldsymbol{\beta}}b(\hat{\boldsymbol{\beta}}) + \tilde{\mathcal{O}}\left(N_S^{-1}\right)\mathcal{E}_i^{ab},$$

where again $\mathcal{E}_i^{ab}$ is finitely square-summable with probability approaching one. Combining gives

$$N_S\mathrm{Cov}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left(a_i(\boldsymbol{\beta}), b(\boldsymbol{\beta})\right) = \frac{1}{2}\nabla_{\boldsymbol{\beta}}a_i(\hat{\boldsymbol{\beta}})^\intercal\hat{\mathcal{I}}^{-1}\nabla_{\boldsymbol{\beta}}b(\hat{\boldsymbol{\beta}}) + \tilde{\mathcal{O}}\left(N_S^{-1}\right)\left(\mathcal{E}_i^{ab} - \mathcal{E}_i^a\mathcal{E}^b\right).$$

Furthermore, setting $\mathcal{E}_i^{cov} := \mathcal{E}_i^{ab} - \mathcal{E}_i^a\mathcal{E}^b$, we have $\frac{1}{N_S}\sum_{i\in[N_S]}\left(\mathcal{E}_i^{cov}\right)^2 = \tilde{\mathcal{O}}_p\left(1\right)$ by Cauchy-Schwartz. □

Proof of Theorem 4.1. Let ψ_i denote the classical empirical influence function for $\hat{\mu}^{\mathrm{MrP}}$ as defined in Giordano and Broderick (2024),

$$\psi_i := N_S\mathrm{Cov}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left(g(\boldsymbol{\beta}), \ell(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})\right).$$

Recall also from the statement of Theorem 4.1 that

$$N_S w_i^{\mathrm{MrP}}\varepsilon_i = N_S\mathrm{Cov}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left(g(\boldsymbol{\beta}), \boldsymbol{x}_i^\intercal\boldsymbol{\beta}\right)(y_i - \hat{y}_i),$$

where

$$\hat{y}_i = \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left[m(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)\right] .$$

The proof will show that $\psi_i \approx N_S w_i^{\text{MrP}} \varepsilon_i$ to a sufficiently high degree of accuracy, and then to apply Giordano and Broderick (2024) Theorem 2.

Define $\nabla_{\boldsymbol{\beta}} g(\boldsymbol{\beta})$ the derviative of $g(\boldsymbol{\beta})$. By Theorem 1 of Giordano and Broderick (2024)

$$N_S w_i^{\text{MrP}} \varepsilon_i - \psi_i = N_S \text{Cov}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left(g(\boldsymbol{\beta}), A(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i) - \hat{y}_i \boldsymbol{\beta}^\intercal \boldsymbol{x}_i\right)$$

By Lemma B.4, we can write

$$\begin{aligned} &N_S \text{Cov}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left(g(\boldsymbol{\beta}), A(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i) - \hat{y}_i \boldsymbol{\beta}^\intercal \boldsymbol{x}_i\right) = \\ &\quad \nabla_{\boldsymbol{\beta}} g(\hat{\boldsymbol{\beta}})^\intercal \left(\nabla_\eta^1 A(\hat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i - \hat{y}_i \boldsymbol{x}_i\right) + \tilde{\mathcal{O}}_p\left(N^{-1}\right) \mathcal{E}_i^{cov}. \end{aligned}$$

Additionally, by Theorem 1 of Giordano and Broderick (2024),

$$\begin{aligned} \hat{y}_i &= \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}\left[m(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)\right] \\ &= m(\hat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_j) + \tilde{\mathcal{O}}_p\left(N^{-1}\right) \mathcal{E}_i^y \\ &= \nabla_\eta^1 A(\hat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_j) + \tilde{\mathcal{O}}_p\left(N^{-1}\right) \mathcal{E}_i^y, \end{aligned}$$

where the final line follows from the exponential family fact that $\nabla_\eta^1 A(\cdot) = m(\cdot)$. It follows that

$$\nabla_\eta^1 A(\hat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_i)\boldsymbol{x}_i - \hat{y}_i \boldsymbol{x}_i = \tilde{\mathcal{O}}_p\left(N^{-1}\right) \mathcal{E}_i^y \boldsymbol{x}_i$$

Combining,

$$\begin{aligned} N_S w_i^{\text{MrP}} \varepsilon_i - \psi_i &= \tilde{\mathcal{O}}\left(N^{-1}\right)\left(\mathcal{E}_i^{cov} + \mathcal{E}_i^y \nabla_{\boldsymbol{\beta}} g(\hat{\boldsymbol{\beta}})^\intercal \boldsymbol{x}_i\right) \\ &=: \tilde{\mathcal{O}}\left(N^{-1}\right) \mathcal{E}_i^{err}. \end{aligned}$$

From this it follows that

$$\begin{aligned} \overline{w^{\text{MrP}}\varepsilon} &= \frac{1}{N_S} \sum_{i \in [N_S]} N_S w_i^{\text{MrP}} \varepsilon_i \\ &= \overline{\psi} + \tilde{\mathcal{O}}\left(N^{-1}\right) \frac{1}{N_S} \sum_{i \in [N_S]} \mathcal{E}_i^{err} \\ \frac{1}{N_S} \sum_{i \in [N_S]} \left(N_S w_i^{\text{MrP}} \varepsilon_i\right)^2 &= \frac{1}{N_S} \sum_{i \in [N_S]} \psi_i^2 + \\ &\quad 2\tilde{\mathcal{O}}\left(N^{-1}\right) \frac{1}{N_S} \sum_{i \in [N_S]} \mathcal{E}_i^{err} \psi_i + \tilde{\mathcal{O}}\left(N^{-2}\right) \frac{1}{N_S} \sum_{i \in [N_S]} (\mathcal{E}_i^{err})^2 \end{aligned}$$

and so

$$\frac{1}{N_S} \sum_{i \in [N_S]} \left(N_S w_i^{\text{MrP}} \varepsilon_i - N_S \overline{w^{\text{MrP}}\varepsilon}\right)^2 = \frac{1}{N_S} \sum_{i \in [N_S]} \left(\psi_i - \overline{\psi}\right)^2 + \tilde{\mathcal{O}}_p\left(N^{-1}\right).$$

The final result follows from Theorem 2 of Giordano and Broderick (2024), which states that the right hand side of the preceding display is a consistent estimate of $\boldsymbol{V}$. □

C Proof of Theorem 4.2

For any r, take $\boldsymbol{Y}(\delta r) = \boldsymbol{Y} + \delta \boldsymbol{R}$.

$$\hat{\mu}^{\mathrm{MrP}}(\tilde{\boldsymbol{Y}}) - \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}) = \left.\frac{\partial \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}(\delta))}{\partial \delta}\right|_{\delta=0} (\delta - 0) + \frac{1}{2} \left.\frac{\partial^2 \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}(\delta r))}{\partial \delta^2}\right|_{\delta=\tilde{\delta}} (\delta - 0)^2$$

for some $\tilde{\delta} \in [0, \delta]$. The first term is given by

$$\left.\frac{\partial \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}(\delta r))}{\partial \delta}\right|_{\delta=0} = \sum_{i \in [N_S]} \frac{\partial \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y})}{\partial y_i} r_i = \delta \frac{1}{N_S} \sum_{i \in [N_S]} w_i^{\mathrm{MrP}} r_i.$$

The second term is given by

$$\left.\frac{\partial^2 \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}(\delta r))}{\partial \delta^2}\right|_{\delta=\tilde{\delta}} = \left.\frac{\partial^2 \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\delta r)}\left[g(\boldsymbol{\beta})\right]}{\partial \delta^2}\right|_{\delta=\tilde{\delta}}.$$

So we would like to control the error

$$\begin{aligned} &\sup_{r \in \mathcal{R}} \frac{1}{\delta} \left| \hat{\mu}^{\mathrm{MrP}}(\tilde{\boldsymbol{Y}}) - \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y}) - \delta \frac{1}{N_S} \sum_{i \in [N_S]} w_i^{\mathrm{MrP}} r_i \right| \leq \\ &\qquad \delta \sup_{r \in \mathcal{R}} \sup_{\tilde{\delta} \in [0,\delta]} \frac{1}{2} \left| \left.\frac{\partial^2 \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\delta r)}\left[g(\boldsymbol{\beta})\right]}{\partial \delta^2}\right|_{\delta=\tilde{\delta}} \right|. \end{aligned} \tag{36}$$

Let $\mathcal{K}_{\mathcal{P}(\boldsymbol{\beta})}(\cdot)$ denotes the third–order cumulant of $\mathcal{P}(\boldsymbol{\beta})$,

$$\begin{aligned} &\mathcal{K}_{\mathcal{P}(\boldsymbol{\beta})}\left(a(\boldsymbol{\beta}), b(\boldsymbol{\beta}), c(\boldsymbol{\beta})\right) := \\ &\quad \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta})}\left[(a(\boldsymbol{\beta}) - \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta})}\left[a(\boldsymbol{\beta})\right]) (b(\boldsymbol{\beta}) - \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta})}\left[b(\boldsymbol{\beta})\right]) (c(\boldsymbol{\beta}) - \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta})}\left[c(\boldsymbol{\beta})\right]) \right]. \end{aligned}$$

Then, for the right hand side of eq. (36), note that

$$\begin{aligned} \left.\frac{\partial^2 \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\delta r)}\left[g(\boldsymbol{\beta})\right]}{\partial \delta^2}\right|_{\delta} &= \mathcal{K}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\delta r)}\left(g(\boldsymbol{\beta}), \left.\frac{\partial \ell(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}; \delta r)}{\partial \delta}\right|_{\delta}, \left.\frac{\partial \ell(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}; \delta r)}{\partial \delta}\right|_{\delta} \right) \\ &= \mathcal{K}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\delta r)}\left(g(\boldsymbol{\beta}), \sum_{i \in [N_S]} r_i \boldsymbol{x}_i^{\intercal} \boldsymbol{\beta}, \sum_{i \in [N_S]} r_i \boldsymbol{x}_i^{\intercal} \boldsymbol{\beta} \right) \\ &= \sum_{i \in [N_S]} \sum_{i' \in [N_S]} r_{i'} \boldsymbol{x}_{i'}^{\intercal} \mathcal{K}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\delta r)}\left(g(\boldsymbol{\beta}), \boldsymbol{\beta}, \boldsymbol{\beta}\right) \boldsymbol{x}_i r_i. \end{aligned}$$

Plugging in,

$$\sup_{r\in\mathcal{R}}\sup_{\tilde{\delta}\in[0,\delta]}\left|\frac{\partial^2\mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\delta r)}\left[g(\boldsymbol{\beta})\right]}{\partial\delta^2}\bigg|_{\delta=\tilde{\delta}}\right| =$$

$$\sup_{r\in\mathcal{R}}\sup_{\tilde{\delta}\in[0,\delta]}\left|\sum_{i\in[N_S]}\sum_{i'\in[N_S]} r_{i'}\boldsymbol{x}_{i'}^{\intercal}\mathcal{K}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\tilde{\delta}r)}\left(g(\boldsymbol{\beta}),\boldsymbol{\beta},\boldsymbol{\beta}\right)\boldsymbol{x}_i r_i\right| \leq$$

$$\sup_{r\in\mathcal{R}}\sup_{\tilde{\delta}\in[0,\delta]} N_S^2\left\|\mathcal{K}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\tilde{\delta}r)}\left(g(\boldsymbol{\beta}),\boldsymbol{\beta},\boldsymbol{\beta}\right)\right\|_2\left\|\frac{1}{N_S^2}\sum_{i\in[N_S]}\boldsymbol{x}_i^{\intercal}r_i\sum_{i'\in[N_S]}\boldsymbol{x}_{i'}r_{i'}\right\|_2 =$$

$$\sup_{r\in\mathcal{R}}\sup_{\tilde{\delta}\in[0,\delta]} N_S^2\left\|\mathcal{K}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\tilde{\delta}r)}\left(g(\boldsymbol{\beta}),\boldsymbol{\beta},\boldsymbol{\beta}\right)\right\|_2\left\|\frac{1}{N_S}\sum_{i\in[N_S]}\boldsymbol{x}_i r_i\right\|_2^2 \leq$$

$$\sup_{r\in\mathcal{R}}\sup_{\tilde{\delta}\in[0,\delta]} N_S^2\left\|\mathcal{K}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\tilde{\delta}r)}\left(g(\boldsymbol{\beta}),\boldsymbol{\beta},\boldsymbol{\beta}\right)\right\|_2\frac{1}{N_S}\sum_{i\in[N_S]}\|\boldsymbol{x}_i r_i\|_2^2 \leq$$

$$\sup_{r\in\mathcal{R}}\sup_{\tilde{\delta}\in[0,\delta]} N_S^2\left\|\mathcal{K}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\tilde{\delta}r)}\left(g(\boldsymbol{\beta}),\boldsymbol{\beta},\boldsymbol{\beta}\right)\right\|_2\sup_{r\in\mathcal{R}}\frac{1}{N_S}\sum_{i\in[N_S]}\|\boldsymbol{x}_i r_i\|_2^2.$$

Here, since $\mathcal{R}$ is Donkser and satisfies $\sup_{r\in\mathcal{R}}\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\boldsymbol{x}r(\boldsymbol{x})\right]_2^2 < \infty$ by Assumption 4.4, by a uniform law of large numbers we have

$$\sup_{r\in\mathcal{R}}\frac{1}{N_S}\sum_{i\in[N_S]}\|\boldsymbol{x}_i r_i\|_2^2 \to \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\|\boldsymbol{x}_i r_i\|_2^2\right] \leq \mathcal{R}_{\max}^2.$$

Therefore it will suffice to show that

$$\sup_{r\in\mathcal{R}}\sup_{\tilde{\delta}\in[0,\delta]}\left\|\mathcal{K}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\tilde{\delta}r)}\left(g(\boldsymbol{\beta}),\boldsymbol{\beta},\boldsymbol{\beta}\right)\right\|_2 = \tilde{\mathcal{O}}_p\left(N_S^{-2}\right).$$

To show this, we need to establish a version of Theorem 1 of Giordano and Broderick (2024) that holds uniformly over δ and r.

C.1 Uniform posterior expansions

Assumption C.1. Define a parameterized log likelihood function $\ell(\boldsymbol{\beta}|y,t)$, for parameter $t \in \Omega_t$. Let Assumption 1 of Giordano and Broderick (2024) hold for each $t \in \Omega_t$ with constants depending on t. Additionally assume that items 4, 5, and 6 of Assumption 1 hold uniformly in t in the following sense:

- **Item 4**: The log likelihood $\ell(\boldsymbol{\beta}|y,t)$ and its first four partial derivatives are each uniformly continuous over $t \in \Omega_t$.
- **Item 5**: The bound $M(y;t)$ from Item 5 holds uniformly in the sense that $\mathbb{E}_{\mathcal{P}_S(y)}\left[\sup_{t\in\Omega_t} M(y;t)^2\right] < \infty$.
- **Item 6 modification 1**: Letting $\lambda_{\mathcal{I},t}$ denote the minimum eigenvalue of $\mathcal{I}_t$, assume that $\inf_{t\in\Omega_t}\lambda_{\mathcal{I},t} > 0$

- **Item 6 modification 2**: The empirical log likelihood bound satisfies $\inf_{t\in\Omega_t} \varepsilon(t) > 0$.

□

Lemma C.1. *Let Assumption C.1 hold. Let $\phi(\boldsymbol{\beta})$ be a function not depending on t or y that is third-order BCLT okay ((Giordano and Broderick, 2024) Definition 2). Then Giordano and Broderick (2024) Theorem 1 holds uniformly over $t \in \Omega_t$ in the following sense. For any target probability $0 < \rho < 1$, there exists C^* and N^* not depending on t for which $N_S > N^*$ implies that*

$$\sup_{t\in\Omega_t} \Bigg| \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y},t)}\left[\phi(\boldsymbol{\beta})\right] - \phi(\hat{\boldsymbol{\beta}}_t) - N_S^{-1}\left(\frac{1}{2}\nabla^2_{\boldsymbol{\beta}}\phi(\hat{\boldsymbol{\beta}}_t)\hat{\mathcal{I}}_t^{-1} + \frac{1}{6}\nabla_{\boldsymbol{\beta}}\phi(\hat{\boldsymbol{\beta}}_t)\nabla^3_{\boldsymbol{\beta}}\hat{\mathcal{L}}_t(\hat{\boldsymbol{\beta}}_t)\hat{\mathcal{M}}\right)\Bigg| \leq N_S^{-2}C^* \tag{37}$$

with probability at least ρ.

Proof. By assumption, eq. (37) holds for each $t \in \Omega_t$ by Giordano and Broderick (2024) Theorem 1 with some C_t^* and N_t^* depending on t. We now proceed step-by-step through the proof of Giordano and Broderick (2024) Theorem 1 and show that the constant C_t^* and N_t^* depend only on the quantities uniformly controlled by assumption, from which we have

$$C^* := \sup_{t\in\Omega_t} C_t^* < \infty \quad \text{and} \quad N^* := \sup_{t\in\Omega_t} N_t^* < \infty.$$

We proceed lemma by lemma, indicating how the proof needs to be modified slightly in order to achieve uniform control over $t \in \Omega_t$.

Lemma 2. We replace the assumption of Lemma 2 with

$$\sup_{t\in\Omega_t}\sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta} \left\|\nabla^k_{\boldsymbol{\beta}}\ell(y|\boldsymbol{\beta})\right\|_2 \leq \sup_{t\in\Omega_t} M(y;t) := M(y) \quad \text{with} \quad \mathbb{E}_{\mathcal{P}_S(y)}\left[M(y)^2\right] < \infty,$$

from which the conclusions of Lemma 2 (eqs. 20,21, and 22) apply uniformly in $t \in \Omega_t$.

Lemma 3. In Lemma 3, replace ε with $\inf_{t\in\Omega_t}\varepsilon_t$ and $\lambda_{\mathcal{I}}$ with $\inf_{t\in\Omega_t}\lambda_{\mathcal{I},t}$. Then by the same proof, the conclusion of Lemma 3 holds uniformly in t in the sense that

$$\sup_{t\in\Omega_t}\left\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_{\infty,t}\right\|_2 \to 0 \quad \text{and} \quad \inf_{t\in\Omega_t}\left\|\hat{\mathcal{I}}_t\right\|_{op} \geq 2\inf_{t\in\Omega_t}\lambda_{\mathcal{I},t} := \lambda_{\mathcal{I}} > 0,$$

both in probability as $N_S \to \infty$.

Lemma 4. Lemma 4 is essentially a convenient re-statement of lemmas 2 and 3. It follows that, for any $0 < \rho < 1$, there exists an N^* not depending on t such that each event of Lemma 4 holds for all $t \in \Omega_t$ with probability at least ρ when $N_S > N^*$. Specifically, items 2, 3, and 4 hold with the corresponding inequalities and $\sup_{t\in\Omega_t}$, and items 5 and 6 hold with the corresponding inequalities and $\inf_{t\in\Omega_t}$. There is one caveat — when the function given in Item 4 of Lemma 4 depends on t, so that the corresponding neighborhood $\delta_{LLN,t}$ depends on t, we must have $\inf_t \delta_{LLN,t} > 0$ for each ϵ_U.

Lemma 5. With the above modifications, the upper bound Lemma 5 holds uniformly in Ω_t.

Lemma 6. The quantities δ_2 and ϵ_U given in Lemma 6 depend only on the occurrence of the events given in Lemma 4, and being able to choose δ_2 small enough to make the following quantities small:

$$\sup_{t\in\Omega_t}\sup_{\boldsymbol{\beta}\in R_{2,t}}\left\|\nabla^3_{\boldsymbol{\beta}}\mathcal{L}_t(\boldsymbol{\beta})\right\|_2 \quad \text{and} \quad \sup_{t\in\Omega_t}\sup_{\boldsymbol{\beta}\in R_{2,t}}\left\|\nabla^4_{\boldsymbol{\beta}}\mathcal{L}_t(\boldsymbol{\beta})\right\|_2.$$

Since, by assumption, $\nabla^k_{\boldsymbol{\beta}}\mathcal{L}_t(\boldsymbol{\beta})$ is uniformly continuous over $t\in\Omega_t$ for $k=3,4$, δ_2 can be chosen to make the expressions in the preceding display to satisfy Lemma 6 uniformly in $t\in\Omega_t$.

Lemma 7 simply relies on Lemma 4.

Lemma 8 follows from Lemma 4, again using uniform continuity of $\nabla^3_{\boldsymbol{\beta}}\mathcal{L}_t(\boldsymbol{\beta})$ and $\nabla^4_{\boldsymbol{\beta}}\mathcal{L}_t(\boldsymbol{\beta})$.

Lemmas 9, 10, and 11 mostly rearrange terms, relying again on Lemma 4.

The remainder of the proof does not require modification, since we don't need to consider data dependence in $\phi(\boldsymbol{\beta})$.

□

Lemma C.2. *Let Assumption C.1 hold. Consider the posterior cumulant $\mathcal{K}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y},t)}\left(a(\boldsymbol{\beta}),b(\boldsymbol{\beta}),c(\boldsymbol{\beta})\right)$, where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ do not depend on t. Assume that each of the functions $a(\boldsymbol{\beta})$, $b(\boldsymbol{\beta})$, $c(\boldsymbol{\beta})$, their pairwise products $a(\boldsymbol{\beta})b(\boldsymbol{\beta})$, $a(\boldsymbol{\beta})c(\boldsymbol{\beta})$, and $c(\boldsymbol{\beta})$, and the three-way product $a(\boldsymbol{\beta})b(\boldsymbol{\beta})c(\boldsymbol{\beta})$ are all third-order BCLT okay ((Giordano and Broderick, 2024) Definition 2).*

Then, for any probability $0<\rho<1$ there exists an N^* and C^* such that $N_S > N^*$ implies that

$$N_S^2 \sup_{t\in\Omega_t}\left|\mathcal{K}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y},t)}\left(a(\boldsymbol{\beta}),b(\boldsymbol{\beta}),c(\boldsymbol{\beta})\right)\right| \le C^*$$

with probability at least ρ.

Proof. First, for each of a, b, and c, define

$$\begin{aligned}\Delta^a &:= a(\hat{\boldsymbol{\beta}}_t) - \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y},t)}\left[a(\boldsymbol{\beta})\right]\\ \overline{a}\left(\boldsymbol{\beta}\right) &:= a(\boldsymbol{\beta}) - a(\hat{\boldsymbol{\beta}}_t),\end{aligned}$$

and so on. Note that $\mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})t}\left[\overline{a}\right] = -\Delta^a$. Then

$$\begin{aligned}&\mathcal{K}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y},t)}\left(a(\boldsymbol{\beta}),b(\boldsymbol{\beta}),c(\boldsymbol{\beta})\right) =\\ &\quad\mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y},t)}\left[\left(\overline{a}+\Delta^a\right)\left(\overline{b}+\Delta^b\right)\left(\overline{c}+\Delta^c\right)\right] =\\ &\quad\mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y},t)}\left[\overline{a}\,\overline{b}\,\overline{c}\right] +\\ &\quad\mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y},t)}\left[\overline{a}\,\overline{b}\right]\Delta^c + \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y},t)}\left[\overline{a}\,\overline{c}\right]\Delta^b + \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y},t)}\left[\overline{b}\,\overline{c}\right]\Delta^a - 2\Delta^a\Delta^b\Delta^c.\end{aligned}$$

Let us first consider $N_S\Delta^a$. Applying Lemma C.1 with constant C^a_Δ,

$$-N_S^{-1}C^a_\Delta - \frac{1}{2}\nabla^2_{\boldsymbol{\beta}}\phi(\hat{\boldsymbol{\beta}}_t)\hat{\mathcal{I}}_t^{-1} - \frac{1}{6}\nabla_{\boldsymbol{\beta}}\phi(\hat{\boldsymbol{\beta}}_t)\nabla^3_{\boldsymbol{\beta}}\hat{\mathcal{L}}_t(\hat{\boldsymbol{\beta}}_t)\hat{\mathcal{M}} \leq N_S\Delta^a \leq N_S^{-1}C^a_\Delta - \frac{1}{2}\nabla^2_{\boldsymbol{\beta}}\phi(\hat{\boldsymbol{\beta}}_t)\hat{\mathcal{I}}_t^{-1} - \frac{1}{6}\nabla_{\boldsymbol{\beta}}\phi(\hat{\boldsymbol{\beta}}_t)\nabla^3_{\boldsymbol{\beta}}\hat{\mathcal{L}}_t(\hat{\boldsymbol{\beta}}_t)\hat{\mathcal{M}}.$$

By the modification Giordano and Broderick (2024) Lemma 3 given in the proof of Lemma C.1, we have that $\sup_t \left\|\hat{\boldsymbol{\beta}}_t\right\|_2$. From this, and continuity of ϕ and its derivatives, we have

$$\sup_{t\in\Omega_t}\left\|\phi(\hat{\boldsymbol{\beta}}_t)\right\| = \tilde{\mathcal{O}}_p(1) \quad , \quad \sup_{t\in\Omega_t}\left\|\nabla_{\boldsymbol{\beta}}\phi(\hat{\boldsymbol{\beta}}_t)\right\| = \tilde{\mathcal{O}}_p(1) \quad , \quad \sup_{t\in\Omega_t}\left\|\hat{\mathcal{I}}_t^{-1}\right\|_{op} = \tilde{\mathcal{O}}_p(1).$$

Similarly, by the modification of Lemma 2 in Lemma C.1, together with Assumption C.1 giving uniform continuity of $\mathcal{L}_t$,

$$\sup_{t\in\Omega_t}\hat{\mathcal{L}}_t(\hat{\boldsymbol{\beta}}_t) = \tilde{\mathcal{O}}_p(1).$$

It follows that $\sup_{t\in\Omega_t}\left|N_S^{-1}\Delta^a\right| = \tilde{\mathcal{O}}_p(1)$. Analogous results hold for Δ^b and Δ^c.

Similarly, note that $\overline{a}(\hat{\boldsymbol{\beta}}_t)\overline{b}(\hat{\boldsymbol{\beta}}_t) = 0$, and $\left.\frac{\partial\overline{a}(\boldsymbol{\beta})\overline{b}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_t} = 0$, from which we have

$$N_S^2\sup_{t\in\Omega_t}\left|\overline{a}\,\overline{b}\right| = \tilde{\mathcal{O}}_p(1),$$

with analogous results for $\overline{a}\,\overline{c}$ and $\overline{b}\,\overline{c}$.

Finally, $\overline{a}(\hat{\boldsymbol{\beta}}_t)\overline{b}(\hat{\boldsymbol{\beta}}_t)\overline{c}(\hat{\boldsymbol{\beta}}_t) = \mathbf{0}$, and

$$\left.\frac{\partial\overline{a}(\boldsymbol{\beta})\overline{b}(\boldsymbol{\beta})\overline{c}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_t} = \mathbf{0}$$

so that

$$N_S^2\sup_{t\in\Omega_t}\left|\overline{a}\,\overline{b}\,\overline{c}\right| = \tilde{\mathcal{O}}_p(1).$$

Combining gives the desired result. □

C.2 Showing that the regressor balance satisfies the conditions

By Lemma C.2, it only remains to show that $\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y};\delta r)$ satisfies Assumption C.1, where $t = (\delta, r)$ and $\Omega_t = (0,\delta_+)\times\mathcal{R}$ for some sufficiently small δ_+.

Lemma C.3. *Let Assumption B.1 hold for the original model, and let $r \in \mathcal{R}$ (Assumption 4.4). Recalling Definition 4.1, define*

$$\ell(\boldsymbol{\beta};\delta r) := \mathbb{E}_{\mathcal{P}_S(y,\boldsymbol{x})}\left[\ell(y|\boldsymbol{x},\boldsymbol{\beta},\delta r)\right].$$

Then $\ell(\boldsymbol{\beta}; \delta r)$ satisfies Assumption 4.2 uniformly over $\delta, r \in (0, \delta_+) \times \mathcal{R}$ in the following sense. Define

$$\boldsymbol{\beta}^*_{\delta r} := \operatorname{argmax}_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \delta r) \quad \textit{and} \quad \mathcal{I}_{\delta r} := \left. \frac{\ell(\boldsymbol{\beta}; \delta r)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\intercal} \right|_{\boldsymbol{\beta}^*_{\delta r}}.$$

Let $\lambda_{\delta r}$ denote the minimum eigenvalue of $\mathcal{I}_{\delta r}$. Then there exists a δ_+ such that $\boldsymbol{\beta}^*_{\delta r}$ is unique for all $\delta, r \in (0, \delta_+) \times \mathcal{R}$, and that

$$\inf_{\delta, r \in (0, \delta_+) \times \mathcal{R}} \lambda_{\mathcal{I}, \delta r} \geq \lambda_{\mathcal{I}} > 0.$$

Furthermore, $\ell(\boldsymbol{\beta}; \delta r)$ and its derivatives are uniformly continuous as a function of $\boldsymbol{\beta}$ for $\delta, r \in (0, \delta_+) \times \mathcal{R}$.

Expanding,

$$\ell(\boldsymbol{\beta}; \delta r) = \ell(y | \boldsymbol{x}, \boldsymbol{\beta}) + \delta \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})} \left[r(\boldsymbol{x}) \boldsymbol{x}^\intercal \right] \boldsymbol{\beta}.$$

Since $\sup_{\delta, r \in (0, \delta_+) \times \mathcal{R}} \left| \delta \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})} \left[r(\boldsymbol{x}) \boldsymbol{x}^\intercal \right] \right| \leq \delta_+ \mathcal{R}_{\max}$ by Assumption 4.4, $\ell(\boldsymbol{\beta}; \delta r)$ and its derivatives are uniformly continuous as a function of $\boldsymbol{\beta}$ for $\delta, r \in (0, \delta_+) \times \mathcal{R}$ by Lemma B.1 (via the dominated convergence theorem).

We now adapt Lemma B.1 to $\ell(\boldsymbol{\beta}; \delta r)$. First, for $k \geq 2$, $\nabla^k_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \delta r) = \nabla^k_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$, so eq. (34) holds without modification for $k \geq 2$.

We now adapt the first and zeroth derivatives. Note that by Jensen's inequality,

$$\sup_{r \in \mathcal{R}} \left\| \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})} \left[\boldsymbol{x} r(\boldsymbol{x}) \right] \right\|_2^2 \leq \sup_{r \in \mathcal{R}} \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})} \left[\left\| \boldsymbol{x} r(\boldsymbol{x}) \right\|_2^2 \right] \leq \mathcal{R}^2_{\max}. \tag{38}$$

Similarly, for $f(\boldsymbol{\beta}) \geq 0$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x}, y)} \left[\sup_{\boldsymbol{\beta} \in \mathcal{B}_\Delta} f(\boldsymbol{\beta}) \right] = \\
&\quad \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x}, y)} \left[\sup_{\boldsymbol{\beta} \in \mathcal{B}_\Delta} f(\boldsymbol{\beta}) (\mathbb{I}\left(f(\boldsymbol{\beta}) \geq 1\right) + \mathbb{I}\left(f(\boldsymbol{\beta}) < 1\right)) \right] \\
&\leq \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x}, y)} \left[\sup_{\boldsymbol{\beta} \in \mathcal{B}_\Delta} f(\boldsymbol{\beta})^2 \mathbb{I}\left(f(\boldsymbol{\beta}) \geq 1\right) \right] + \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x}, y)} \left[\sup_{\boldsymbol{\beta} \in \mathcal{B}_\Delta} 1 \mathbb{I}\left(f(\boldsymbol{\beta}) < 1\right) \right] \\
&\leq \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x}, y)} \left[\sup_{\boldsymbol{\beta} \in \mathcal{B}_\Delta} f(\boldsymbol{\beta})^2 \right] + 1.
\end{aligned} \tag{39}$$

Then

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \delta r) &= \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) + \delta \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})} \left[r(\boldsymbol{x}) \boldsymbol{x}^\intercal \right] \Rightarrow \\
\left\| \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \delta r) \right\|_2^2 &= \left\| \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) \right\|_2^2 + \delta^2 \left\| \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})} \left[r(\boldsymbol{x}) \boldsymbol{x}^\intercal \right] \right\|_2^2 + 2\delta \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})} \left[r(\boldsymbol{x}) \boldsymbol{x}^\intercal \right] \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) \\
&\leq \left\| \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) \right\|_2^2 + \delta_+^2 \mathcal{R}^2_{\max} + 2\delta_+ \mathcal{R}_{\max} \left\| \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) \right\|_2.
\end{aligned}$$

The right hand side of the preceding display does not depend on r or δ, and has finite expectation under

$\mathcal{P}_S(\boldsymbol{x}, y)$ by Lemma B.1 and eqs. (38) and (39), and so we have identified $M_1'(\boldsymbol{x}, y)$ such that

$$\sup_{\delta, r \in (0,\delta_+)\times\mathcal{R}} \sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta} \left\|\nabla^k_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \delta r)\right\|_2^2 \leq \tilde{M}_1(\boldsymbol{x}, y) \quad \text{and} \quad \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[\tilde{M}_1(\boldsymbol{x}, y)\right] < \infty.$$

Similarly,

$$\begin{aligned}\ell(\boldsymbol{\beta}; \delta r)^2 &= \ell(\boldsymbol{\beta})^2 + \delta^2 \text{Trace}\left(\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[r(\boldsymbol{x})\boldsymbol{x}\right] \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[r(\boldsymbol{x})\boldsymbol{x}^\intercal\right] \boldsymbol{\beta}\boldsymbol{\beta}^\intercal\right) + \\ &\quad 2\delta\ell(\boldsymbol{\beta})\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[r(\boldsymbol{x})\boldsymbol{x}^\intercal\right] \boldsymbol{\beta} \\ &\leq \ell(\boldsymbol{\beta})^2 + \delta_+^2 \mathcal{R}_{\max}^2 \left\|\boldsymbol{\beta}\right\|_2^2 + 2\ell(\boldsymbol{\beta})\mathcal{R}_{\max}\left\|\boldsymbol{\beta}\right\|.\end{aligned}$$

As before, we have identified $\tilde{M}(\boldsymbol{x}, y)$ not depending on r or δ satisfying

$$\sup_{\delta, r \in (0,\delta_+)\times\mathcal{R}} \sup_{\boldsymbol{\beta}\in\mathcal{B}_\Delta} \ell(\boldsymbol{\beta}; \delta r)_2^2 \leq \tilde{M}(\boldsymbol{x}, y) \quad \text{and} \quad \mathbb{E}_{\mathcal{P}_S(\boldsymbol{x},y)}\left[\tilde{M}(\boldsymbol{x}, y)\right] < \infty.$$

We now show that we can control $\inf_{\delta, r\in(0,\delta_+)\times\mathcal{R}} \lambda_{\delta r}$. Let $\lambda_{\min}(\cdot)$ denote the minimum eigenvalue of the argument. Now, we have that

$$\lambda_{\delta r} = \lambda_{\min}\left(\mathcal{I}_{\delta r}\right) = \lambda_{\min}\left(\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\nabla^2_\eta A(\boldsymbol{x}^\intercal\boldsymbol{\beta}^*_{\delta r})\boldsymbol{x}^{\otimes 2}\right]\right).$$

Since by Assumption B.1, $\lambda_{\min}(\mathcal{I}) = \lambda_{0\cdot r} > 0$, and by Lemma B.1 the map

$$\boldsymbol{\beta} \mapsto \lambda_{\min}\left(\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\nabla^2_\eta A(\boldsymbol{x}^\intercal\boldsymbol{\beta})\boldsymbol{x}^{\otimes 2}\right]\right)$$

is continuous at $\boldsymbol{\beta}^*$, we can choose $\Delta > 0$ small enough that

$$\boldsymbol{\beta} \in \mathcal{B}_\Delta \quad \Rightarrow \quad \lambda_{\min}\left(\mathbb{E}_{\mathcal{P}_S(\boldsymbol{x})}\left[\nabla^2_\eta A(\boldsymbol{x}^\intercal\boldsymbol{\beta})\boldsymbol{x}^{\otimes 2}\right]\right) \geq \frac{1}{2}\lambda_{\mathcal{I}} > 0.$$

So it suffices to take such a Δ and show that there exists δ_+ sufficiently small that

$$\boldsymbol{\beta}^*_{\delta r} \in \mathcal{B}_\Delta \quad \text{for all} \quad \delta, r \in [0, \delta_+].$$

Since $\ell(\boldsymbol{\beta}; \delta r)$ is continuously differentiable, $\boldsymbol{\beta}^*$ is a solution to

$$\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^*_{\delta r}; \delta r) = \nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^*_{\delta r}) + \delta\mathbb{E}_{\mathcal{P}_S}\left[r(\boldsymbol{x})\boldsymbol{x}\right] = \mathbf{0}.$$

Note that $\left\|\delta\mathbb{E}_{\mathcal{P}_S}\left[r(\boldsymbol{x})\boldsymbol{x}\right]\right\|_2 \leq \delta_+\mathcal{R}_{\max}$ by Assumption 4.4, so we can define $\boldsymbol{\beta}^*_{\mathbf{v}}$ as the solution to

$$\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^*_{\mathbf{v}}) + \mathbf{v} = \mathbf{0} \text{ for } \left\|\mathbf{v}\right\|_2 \leq \delta_+\mathcal{R}_{\max}.$$

Since $\mathbf{v} = \mathbf{0}$ corresponds to $\boldsymbol{\beta}^*$, at which $\nabla^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}^*_{\mathbf{v}})$ is positive definite, the inverse function theorem implies that there exists a neighborhood $\{\mathbf{v} : \left\|\mathbf{v}\right\|_2 < \Delta'\}$ such that the map $\mathbf{v} \mapsto \boldsymbol{\beta}^*_{\mathbf{v}}$ is continuous (Krantz and Parks, 2002, Theorem 3.3.2). Taking

$$\delta_+ \leq \frac{\min\left(\Delta, \Delta'\right)}{\mathcal{R}_{\max}}$$

thus suffices to guarantee that

$$\inf_{\delta,r\in(0,\delta_+)\times\mathcal{R}} \lambda_{\delta r} \geq \frac{1}{2}\lambda_{\mathcal{I}} > 0.$$

We have one more condition of Assumption C.1 to satisfy, Item 6. A small modification of the proof of Lemma B.2 gives that, for sufficiently large N_S, there exists a $\gamma > 0$ such that

$$\sup_{\delta,r\in(0,\delta_+)\times\mathcal{R}} \left(\hat{\ell}_{N_S}(\boldsymbol{\beta}^*_{\delta r}; \delta r) - \sup_{\boldsymbol{\beta}\in\partial\mathcal{B}_\Delta(\delta r)} \hat{\ell}_{N_S}(\boldsymbol{\beta}; \delta, r) \right) > \gamma > 0,$$

where here $\partial\mathcal{B}_\Delta(\delta r) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*_{\delta r}\|_2 \leq \Delta\}$. This follows by a uniform law of large numbers applied to both $\boldsymbol{\beta}$ and to δ, r. The remainder of the proof of Lemma B.2 is unchanged, giving

$$\sup_{\delta,r\in(0,\delta_+)\times\mathcal{R}} \sup_{\boldsymbol{\beta}\in\mathbb{R}^{D_{\boldsymbol{\beta}}}\setminus\mathcal{B}_\Delta} \left(\frac{1}{N_S}\sum_{i\in[N_S]} \ell(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}^*_{\delta r}; \delta r) - \frac{1}{N_S}\sum_{i\in[N_S]} \ell(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}; \delta r) \right) \geq \gamma > 0.$$

It follows that Assumption C.1 is satisfied, and Lemma C.2 gives the desired result.

D Generating perturbed binary datasets

In this section, we describe one method for constructing randomized binary datasets that approximately represent a particular continuous perturbation. How to do so optimally seems to be an interesting topic for future work. We then rerun MCMC for a particular draw of the binary dataset to see whether the linear prediction given by MrPlew can extrapolate to analytically meaningful differences in the MrP estimates. Our results are mixed, and so we strongly recommend that balance and subgroup contribution plots be used as computationally efficient but approximate ways to explore the space of potentially problematic datasets, which are then verified by MCMC.

D.1 Constructing a binary response vector

Recall that our goal, by eq. (21), is to identify a binary random variable which we will call $\breve{y}$, such that

$$\mathbb{E}_{\mathcal{P}(\breve{y}|\boldsymbol{x})}[\breve{y}] \approx \mathbb{E}_{\mathcal{P}(\tilde{y}|\boldsymbol{x})}[\tilde{y}] = \mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[y] + \delta r. \tag{40}$$

Our motivation for doing so will be that, if $\hat{\mu}^{\mathrm{MrP}}(\tilde{\boldsymbol{Y}}) - \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y})$ is large, so that r appears to be "imbalanced" according to Theorem 4.2, then we hope that the realizable $\hat{\mu}^{\mathrm{MrP}}(\breve{\boldsymbol{Y}}) - \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y})$ may be "imbalanced" as well. We are further motivated to find a $\breve{\boldsymbol{Y}}$ that is "close" to $\boldsymbol{Y}$ so that the approximation eq. (19) remains approximately valid for $\hat{\mu}^{\mathrm{MrP}}(\breve{\boldsymbol{Y}}) - \hat{\mu}^{\mathrm{MrP}}(\boldsymbol{Y})$.

One way to proceed is via a perturbation to the parametric bootstrap (Efron and Tibshirani, 1994). We begin with a preliminary guess $\hat{m}_i := \hat{m}(\boldsymbol{x}_i) \approx \mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[y]$ for each i in the survey. If the guess $\hat{m}(\cdot)$ is poor, then the perturbed data will not bear the same relationship to the regressors as would a genuine alternative draw of the data. However, we emphasize that the role of $\hat{m}(\boldsymbol{x}_i)$ is to *define a data perturbation of interest*, rather than to perform formal inference. A reasonable choice might be to take $\hat{\boldsymbol{\beta}} := \mathbb{E}_{\mathcal{P}(\boldsymbol{\beta}|\boldsymbol{Y})}[\boldsymbol{\beta}]$

and $\hat{m}(\boldsymbol{x}_i) = m(\hat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_i)$. Two other extremes would be to take $m(\boldsymbol{x}_i) = \frac{1}{N_S}\sum_{i\in[N_S]} y_i$, or to take $m(\boldsymbol{x}_i) = y_i$. As we will see below, these extremes represent tradeoffs in their ability to plausibly reproduce the change eq. (21) on one hand, and to produce estimates that are close to $\boldsymbol{Y}$ on the other.

Supposing that $\mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[y] = \hat{m}(\boldsymbol{x})$, we wish to draw a binary $\breve{y}$ that is "not far" from y, but which has expectation $\hat{m}(\boldsymbol{x}) + \delta r$. One way to do that is to couple y and $\breve{y}$ with a common uniform random variable $u \sim \text{Uniform}(0,1)$: $y = \mathbb{I}(u < \hat{m}(\boldsymbol{x}))$ and $\breve{y} = \mathbb{I}(u < \hat{m}(\boldsymbol{x}) + \delta r)$. Under this construction, $\mathbb{E}[\breve{y}] = \hat{m}(\boldsymbol{x}) + \delta r$ and $\mathbb{E}[y] = \hat{m}(\boldsymbol{x})$ as desired, as long as $\hat{m}(\boldsymbol{x}) + \delta r \in [0,1]$ and $\hat{m}(\boldsymbol{x}) \in [0,1]$. Given y and $\boldsymbol{x}$, the conditional distribution of u_i is then

$$\mathcal{P}(u|y,\boldsymbol{x}) = \begin{cases} \text{Uniform}(0, \hat{m}(\boldsymbol{x})) & \text{if } y = 1 \\ \text{Uniform}(\hat{m}(\boldsymbol{x}), 1) & \text{if } y = 0. \end{cases}$$

Then if we draw $u \sim \mathcal{P}(u|y,\boldsymbol{x})$ and set $\breve{y} = \mathbb{I}(u \le \hat{m}(\boldsymbol{x}) + \delta r)$, then $\breve{y}$ is highly correlated with y marginally, and satisfies eq. (21) when $\hat{m}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[y]$. The procedure, which would be repeated for each i, is shown in Algorithm 1.

Algorithm 1: Draw perturbed binary data

Input: Estimate $\hat{m}_i \approx \mathbb{E}_{\mathcal{P}(y|\boldsymbol{x}_i)}[y]$, perturbation δr_i, original response y_i
Output: Vector of $\tilde{y}_i \in \{0,1\}$ with $\mathbb{E}_{\mathcal{P}(\tilde{y}|\boldsymbol{x}_i)}[\tilde{y}] \approx \mathbb{E}_{\mathcal{P}(y|\boldsymbol{x})}[y_i] + \delta r_i$
assert $\hat{m}_i + \delta r_i \in [0,1]$
`// Draw` $u_i \sim p(u \mid y_i)$
if $y_i = 1$ **then**
  $u_i \sim \text{Uniform}(\hat{m}_i, 1)$;
else
  $u_i \sim \text{Uniform}(0, \hat{m}_i)$;
`// Draw` $\breve{y}_i \sim p(\breve{y} \mid y_i)$
$\breve{y}_i \leftarrow \mathbb{I}(u_i \ge \hat{m}_i + \delta r_i)$

Note also that algorithm 1 is random. In general, for a given $\hat{m}(\cdot)$, there are many binary datasets that are consistent with a particular mean perturbation.

We note that choices of $\hat{m}(\boldsymbol{x}_i)$ like $\frac{1}{N_S}\sum_{i\in[N_S]} y_i$ that "underfit" the data permit larger perturbations, since δ may be larger before $\hat{m}(\boldsymbol{x}_i)$ leaves $[0,1]$, but at the risk of failing to reproduce the intended relationship between $\boldsymbol{x}_i$ and $\breve{y}_i$. On the other hand, "overfitting" the data, say by choosing $\hat{m}(\boldsymbol{x}_i) = y_i$ perfectly reproduces the relationship between $\boldsymbol{x}_i$ and y_i, but there may be no non-zero δ satisfying $\hat{m}(\boldsymbol{x}_i) + \delta r_i \in [0,1]$ — for example, if $r_i > 0$ and $y_i = 1$, then we must have $\delta = 0$. Ultimately, the tradeoff between fidelity to the data generating distribution and the ability to produce perturbed datasets is a judgement call that defines the robustness question that is being asked.

E Additional Experimental Results

E.1 Supplementary Graphs

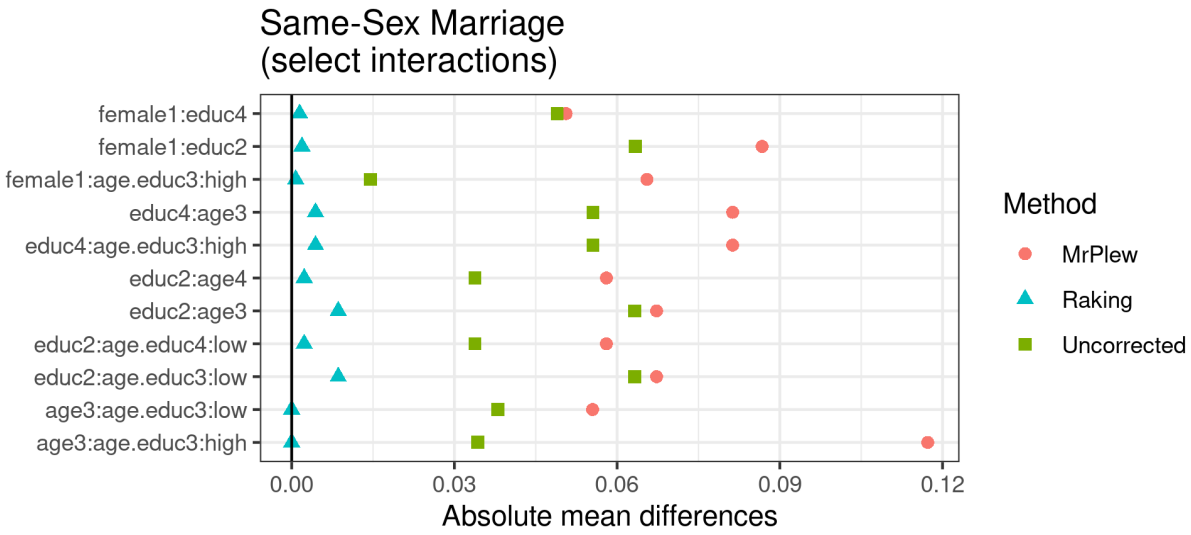


Figure 9: Balance

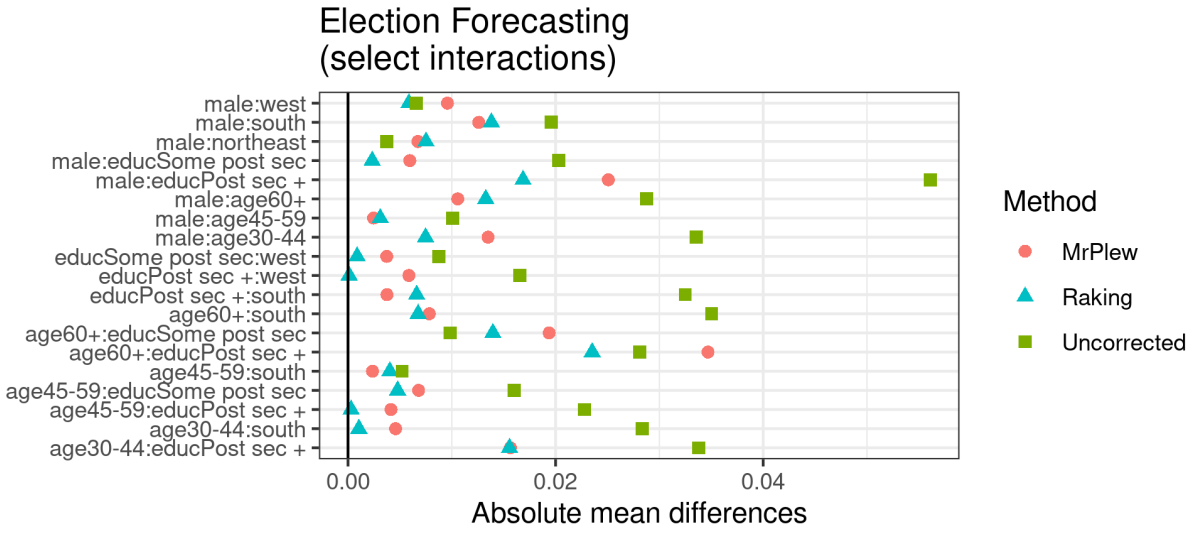


Figure 10: Balance

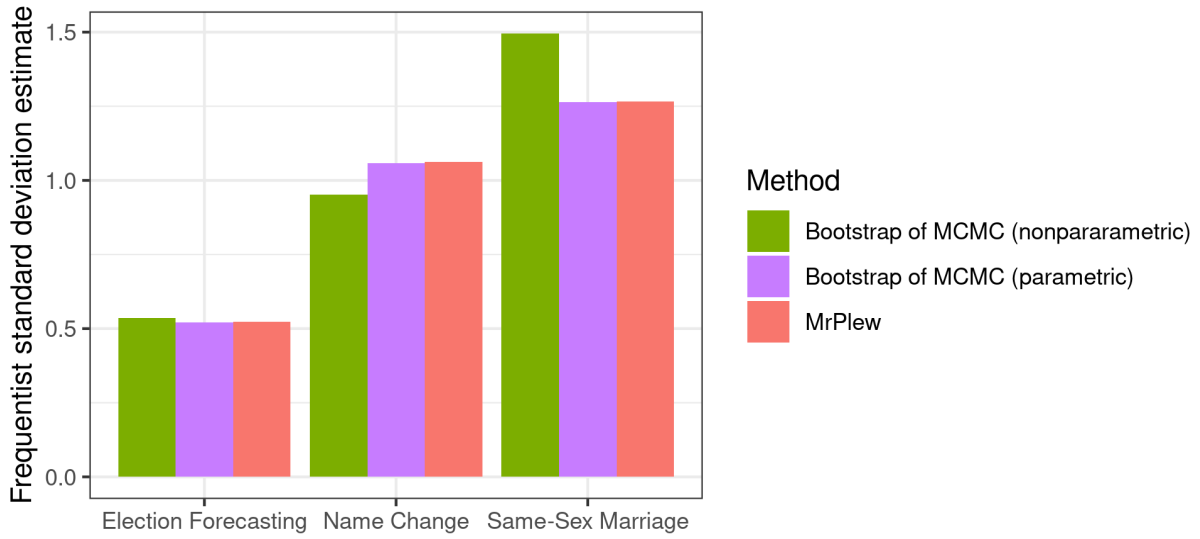


Figure 11: Estimates of the frequentist standard deviation of $\sqrt{N_S}\hat{\mu}^{\mathrm{MrP}}$

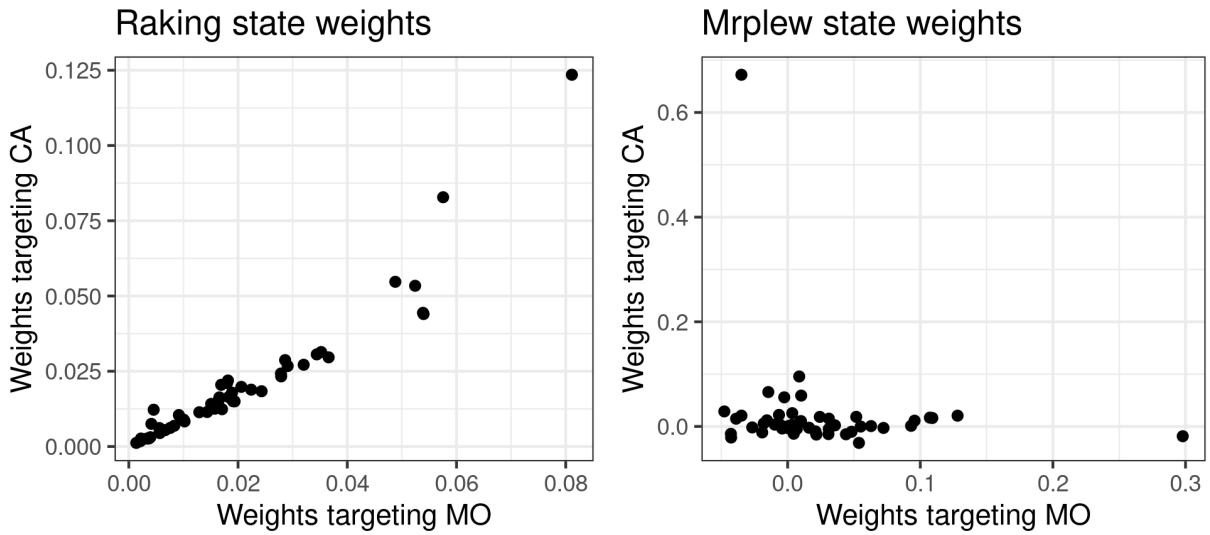


Figure 12: Different variability in per-state subgroup contribution weights for Same-Sex Marriage. Each point is a state.

E.2 Extrapolating to assess non-local robustness

Following the setup in Section 6.2, we investigate how well the MrPlew balance checks assess non-local robustness after extrapolating along a chosen dimension. For our given outcome prediction $\hat{m}(\cdot)$, we investigate Equation (24) by re-running MCMC. Specifically, we choose a range of δ, up to the largest possible step size $\delta_{\max}$ that result in $\hat{m}_i + \delta r_i \in [0, 1]$ for all i. At the most extreme points, which effectively set all the responses in the given category to 1, we have changed 10% of the responses in the Same-Sex Marriage and

2% of the responses in the Name Change dataset. For each δ in each analysis, we ran algorithm 1 from Appendix D to produce a draw $\breve{\boldsymbol{Y}}$, and then re-ran MCMC to compute $\hat{\mu}^{\text{MrP}}(\breve{\boldsymbol{Y}})$.

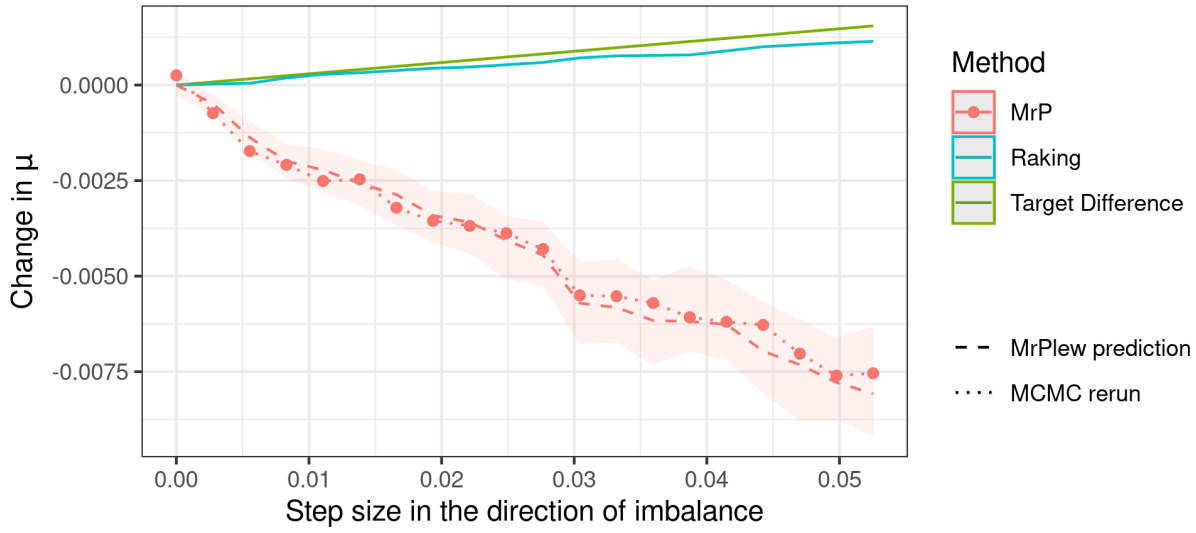


Figure 13: Refit

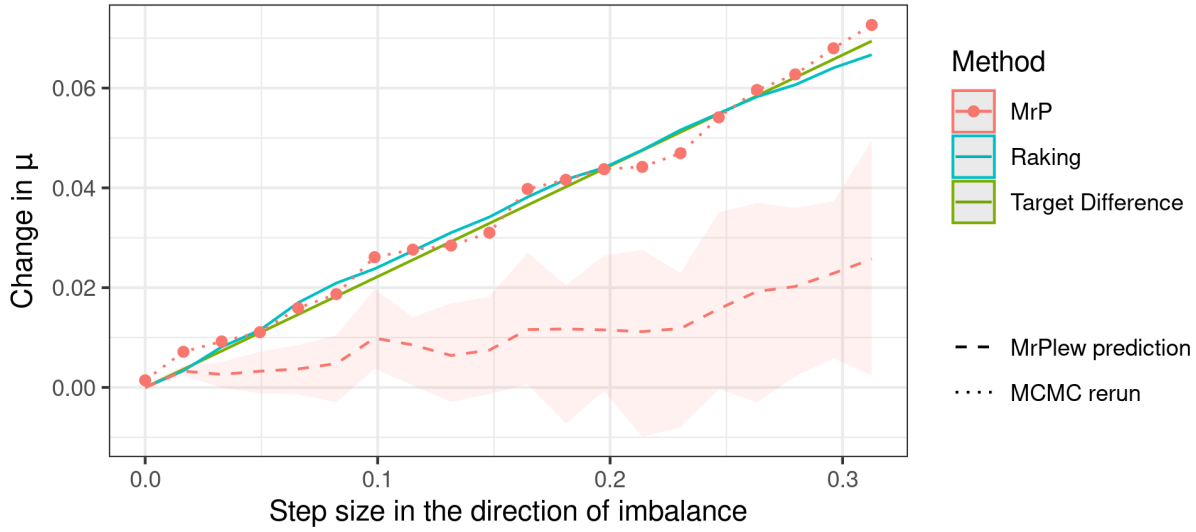


Figure 14: Refit

E.3 Nonlinearity in Same-Sex Marriage analysis

In this section we briefly demonstrate that posterior is in fact locally linear in fig. 14, but that even the least-perturbed binary vectors leave the domain of linearity quickly due to a large degree of posterior curvature.

We selected the smallest $\breve{\boldsymbol{Y}}$ not equal to $\boldsymbol{Y}$ from fig. 14. This $\breve{\boldsymbol{Y}}$ corresponds to the second datapoint from the left. We then define the MrP estimate

$$\hat{\mu}^{\mathrm{MrP}}(\epsilon) := \hat{\mu}^{\mathrm{MrP}}((1-\epsilon)\boldsymbol{Y} + \epsilon\breve{\boldsymbol{Y}}). \tag{41}$$

Both $\boldsymbol{Y}$ and $\breve{\boldsymbol{Y}}$ are binary vectors, and eq. (41) defines a smooth path between them as ϵ varies from 0 to 1. For sufficiently small ϵ we can estimate $\hat{\mu}^{\mathrm{MrP}}(\epsilon)$ with self-normalized importance sampling, and we consider ϵ for which there are at least 1000 effective samples in the importance-reweighted posterior.

Figure 15 shows the path of both the MrPlew linear approximation to $\hat{\mu}^{\mathrm{MrP}}(\epsilon) - \hat{\mu}^{\mathrm{MrP}}$ and the importance sampling posterior estimate. The local linearity is evident for very small ϵ, but the posterior rapidly deviates from the linear approximation long before $\epsilon = 1$.

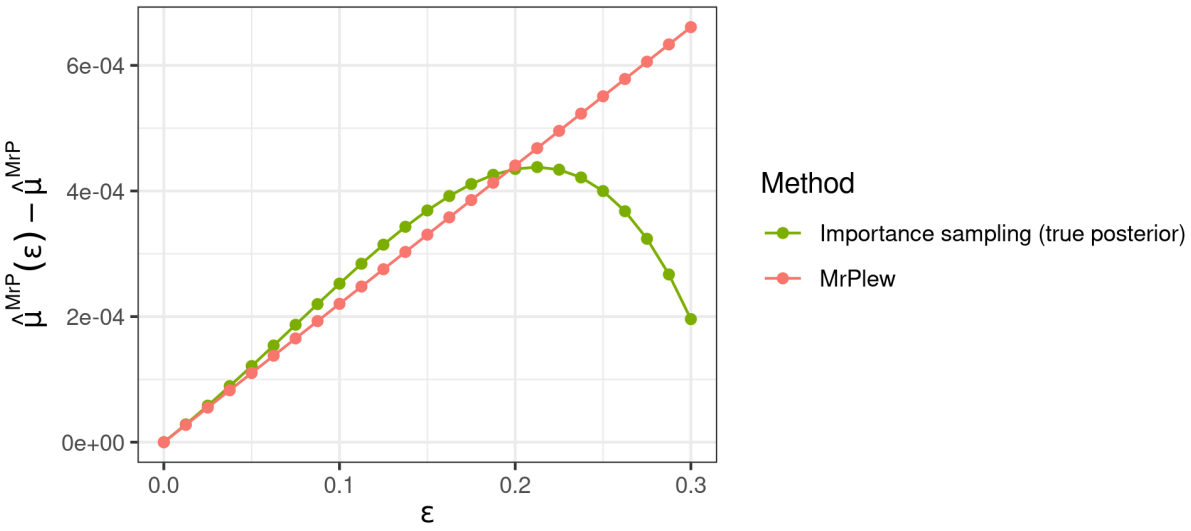


Figure 15: Local behavior of the Same-Sex Marriage perturbation. We only consider ϵ with at least 1000 effective self-normalized importance samples.

Though not shown, a randomly selected binary response vector that changes the same number of entries as $\breve{\boldsymbol{Y}}$ exhibits very little curvature. There may be something about the fact that we are perturbing the posterior in a direction of imbalance that is causing the severe non-linearity. Theoretically and practically understanding why the Same-Sex Marriage analysis exhibits such strong curvature but the Name Change analysis does not remains important future work.