

# Calibrated Inference: Statistical Inference that Accounts for Both Sampling Uncertainty and Distributional Uncertainty

**Yujin Jeong**

*Department of Statistics  
Stanford University  
Stanford, CA 94305, USA*

YUJINJ@STANFORD.EDU

**Dominik Rothenhäusler**

*Department of Statistics  
Stanford University  
Stanford, CA 94305, USA*

RDOMINIK@STANFORD.EDU

**Editor:** Genevera Allen

## Abstract

How can we draw trustworthy scientific conclusions? One criterion is that a study can be replicated by independent teams. While replication is critically important, it is arguably insufficient. If a study is biased for some reason and other studies recapitulate the approach then findings might be consistently incorrect. It has been argued that trustworthy scientific conclusions require disparate sources of evidence. However, different methods might have shared biases, making it difficult to judge the trustworthiness of a result. We formalize this issue by introducing a “distributional uncertainty model”, wherein dense distributional shifts emerge as the superposition of numerous small random changes. The distributional perturbation model arises under a symmetry assumption on distributional shifts and is strictly weaker than assuming that the data is i.i.d. from the target distribution. We show that a stability analysis on a single data set allows us to construct confidence intervals that account for both sampling uncertainty and distributional uncertainty.

**Keywords:** Distributional Stability, Distributional Robustness, Distributional Shifts, Generalizability, Uncertainty Quantification

## 1. Introduction

Statistical inferences can be fragile. If we compare two analyses conducted by different data scientists on different data sets, variation can be due to sampling, due to distribution shift, or due to a change in methodology. These issues raise a fundamental question: How can we draw trustworthy scientific conclusions? A common recommendation is to have independent teams attempt to replicate the findings of others. While replication is critically important, it is arguably insufficient. If a study suffers from biases for some reason and replication studies emulate the study, the findings will be consistently incorrect (Munafò and Smith, 2018). To solve this issue, researchers have advocated investigating independent lines of evidence (Denzin, 1970; Freedman, 1991; Rosenbaum, 2010; Munafò and Smith, 2018). Ideally, these different lines of evidence are susceptible to different biases. Intuitively, if results agree across different methodologies, then a statistical finding is less likely to be an

artifact. However, it might be expensive and impractical to ask several researchers to run studies independently.

Can we emulate this strategy on a single data set? In fact, stability analyses have been advocated by many researchers. To be more precise, it has been recommended to evaluate several reasonable modelling choices for one single data set (Leamer, 1983; Rosenbaum, 2010; Patel et al., 2015; Steegen et al., 2016; Yu and Kumbier, 2020). Practitioners often compute multiple estimators for a single target quantity by running differently specified regressions or considering the perturbations induced by various forms of data pre-processing. If the estimator-to-estimator variability is high, then the analyst has reason to distrust the estimates.

This practice — computing multiple estimators for a single target quantity and studying their estimator-to-estimator variability — warrants an investigation into its theoretical properties. If the estimator-to-estimator variability is high, it may raise concerns about the reliability of the estimates. However, what criterion tells us whether we should be concerned about such variability? Does this practice come with any guarantees, and if so, which ones? What mathematical problems do we address by examining the estimator-to-estimator variability? Often the decision on what qualifies as a “stable result” is left up to the individual judgments of the analyst. One could analyze the estimators with a random effects model, but since the estimators are computed on the same data set, they might share biases. More concretely speaking, since the structure of the biases is generally unknown, it is not clear how to define a random effects model that captures the correlation structure of the estimators’ biases.

In order to discuss these questions in a formal framework, we take a distributional perspective. We consider a setting where the data is drawn from a “perturbed” or “contaminated” distribution, while the goal is to infer some properties of the uncontaminated distribution. The classical robust statistics literature (Huber, 1981) addresses distributional perturbations by investigating the worst-case behavior of a statistical functional over a fixed neighborhood of the model. More recently, distributional uncertainty sets based on  $f$ -divergence have been linked to distributionally robust optimization (Ben-Tal et al., 2013; Duchi et al., 2021). In such models, it is challenging to choose the appropriate set of distributions, since the size of the perturbation is generally unknown.

Selection bias, confounding variables, or batch effects may be seen as sparse distribution shifts, where the shift affects only specific parts of the data generation process, while other parts stay invariant.

In contrast to sparse distribution shifts, we consider dense distribution shifts where the shift arises as the superposition of many small random changes that affect all parts of the data generating process. Here is an example that illustrates how dense distribution shifts may arise in practice: In clinical trials, distribution shifts can arise from complex, unobservable factors in the underlying population. Consider a scenario where a trial is conducted in 2022. During this time, various factors influence the population’s health, such as the severity of the flu that year, local demographic patterns, and other environmental or social variables. When attempting to replicate the trial in a later period, the distribution of these factors may have changed significantly. However, because there are so many contributing variables—many of which are difficult or impossible to measure—the cumulative effect of these changes cannot be precisely quantified. This scenario exemplifies a dense shift: the

overall distribution is shaped by distribution shifts in a multitude of nonrandom factors, whose combined effects are so intricate and unpredictable that they can be modeled as effectively random for practical purposes. We model dense distribution shifts as random and symmetric by randomly up-weighting and down-weighting different parts of the target distribution, capturing the inherent unpredictability and complexity of dense shifts. As another example, in economics, business cycles are often seen as driven by ‘random summation of random causes’ (Drautzburg, 2019). One potential interpretation of this is that in complex social and economic systems, there might be dense and unpredictable shifts driven by a confluence of broader social, economic, and natural forces.

Motivated by an empirical phenomenon observed in several real-world data sets, we define a family of random symmetric perturbations. While the symmetry assumption is strong, it is strictly weaker than assuming the data is i.i.d. from the target distribution. Distribution shift observed in the real world may involve a combination of sparse and dense shifts and we may need a hybrid approach in the future. As a first stepping stone, we aim to establish theoretical foundations addressing dense, symmetric distribution shifts.

Finally, we show that modeling distributional perturbations as random and symmetric has an intriguing consequence: using a stability analysis, it is possible to estimate the strength of the distributional perturbation. Based on an estimate of the distributional perturbation strength, we propose confidence intervals that capture both sampling uncertainty and distributional uncertainty.

## 1.1 Related Work

Considerations of model stability have emerged in Bayesian statistics (Box, 1980; Skene et al., 1986), causal inference (Leamer, 1983; LaLonde, 1986; Rosenbaum, 1987; Imbens and Rubin, 2015) and in discussions about the data science life-cycle (Yu, 2013; Steegen et al., 2016; Yu and Kumbier, 2020). Using different estimation strategies is commonly recommended to corroborate a causal hypothesis (Freedman, 1991; Rosenbaum, 2010; Karmakar et al., 2019). In particular, to evaluate omitted variable bias, it is a common recommendation to consider the between-estimator variation of several adjusted regressions (Oster, 2019). Sensitivity analysis bounds the influence of confounders that have been omitted in a regression or matching procedure and has played an influential role in increasing trustworthiness of causal inference from observational data (Cornfield et al., 1959; Rosenbaum and Rubin, 1983; VanderWeele and Ding, 2017). It has been argued that causal mechanisms are expected to lead to stable associations across settings, if the same mechanism is shared across settings. Based on this observation, stability principles have been employed to discover causal relationships based on heterogeneous data sets (Peters et al., 2016; Rothenhäusler et al., 2015; Bühlmann, 2020; Pfister et al., 2021). Stability principles are heavily used in machine learning, often with the goal of variance reduction. For example, some tree-based methods employ feature bagging, which can be seen as averaging over differently specified prediction models (Breiman, 1996, 2001). Dropout in neural networks is another form of algorithm perturbation (Srivastava et al., 2014). Distributional uncertainty sets based on  $f$ -divergences have been linked to distributionally robust optimization (Ben-Tal et al., 2013; Duchi et al., 2021). In the context of prediction under distribution shift, stability or invariance principles have been employed to learn prediction mechanisms that

generalize to new settings (Schölkopf et al., 2012; Zhang et al., 2013; Rojas-Carulla et al., 2018; Heinze-Deml and Meinshausen, 2021; Rothenhäusler et al., 2021). Quasi-likelihoods (Wedderburn, 1974) are a way to allow greater variability in the data than what is expected from the model. However, uncertainty quantification in quasi-likelihoods still only deals with sampling uncertainty, while we aim to quantify uncertainty due to both sampling and distributional uncertainty.

## 1.2 Outline of The Paper

In Section 1.3, we will quickly review standard practice for forming confidence intervals. In Section 2, we introduce the setting of the paper and discuss why standard statistical practice does not account for all types of uncertainty in this setting. The setting of our paper arises under a distributional perturbation model described in Section 2 and sampling procedures described in the Appendix. We then turn to statistical inference. In Section 3, we discuss how to form confidence intervals in our setting. This completes the picture from an inferential viewpoint. In Section 4, we evaluate the performance of the proposed procedure on a simulated example from causal inference. In Section 5 we demonstrate that the proposed procedure can increase the stability of decision-making based on real-world data. We conclude in Section 6.

## 1.3 Standard Approach

Let us consider estimation of the mean  $\theta^0 = \mathbb{E}[D]$  of a square-integrable real-valued random variable  $D \in \mathcal{D}$ ,  $D \sim \mathbb{P}$ . Assume that we are given data  $(D_i)_{i=1,\dots,n} \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$  with  $\text{Var}(D_i) = \sigma^2 \in (0, \infty)$ . We can estimate  $\sigma^2$  via  $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2$  to form asymptotically valid confidence intervals that means

$$\mathbb{P}(\bar{D} - z_{1-\alpha/2} \hat{\sigma} / \sqrt{n} \leq \theta^0 \leq \bar{D} + z_{1-\alpha/2} \hat{\sigma} / \sqrt{n}) \rightarrow 1 - \alpha, \quad (1)$$

where  $z_{1-\alpha/2}$  is the  $1 - \alpha/2$  quantile of a standard Gaussian random variable. This practice is justified by the central limit theorem which implies

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (D_i - \mathbb{E}[D]) \xrightarrow{d} \mathcal{N}(0, \text{Var}(D)). \quad (2)$$

More generally, for some vector-valued data  $D_i \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$  consider a parametrized model  $\{p_\theta, \theta \in \Omega\}$  of positive probability densities  $p_\theta$  with respect to some  $\sigma$ -finite measure  $\mu$ . Assume that the parameter space  $\Omega$  is an open subset of  $\mathbb{R}^d$ . We consider the maximum-likelihood estimator

$$\hat{\theta} = \arg \max \sum_{i=1}^n \log p_\theta(D_i),$$

for some unknown target parameter  $\theta^0(\mathbb{P}) = \arg \max \mathbb{E}[\log p_\theta(D)]$ , where  $D \sim \mathbb{P}$ . Under regularity assumptions (Van der Vaart, 2000; Tsiatis, 2006), for  $n \rightarrow \infty$ ,

$$\sqrt{n}(\hat{\theta} - \theta^0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n -\mathbb{E}[\partial_\theta^2 \log p_{\theta^0}(D)]^{-1} \partial_\theta \log p_{\theta^0}(D_i) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

where  $\Sigma = \mathbb{E}[\partial_\theta^2 \log p_{\theta^0}(D)]^{-1} \text{Var}(\partial_\theta \log p_{\theta^0}(D)) \mathbb{E}[\partial_\theta^2 \log p_{\theta^0}(D)]^{-1}$ . Thus, based on a consistent estimator  $\hat{\Sigma} \rightarrow \Sigma$ , one can form asymptotically valid confidence intervals via

$$\hat{\theta}_k \pm z_{1-\alpha/2} \frac{\sqrt{\hat{\Sigma}_{kk}}}{\sqrt{n}}. \quad (3)$$

A similar approach can be used to construct asymptotically valid confidence intervals for  $M$ -estimators. In the following, we discuss situations in which this approach does not have the desired coverage.

## 2. Distributional Uncertainty

There are several reasons why the coverage in equation (1) might not hold. The main focus of this paper will be violations of (2) due to what we call distributional uncertainty. Due to distribution shifts, the data scientist might not draw a sample from the target distribution  $\mathbb{P}$  but from some  $\mathbb{P}^\xi \neq \mathbb{P}$ . The data analyst may try to address the source of bias by re-weighting, regression adjustment, random effect modeling, a bias correction, or other statistical techniques. Our viewpoint is that when using such techniques, it is likely that some residual error remains which we might want to address by scaling the confidence intervals. Ideally, we would like to construct confidence intervals that detect residual errors due to distributional perturbations, and account for them, if necessary. We model the variation due to distributional changes as random. This will allow us to integrate both distributional uncertainty and sampling uncertainty in a natural fashion.

Let us make this more concrete by returning to the example of estimating the mean. Due to a superposition of small random errors, the data scientist might not draw a sample from the target distribution  $\mathbb{P}$ . Instead, the data  $(D_i)_{i=1,\dots,n}$  might be drawn i.i.d. from some perturbed distribution  $\mathbb{P}^\xi \neq \mathbb{P}$ , where  $\xi$  is a random variable and  $\mathbb{P}^\bullet$  is a probability distribution for each fixed  $\bullet \in \text{range}(\xi)$ . Then the error of the empirical mean can be decomposed:

$$\frac{1}{n} \sum_{i=1}^n D_i - \mathbb{E}[D] = \underbrace{\frac{1}{n} \sum_{i=1}^n D_i - \mathbb{E}^\xi[D]}_{\text{variation due to sampling}} + \underbrace{\mathbb{E}^\xi[D] - \mathbb{E}[D]}_{\text{variation due to distributional perturbation}}. \quad (4)$$

Here,  $\mathbb{E}$  denotes the expectation under the target distribution  $\mathbb{P}$  and  $\mathbb{E}^\xi$  denotes the expectation under the perturbed distribution  $\mathbb{P}^\xi$ . Equation (2) usually does not hold in this setting as distributional perturbations induce additional variation.

In the following, we focus on the regime where the variation due to sampling and the variation due to distributional perturbations are both of the order  $1/\sqrt{n}$ . This choice is motivated by observations from real-world data sets, where these variations often appear to be of similar order (see Figure 1). There may be situations where higher-order bias exists and should be removed when possible. However, even after accounting for all estimable bias components, residual data quality issues can remain. We would like to address these residual errors by scaling confidence intervals.

**Notation.** Let  $\mathbb{P}$  denote an unknown fixed target probability measure on  $\mathcal{D}$ . For each fixed  $n$  the random variable  $\xi(n) \in \Xi$  encodes the distributional perturbation. Formally,  $\mathbb{P}^\bullet$ ,  $\bullet \in \Xi$ , is a stochastic kernel with the target space  $\mathcal{D}$ . Conditionally on  $\xi(n)$ , we draw an i.i.d. sample  $(D_1^n, \dots, D_n^n)$  from  $\mathbb{P}^{\xi(n)}$ .  $\xi(n)$  might depend on  $n$  but we suppress this in the notation and simply write  $\xi$ . Similarly, we sometimes suppress the dependence of  $(D_1^n, \dots, D_n^n)$  on  $n$  and simply write  $(D_1, \dots, D_n)$ . We write  $P$  for the marginal distribution of  $(D_1, \dots, D_n, \xi)$ . We denote  $E$  as the marginal expectation under  $P$ ,  $\mathbb{E}$  as the expectation under the target distribution  $\mathbb{P}$ , and  $\mathbb{E}^\xi$  as the expectation under  $\mathbb{P}^\xi$  (conditioned on  $\xi$ ). We write  $\text{Var}_P$  for the variance under  $P$  and  $\text{Var}_{\mathbb{P}}$  for the variance under  $\mathbb{P}$ .

## 2.1 Empirical examples

What is a reasonable model for the distributional perturbation? We draw inspiration from direct replication studies and the GTEx<sup>1</sup> gene expression data.

For the two direct replication studies in Prochazka et al. (2022), we assume that the first replication study  $D_1, \dots, D_{n_1}$  is drawn i.i.d. from the target distribution  $\mathbb{P}$ , and the data set of the second replication study consists of  $n_2$  samples  $(D'_1, \dots, D'_{n_2})$  drawn from the perturbed distribution  $\mathbb{P}^\xi$ . Analogously, for GTEx gene expression data (V6), we randomly selected the tissue “Liver” and consider the data set from the tissue as  $n_2$  samples from the perturbed distribution. Then we define the target distribution as the distribution of gene expression across the remaining tissues.

From a statistical perspective, it is natural to study the distribution of the standardized mean difference:

$$\left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{-1/2} \left( \frac{\frac{1}{n_2} \sum_{i=1}^{n_2} \psi(D'_i) - \frac{1}{n_1} \sum_{i=1}^{n_1} \psi(D_i)}{\hat{\text{sd}}_{\mathbb{P}}(\psi(D))} \right) \text{ for some test function } \psi. \quad (5)$$

If there were no distributional shifts across different replication studies or different tissues and all data were sampled i.i.d. from  $\mathbb{P}^\xi = \mathbb{P}$ , for fixed  $\psi$  one would expect the ratio to follow roughly a standard Gaussian distribution.

To investigate the distribution of equation (5) in practice, we define  $\psi_\ell$  as follows. For the replication studies in Prochazka et al. (2022), we define  $\psi_\ell$  as either the covariate mean or the sign-flipped covariate mean for either the treatment or control group, resulting in 180 test functions. For GTEx gene expression data, the covariate means are standardized in a pre-processing step. Thus, we study cross-products. We randomly select 1000 gene pairs and define  $\psi_\ell$  as the product of gene-expressions for the  $\ell$ -th pair. We illustrate the behavior of (5) from replication studies and GTEx data on the left side of Figure 1 using QQ plots.

**Unexpected: distribution shift on the line.** Surprisingly, the QQ plots in Figure 1 indicate that the statistics in (5) follow a Gaussian distribution. The slopes of the black QQ lines are larger than 1, which indicate excess variation compared to i.i.d. sampling from  $\mathbb{P}$ .

1. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data set used for the analyses described in this manuscript is version 6 and can be downloaded in the GTEx Portal: [www.gtexportal.org](http://www.gtexportal.org).

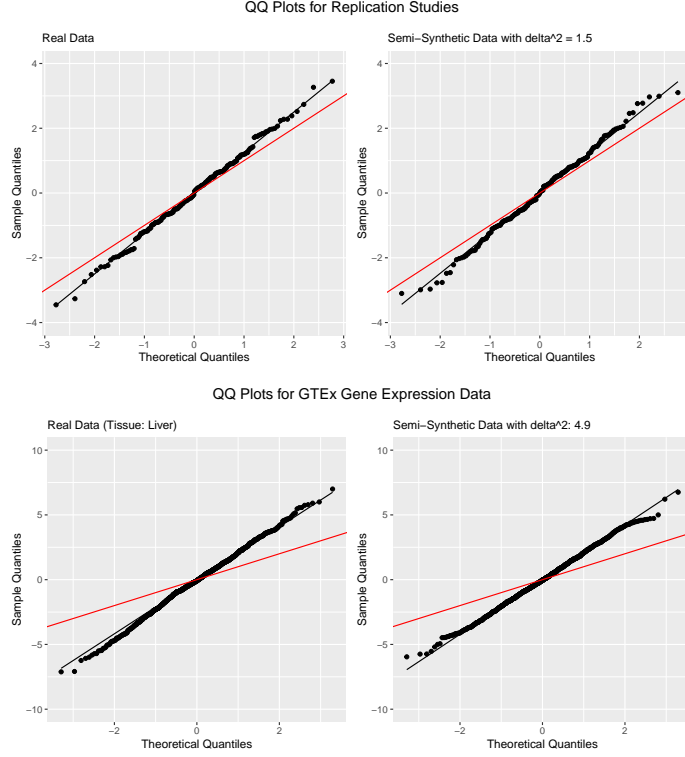


Figure 1: QQ plots of (5) for various test functions  $\psi_l$  on replication studies in Prochazka et al. (2022) (above) and GTEx gene expression data set (below). The QQ plots on the left side are from real-world data sets. The red line represents the expected QQ line if the data were all drawn i.i.d. from some (unperturbed) distribution  $\mathbb{P}$ . Perhaps surprisingly, the standardized means are on a line, indicating that the distribution shift has some structure that we can exploit for estimation and inference. The QQ plots on the right side are computed on the data drawn from our model (6) with estimated variance inflation factor  $\delta$  from the left side. The simulated shifts on the right closely match the pattern observed on the left side.

This raises the question of whether we can model such (moderate) variance inflation with a statistical model for distribution shift.

**‘Distribution shift on the line’ implies isotropic perturbations.** One can show that constant variance inflation as in Figure 1 implies that the shifted distribution arises from randomly re-weighting the original distribution with uncorrelated weights with equal variance. To not interrupt the flow of the discussion, we provide the justification for this claim in Appendix A.1.

**Isotropic perturbations imply ‘distribution shift on the line’.** In Section 2.2, we will introduce a random perturbation model (equation (6)) that is based on randomly re-weighting the target distribution with independent weights with equal variance.

As a sanity check, we semi-synthetically sample from this model. If the random perturbation model is reasonable, the semi-synthetic data should exhibit similar patterns as the left-hand side of Figure 1. From the target data set, we sample a perturbed data set using the estimated variance inflation factor  $\delta$  obtained from the left side of Figure 1. The QQ plots generated from the semi-synthetic data are presented on the right side of Figure 1. We see that these plots closely resemble the patterns observed in the QQ plots obtained from real-world data sets.

## 2.2 The Isotropic Perturbation Model

In this section, we construct a general isotropic perturbation model for multivariate continuous or discrete random variables and consider the case where the change in measure is a superposition of small incremental changes. We will see that under such a model, we will get a non-standard CLT in the sense that sample means are asymptotically normal, but with a different variance formula compared to the i.i.d. case.

To recap, in a random perturbation model, the data is not directly drawn from the target distribution  $\mathbb{P}$ , but from some random probability measure  $\mathbb{P}^\xi$ , where  $\mathbb{P}^\xi$  is close to  $\mathbb{P}$ . The idea is that due to numerous random distributional changes, the actual sampling distribution  $\mathbb{P}^\xi$  randomly differs from the target distribution  $\mathbb{P}$ . Under  $\mathbb{P}^\xi$ , probabilities of events are slightly up-weighted or down-weighted compared to  $\mathbb{P}$ .

We want to construct a random perturbation model that includes many commonly encountered situations such as distributions on  $\mathbb{R}^d$  or the (infinite-dimensional) space of continuous functions on  $\mathbb{R}$ . A result from probability theory shows that any random variable  $D$  on a finite or countably infinite dimensional probability space can be written as a measurable function  $D \stackrel{d}{=} h(U)$ , where  $U$  is a uniform random variable on  $[0, 1]$ .<sup>2</sup> Thus, without loss of generality we will construct distributional perturbations for a uniform distribution on  $[0, 1]$ . With the transformation  $h(\cdot)$  defined above, this construction generalizes to the general cases by setting

$$\mathbb{P}^\xi(D \in \bullet) = \mathbb{P}^\xi(h(U) \in \bullet).$$

The role of  $h(\cdot)$  is mainly to ensure that the probability space is rich enough to be transformed into a uniform random variable. In principle, this construction is not unique. There may be many possible choices of  $h(\cdot)$  that result in  $D \stackrel{d}{=} h(U)$ . However, as we will see below, the asymptotic behaviour of the perturbation model does not depend on the choice of  $h$ .

Let us now construct the distributional perturbation for a uniform random variable. As discussed in Appendix A.1, the distribution-shift-on-the-line phenomenon observed in Figure 1 suggests that we can think about the shifted distribution as arising from randomly re-weighting the original distribution with (almost) uncorrelated weights with equal variance. Thus, we take  $m$  bins  $I_k = [(k-1)/m, k/m]$  for  $k = 1, \dots, m$ . Let  $W_1, \dots, W_m$  be i.i.d. positive random variables with finite variance. Set  $\xi = (W_1, \dots, W_m)$ . We define the

---

2. For any Borel-measurable random variable  $D$  on a Polish (separable and completely metrizable) space  $\mathcal{D}$ , there exists a Borel-measurable function  $h$  such that  $D \stackrel{d}{=} h(U)$  where  $U$  follows the uniform distribution on  $[0, 1]$  (Dudley, 2018).



randomly perturbed distribution  $\mathbb{P}^\xi$  by setting

$$\mathbb{P}^\xi(U \in \bullet) = \sum_k \mathbb{P}(U \in I_k \cap \bullet) \cdot \frac{W_k}{\sum_{k=1}^m W_k/m}. \quad (6)$$

Let  $m = m(n)$  such that  $\frac{n}{m(n)}$  converges to some limit  $r \in (0, \infty)$ . Note that  $\xi$  depends on  $m$  and thus also on  $n$ . Conditionally on  $\xi$ , let  $(D_1^n, \dots, D_n^n)$  be i.i.d. draws from  $\mathbb{P}^\xi$ .

**Lemma 1 (CLT under distributional uncertainty)** *Under the assumptions mentioned above (Section 2.2), for any Borel-measurable square-integrable function  $\psi : \mathcal{D} \mapsto \mathbb{R}^l$ , we have*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\psi(D_i^n) - \mathbb{E}[\psi(D)]) \xrightarrow{d} \mathcal{N}(0, \delta^2 \text{Var}_{\mathbb{P}}(\psi(D))), \quad (7)$$

with

$$\delta^2 = 1 + \frac{r \text{Var}(W_1)}{E[W_1]^2}.$$

In other words, the marginal distribution of  $\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(D_i^n)$  is asymptotically Gaussian with asymptotic variance containing a scaling factor  $\delta^2$ .

The proof can be found in Appendix B.1. The reason we consider a triangular array of data sets is that, motivated by the empirical example in Figure 1, we consider a setting where the sampling uncertainty and distributional uncertainty are of the same order.

Unless explicitly mentioned otherwise, in the following we assume that the data scientist has access to one such data set  $(D_1^n, \dots, D_n^n)$  for some large  $n$ . Note that if the data  $(D_1, \dots, D_n)$  is  $\overset{\text{i.i.d.}}{\sim} \mathbb{P}$ , then equation (7) holds for  $\delta = 1$  and  $D_i^n = D_i$ . Thus, equation (7) is weaker than assuming that the data is drawn i.i.d. from  $\mathbb{P}$ .

The asymptotic behaviour shown in equation (7) arises not only under the distributional perturbation model but other types of sampling procedures that induce dependence between observations. In Appendix B.2, we discuss other sampling models that give rise to (7).

In the following, we discuss how estimators behave asymptotically under equation (7). It turns out that under some regularity assumptions, maximum likelihood estimators are still consistent and asymptotically normal, but with the scaling factor  $\delta^2$  in the variance formula.

### 2.3 Asymptotic Behaviour of M-estimators

Here we will consider the asymptotic behaviour of estimators  $\hat{\theta} = \arg \min_{\theta \in \Omega} \frac{1}{n} \sum_{i=1}^n L(\theta, D_i^n)$  for a target defined via  $\theta^0 = \arg \min_{\theta \in \Omega} \mathbb{E}[L(\theta, D)]$ , where  $L(\theta, \bullet)$  is a Borel-measurable loss function and  $\Omega$  is an open subset of  $\mathbb{R}^d$ . These estimators include maximum likelihood estimators with  $L(\theta, D) = -\log p_\theta(D)$ .

In classical statistical theory, uncertainty quantification is usually based on showing that the estimator is asymptotically Gaussian. Since we have a different two-stage sampling model, one has to verify that a similar approximation – with a different variance formula – still holds in our setting.

First, we will discuss consistency. Instead of aiming for maximal generality, we will adapt a simple consistency proof from the literature. In particular, we will adapt the

classical consistency result in Van der Vaart (2000), Section 5.2.1. We expect that other consistency proofs can be adapted similarly. The main difference in the proof is that since the data is not i.i.d. from the target distribution we cannot directly rely on the law of large numbers. The proof can be found in Appendix B.6.

**Lemma 2 (Consistency of M-estimators)** *Consider the M-estimator*

$$\hat{\theta} = \arg \min_{\theta \in \Omega} \frac{1}{n} \sum_{i=1}^n L(\theta, D_i^n),$$

and the target  $\theta^0 = \arg \min_{\theta \in \Omega} \mathbb{E}[L(\theta, D)]$ , where  $\Omega$  is a compact subset of  $\mathbb{R}^d$ . Furthermore assume that  $\theta \mapsto L(\theta, D)$  is continuous and that  $\inf_{\|\theta - \theta'\|_2 \leq \delta} L(\theta, D)$  is square-integrable under  $\mathbb{P}$  for every  $\delta$  and  $\theta'$  and that  $\inf_{\theta \in \Omega} L(\theta, D)$  is square integrable. We assume that  $\mathbb{E}[L(\theta, D)]$  has a unique minimum. Then,

$$\hat{\theta} - \theta^0 = o_p(1).$$

Now let us turn to asymptotic normality. We will modify the proof in Van der Vaart (2000), Section 5.6. Similarly as above, the main difference in the proof is since the data is not i.i.d. from the target distribution we cannot directly rely on the law of large numbers or a standard CLT. The proof can be found in Appendix B.6.

**Lemma 3 (Asymptotic normality of M-estimators)** *For each  $\theta$  in an open subset of  $\Omega$ , let  $\theta \mapsto \partial_{\theta} L(\theta, D)$  be twice continuously differentiable in  $\theta$  for every  $D$ . Assume that the matrix  $\mathbb{E}[\partial_{\theta}^2 L(\theta^0, D)]$  exists and is nonsingular. Assume that third order partial derivatives of  $\theta \mapsto L(\theta, D)$  are dominated by a fixed function  $h(\cdot)$  for every  $\theta$  in a neighborhood of  $\theta^0$ . We assume that  $\partial_{\theta} L(\theta^0, D)$ ,  $\partial_{\theta}^2 L(\theta^0, D)$  and  $h(D)$  are square-integrable under  $\mathbb{P}$ . Let  $\hat{\theta} = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n L(\theta, D_i^n)$ . Assume that  $\hat{\theta} - \theta^0 = o_p(1)$ , where  $\theta^0$  satisfies the estimating equation  $\mathbb{E}[\partial_{\theta} L(\theta^0, D)] = 0$ . Then,*

$$\sqrt{n}(\hat{\theta} - \theta^0) = -\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{E}[\partial_{\theta}^2 L(\theta^0, D)]^{-1} \partial_{\theta} L(\theta^0, D_i^n) + o_p(1).$$

In particular, by Lemma 1 we have that  $\sqrt{n}(\hat{\theta} - \theta^0)$  converges in distribution to a normal distribution with mean zero and covariance matrix  $\delta^2 \Sigma$ , where

$$\Sigma = \mathbb{E}[\partial_{\theta}^2 L(\theta^0, D)]^{-1} \mathbb{E}[\partial_{\theta} L(\theta^0, D) \partial_{\theta} L(\theta^0, D)^{\top}] \mathbb{E}[\partial_{\theta}^2 L(\theta^0, D)]^{-1}.$$

The upshot is that  $M$ -estimators are asymptotically unbiased, marginally across both sampling uncertainty and distributional uncertainty. However, the variance formula changes in the sense that there is an (unknown) scaling factor  $\delta^2$ .

## 2.4 The Standard Mode of Inference Fails

Let us quickly sketch why the standard mode of inference fails. Let's consider the case of estimating the mean  $\theta^0 = \mathbb{E}[D]$  via  $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n D_i^n$ . One may be tempted to use the standard variance estimate  $\hat{\sigma}_{\text{naive}}^2/n$ , where

$$\hat{\sigma}_{\text{naive}}^2 = \frac{1}{n} \sum_{i=1}^n \left( D_i^n - \frac{1}{n} \sum_{j=1}^n D_j^n \right)^2.$$

However, a short calculation shows that

$$\hat{\sigma}_{\text{naive}}^2 = \frac{1}{n} \sum_{i=1}^n \left( D_i^n - \frac{1}{n} \sum_{j=1}^n D_j^n \right)^2 = \frac{1}{n} \sum_{i=1}^n (D_i^n)^2 - \left( \frac{1}{n} \sum_{j=1}^n D_j^n \right)^2 = \text{Var}_{\mathbb{P}}(D) + o_P(1).$$

Here, we used equation (7) for  $\psi(D) = D$  and  $\psi(D) = D^2$ . However, as shown in Lemma 1, the asymptotic variance of  $\hat{\theta}$  is  $\frac{\delta^2}{n} \text{Var}_{\mathbb{P}}(D)$ . Thus, the standard approach drastically underestimates variance in our setting. If  $\delta$  is known, one can simply stretch the confidence intervals discussed in Section 1.3 accordingly. However, in general  $\delta$  will be unknown and has to be estimated from data. We will discuss the estimation of  $\delta$  in Section 3.

### 3. Calibrated Inference

We will now discuss how to estimate  $\delta$  and form asymptotically valid confidence intervals for  $\theta^0$ . As discussed earlier, data analysts often have not just one reasonable estimator for a given parameter  $\theta^0$ , but potentially several reasonable estimators  $\hat{\theta}^1, \dots, \hat{\theta}^K$ . For example, these estimators can arise from using different specifications in generalized linear models or by running the analysis for subgroups of the observations.

**Example 1 (OLS with several specifications)** *Let us consider a setting in which the data analyst wants to estimate the causal effect of some variable  $X_1$  on a target variable  $Y$ . On observational data, this is often done by invoking suitable assumptions and regressing  $Y$  on  $X_1$  and a suitable set of covariates. Often, the analyst has several reasonable choices for the set of covariates. Suppose that the data analyst performs ordinary least-squares on  $K$  different subsets of  $X$  that include  $X_1$ , denoted by  $X^{S_1}, X^{S_2}, \dots, X^{S_K}$ . For example,  $X^{S_1}$  can be  $(X_1, X_2, X_3)$ . Now the data analyst has  $K$  different regression coefficients of  $X_1$ ,  $\hat{\theta}^1, \dots, \hat{\theta}^K$  where*

$$\hat{\theta}^k = \left( \sum_{i=1}^n X_i^{S_k} (X_i^{S_k})^\top \right)^{-1}_{1,\bullet} \sum_{i=1}^n X_i^{S_k} Y_i.$$

If the empirical variation between the estimators  $\hat{\theta}^1, \dots, \hat{\theta}^K$  is low, then the analyst may feel more confident about conclusions drawn from these estimates than if the variation between these estimators is very large. As an example, in Chiappori et al. (2012) the authors write “It is reassuring that the estimates are very similar in the standard and the augmented specifications”. We will now look at this practice under the isotropic perturbation model. We will see that in this setting it is possible to construct a consistent estimator of  $\delta$  and form asymptotically valid confidence intervals that account for both sampling uncertainty and distributional perturbations.

If the estimators  $\hat{\theta}^k = \hat{\theta}^k(D_1, \dots, D_n)$  are M-estimators, by Lemma 2 and Lemma 3 the estimators are asymptotically linear in the sense that

$$\hat{\theta}^k - \theta^k = \frac{1}{n} \sum_{i=1}^n \phi^k(D_i) + o_P\left(\frac{1}{\sqrt{n}}\right), \quad (8)$$

for some deterministic  $\theta^k = \arg \min \mathbb{E}[L^k(\theta, D)]$ , where  $L^k$  is the loss function of the estimator  $\theta^k$ .  $\phi^k$  is referred to as the influence function of  $\hat{\theta}^k$  that is assumed to satisfy  $\mathbb{E}[\phi^k(D)] = 0$  and  $\text{Var}_{\mathbb{P}}(\phi^k(D)) \in (0, \infty)$ . Since  $\phi^k(D)$  is square integrable, by Lemma 1 the sequence  $\sqrt{n}(\hat{\theta}^k - \theta^k)$  converges in distribution to a normal distribution with mean zero and covariance  $\delta^2 \text{Var}_{\mathbb{P}}(\phi^k(D))$ . We summarize this behavior of the estimators as the following assumption for the convenience of reference later.

**Assumption 1 (Asymptotic linearity)** *The estimators  $\hat{\theta}^k$ ,  $k = 1, \dots, K$  are asymptotically linear, that is they satisfy equation (8) for influence functions  $\phi^k$  with  $\mathbb{E}[\phi^k(D)] = 0$  and  $0 < \text{Var}_{\mathbb{P}}(\phi^k(D)) < \infty$ .*

As discussed above, for the case of  $M$ -estimators, this assumption can be justified via Lemma 2 and Lemma 3. We will now formalize the premise that the data analyst considers each of the  $\hat{\theta}^k$  a reasonable estimator for the parameter of interest,  $\theta^0$ .

**Assumption 2 (Agreement)** *We have  $\theta^k = \theta^0$  for  $k = 1, \dots, K$ .*

This assumption must be justified with scientific background knowledge. Intuitively, the assumption states that if both sampling uncertainty and distributional uncertainty were negligible, the estimators would agree. In Section 4, we discuss in a numerical example how the choice of such estimators can be justified. If the data scientist does not believe in asymptotic agreement of the estimators, we present conservative confidence intervals in Appendix C.

### 3.1 Confidence Intervals

Now let us turn to constructing confidence intervals for  $\theta^0$ . Assume that the data analyst has access to  $K$  different estimators  $\hat{\theta}^1, \dots, \hat{\theta}^K$ . We assume that these estimators are asymptotically linear for estimating  $\theta^0$  with influence functions  $\phi^1(D), \dots, \phi^K(D)$ , i.e. that equation (8) holds. As discussed above, this can be justified for common estimators using the theory in Section 2. For expository simplicity, for now we assume that their influence functions  $\phi^1(D), \dots, \phi^K(D)$  are uncorrelated and have the same variance  $\sigma^2 > 0$  under  $\mathbb{P}$ . Later in the section, we discuss how to construct confidence intervals for general cases where influence functions are possibly correlated and have different variances. Since the estimators are asymptotically unbiased, uncorrelated, and have the same variance, as the final estimate we consider the mean of estimators,  $\hat{\theta}^{\text{pooled}} = \frac{1}{K} \sum_k \hat{\theta}^k$ . In the following, we will investigate the asymptotic behaviour of this estimator.

By Assumption 1, 2 and Lemma 1, for  $k = 1, \dots, K$ ,

$$\sqrt{n}(\hat{\theta}^k - \theta^0)_{k=1, \dots, K} \stackrel{d}{=} \delta\sigma(Z_k)_{k=1, \dots, K} + o_P(1),$$

where  $Z_k$  are independent standard normal random variables. Thus,

$$\sqrt{n}(\hat{\theta}^{\text{pooled}} - \theta^0) \stackrel{d}{=} \delta\sigma\bar{Z} + o_P(1). \quad (9)$$

On the other hand, define the between-estimator variance

$$\hat{\sigma}_{\text{bet}}^2 = \frac{1}{K-1} \sum_{k=1}^K (\hat{\theta}^k - \hat{\theta}^{\text{pooled}})^2.$$

Then,

$$\hat{\sigma}_{\text{bet}}^2 \stackrel{d}{=} \frac{\delta^2 \sigma^2}{n} \frac{1}{K-1} \sum (Z_k - \bar{Z})^2 + o_P(1/n) \stackrel{d}{=} \frac{\delta^2 \sigma^2}{n} \frac{\chi^2(K-1)}{K-1} + o_P(1/n), \quad (10)$$

where  $\chi^2(K-1)$  is a chi-square random variable with  $K-1$  degrees of freedom. Let us assume for a moment that  $\sigma^2$  is known to the data scientist. In this case, the data scientist may estimate  $\hat{\delta}^2$  via

$$\hat{\delta}^2 := \frac{n\hat{\sigma}_{\text{bet}}^2}{\sigma^2} \xrightarrow{d} \delta^2 \frac{\chi^2(K-1)}{K-1}.$$

Combining equations (9) and (10), we get

$$\frac{\hat{\theta}^{\text{pooled}} - \theta^0}{\hat{\sigma}_{\text{bet}}/\sqrt{K-1}} \xrightarrow{d} t(K-1),$$

where  $t(K-1)$  is a  $t$ -distributed random variable with  $K-1$  degrees of freedom. Note that  $\delta, \sigma$  cancel out.

Without direct estimation of  $\delta$  or  $\sigma$ , we have an  $1 - \alpha$  confidence interval of  $\theta^0$ :

$$\hat{\theta}^{\text{pooled}} \pm t_{K-1, 1-\alpha/2} \frac{\hat{\sigma}_{\text{bet}}}{\sqrt{K-1}}, \quad (11)$$

where  $t_{K-1, 1-\alpha/2}$  is the  $1 - \alpha/2$  quantile of the  $t$ -distribution with  $K-1$  degrees of freedom. Note that the size of the confidence intervals goes to zero with rate  $1/\sqrt{n}$  as  $\hat{\sigma}_{\text{bet}} = O_P(1/\sqrt{n})$ .

Let us make the argument in (11) more general. We will now discuss the case where the estimators  $\hat{\theta}^k$  have potentially different asymptotic variances  $\text{Var}_{\mathbb{P}}(\phi^k(D))$ . Instead of using  $\hat{\theta}^{\text{pooled}} = \frac{1}{K} \sum_k \hat{\theta}^k$  as the final estimate, we recommend inverse variance weighting. Thus, we first need to estimate  $\text{Var}_{\mathbb{P}}(\phi^k(D))$  consistently. While estimating  $\text{Var}_{\mathbb{P}}(\phi^k(D))$  is straightforward under i.i.d. sampling, we also have to verify that this works in our model class. We estimate  $\text{Var}_{\mathbb{P}}(\phi^k(D))$  using plug-in estimators of the influence function  $\hat{\phi}^k(D)$  as

$$\widehat{\text{Var}_{\mathbb{P}}(\phi^k(D))} = \frac{1}{n} \sum_{i=1}^n \left( \hat{\phi}^k(D_i) - \frac{1}{n} \sum_{i=1}^n \hat{\phi}^k(D_i) \right)^2. \quad (12)$$

The following proposition shows that  $\widehat{\text{Var}_{\mathbb{P}}(\phi^k(D))}$  is a consistent estimator of  $\text{Var}_{\mathbb{P}}(\phi^k(D))$ .

**Proposition 4 (Consistency of  $\widehat{\text{Var}_{\mathbb{P}}(\phi^k(D))}$ )** *Suppose that the  $\phi^k(D)$  has finite fourth moments. Furthermore, suppose that the estimation of the influence function is consistent in the sense that*

$$\frac{1}{n} \sum_{i=1}^n (\hat{\phi}^k(D_i) - \phi^k(D_i))^2 = o_p(1). \quad (13)$$

*Then,  $\widehat{\text{Var}_{\mathbb{P}}(\phi^k(D))}$  defined in (12) satisfies that*

$$\widehat{\text{Var}_{\mathbb{P}}(\phi^k(D))} = \text{Var}_{\mathbb{P}}(\phi^k(D)) + o_p(1).$$

**Remark 5 (OLS)** *Note that equation (13) is expected to hold for the plug-in estimators of the influence function under regularity assumptions. Revisiting Example 1, a plug-in estimator of the influence function is*

$$\hat{\phi}^k(D_i) = \left( \frac{1}{n} \sum_{j=1}^n X_j^{S_k} (X_j^{S_k})^\top \right)^{-1}_{1,\bullet} X_i^{S_k} (Y_i - (X_i^{S_k})^\top \hat{\theta}^{k,OLS}),$$

where  $\hat{\theta}^{k,OLS}$  is the OLS estimator computed with covariates  $X^{S^k}$ . One can now justify equation (13) via Lemma 2.

Now we construct asymptotically valid confidence intervals for  $\theta^0$  using  $K$  different estimators  $\hat{\theta}^1, \dots, \hat{\theta}^K$  that are asymptotically linear for estimating  $\theta^0$ . In the following theorem with Remark 7, influence functions of  $K$  different estimators can be correlated and have different variances.

**Theorem 6** *(Asymptotic validity of calibrated confidence interval). Suppose Assumption 1 and 2 hold and the influence functions  $\phi^1(D), \dots, \phi^K(D)$  are uncorrelated. Let  $\hat{\theta}^W = \sum_{k=1}^K \hat{\alpha}_k \hat{\theta}^k$  be the inverse-variance weighted estimator where the weights are*

$$\hat{\alpha}_k = \frac{\frac{1}{\widehat{\text{Var}}_{\mathbb{P}}(\phi^k(D))}}{\sum_{j=1}^K \frac{1}{\widehat{\text{Var}}_{\mathbb{P}}(\phi^j(D))}}, \quad (14)$$

with  $\widehat{\text{Var}}_{\mathbb{P}}(\phi^k(D)) = \text{Var}_{\mathbb{P}}(\phi^k(D)) + o_p(1)$  for  $k = 1, \dots, K$ . Let  $\hat{\sigma}_{bet}$  be the weighted between-estimator variance defined as

$$\hat{\sigma}_{bet}^2 = \sum_k \hat{\alpha}_k (\hat{\theta}^k - \hat{\theta}^W)^2.$$

Then for any  $\alpha \in (0, 1)$ , for fixed  $K$  and as  $n \rightarrow \infty$  we have

$$\mathbb{P} \left( \theta^0 \in \left[ \hat{\theta}^W \pm t_{K-1, 1-\alpha/2} \cdot \frac{\hat{\sigma}_{bet}}{\sqrt{K-1}} \right] \right) \rightarrow 1 - \alpha,$$

where  $t_{K-1, 1-\alpha/2}$  is the  $1-\alpha/2$  quantile of the  $t$  distribution with  $K-1$  degrees of freedom. To be clear, here we marginalize over both the randomness due to sampling and the randomness due to the distributional perturbation.

**Remark 7 (Correlated estimators)** *In practice, the components of  $(\phi^1(D), \dots, \phi^K(D))$  may be correlated. Then, we can apply a linear transformation to the estimators to obtain uncorrelated estimators that are asymptotically unbiased for  $\theta^0$ . We define the transformation matrix  $T_{ij} = \frac{(\hat{\Sigma}^{-1/2})_{ij}}{\sum_{j'} (\hat{\Sigma}^{-1/2})_{ij'}}$ , where  $\hat{\Sigma}$  is an estimate of the covariance matrix of  $\hat{\theta}^1, \dots, \hat{\theta}^K$ . We can then define  $(\hat{\eta}^1, \dots, \hat{\eta}^K)^\top = T \cdot (\hat{\theta}^1, \dots, \hat{\theta}^K)^\top$ . If  $\|\hat{\Sigma} - \Sigma\|_2 = o_p(1)$  and  $\Sigma$  is invertible, then the estimators  $\hat{\eta}^1, \dots, \hat{\eta}^K$  also satisfy Assumption 1 with influence functions that are pairwise uncorrelated. Furthermore, if the  $\hat{\theta}^k$ ,  $k = 1, \dots, K$  satisfy Assumption 2, then also  $\hat{\eta}^k$ ,  $k = 1, \dots, K$  satisfy Assumption 2.*

**Remark 8 (Meta-analysis on a single data set)** *The inverse variance-weighted estimate shares some similarity with a meta-analysis model. In traditional meta-analysis, one accounts for the random distributional variability of estimators obtained across different data sets. In contrast, our method accounts for the variability of multiple estimators obtained on a single data set under the random distribution shift model, where multiple estimators for a single target quantity are subject to shared symmetric distribution shifts. Thus, the random distribution shift model justifies a particular “meta-analysis on a single data set”. The idea that a study can potentially “replicate itself” has appeared in several communities, see the discussion in Section 3.2. In some of this literature, there is an emphasis that estimators should be subject to different biases. Analogously, in our approach, the estimators should have different influence functions, which implies that they will be affected differently by the random distribution shift.*

Below we present an algorithm box that provides a summary of Theorem 6 and Remark 7 to construct calibrated confidence intervals.

---

**Algorithm 1:** Constructing Calibrated Confidence Intervals

---

**Input:**  $K$  different estimators  $\hat{\theta}^1, \dots, \hat{\theta}^K$  and their estimated influence functions  $\hat{\phi}^1, \dots, \hat{\phi}^K$

**Output:** A calibrated confidence interval for  $\theta^0$

- 1 **if**  $\phi_1, \dots, \phi_K$  are correlated **then**
- 2     Estimate the transformation matrix  $T$  as in Remark 7.
- 3     Let  $(\hat{\theta}^1, \dots, \hat{\theta}^K)^\top \leftarrow T \cdot (\hat{\theta}^1, \dots, \hat{\theta}^K)^\top$ .
- 4     Let  $(\hat{\phi}^1, \dots, \hat{\phi}^K)^\top \leftarrow T \cdot (\hat{\phi}^1, \dots, \hat{\phi}^K)^\top$ .
- 5 Estimate the weights  $\hat{\alpha}_k$  for  $k = 1, \dots, K$  by Equation (14).
- 6 Compute the inverse-variance weighted estimator  $\hat{\theta}^W = \sum_{k=1}^K \hat{\alpha}_k \hat{\theta}^k$ .
- 7 Compute the weighted between-estimator variance  $\hat{\sigma}_{\text{bet}}^2 = \sum_{k=1}^K \hat{\alpha}_k (\hat{\theta}^k - \hat{\theta}^W)^2$ .
- 8 Return a calibrated confidence interval for  $\theta^0$  as

$$\hat{\theta}^W \pm t_{K-1, 1-\alpha/2} \cdot \frac{\hat{\sigma}_{\text{bet}}}{\sqrt{K-1}}.$$


---

In some cases, the data analyst may trust one of the estimators  $\hat{\theta}^k$  more than others. For example, the data analyst may be convinced that  $\theta^1 = \theta^0$  but may not be sure whether  $\theta^k = \theta^0$  for  $k \geq 2$ . In this case, it is possible to construct variance estimates that are upwardly biased in the sense that the resulting confidence intervals are expected to be conservative. The data analyst may report the confidence interval for  $\theta^0$  using  $\hat{\theta}^1$  instead of  $\hat{\theta}^W$  with  $\delta$  estimated by computing the between-estimator variance of the remaining  $K - 1$  estimators. As a result, they would lose one degree of freedom in their confidence intervals. The details can be found in Appendix C.

In the final variance estimate, there are two effects that are counteracting each other. Inverse-variance weighting reduces the variance of the final estimate compared to each of the individual estimators  $\hat{\theta}^k$ . On the other hand, the new variance formula accounts for distributional uncertainty and thus potentially inflates the variance.

### 3.2 Practical Implications for Stability Analyses

The proposed model not only leads to a recommendation on how to summarize the between-estimator uncertainty in confidence intervals but also lets us give some additional guidance.

First, note that if all estimators  $\hat{\theta}^k$  have similar influence functions, the proposed method will be unstable since in Remark 7 we invert the estimated covariance matrix. This coincides with the following intuition: Reporting that a large number of extremely similar estimators return similar results does not automatically increase the trustworthiness of a result. To corroborate a hypothesis one should have estimators that are susceptible to different sources of biases. In our model, this corresponds to estimators that are not highly correlated. Ideally, the estimators are independent. Similar arguments have appeared in other parts of the literature. For example, Rosenbaum (2021) writes: “An observational study has two evidence factors if it provides two comparisons susceptible to different biases that may be combined as if from independent studies of different data by different investigators, despite using the same data twice”.

In practice, it may happen that calibrated confidence intervals are very large compared to traditional sampling-based confidence intervals. Apart from the estimation error and small  $K$ , there are two possible explanations.

First, it could be that distributional uncertainty is very large. If distributional uncertainty is much larger than sampling uncertainty, conventional (unadjusted) confidence intervals are of limited value. Similar points have been made in different parts of the literature. For example, Meng (2018) argues that as the sample size grows, data quality becomes more important than data quantity and that standard confidence intervals have to be inflated to account for issues of data quality. In this vein, distributional confidence intervals can be used as a warning signal that we might be in a regime where data quality issues are more pressing than sampling uncertainty.

Secondly, it could be that the assumptions are violated (that means  $\theta^k \neq \theta^{k'}$  for some  $k, k'$ ). If the assumptions are grossly violated, inference will be more conservative. This is further detailed in Appendix C. If the assumptions are correct, inference will be more precise. In other words, the precision of calibrated inference depends on whether Assumption 2 is satisfied or not.

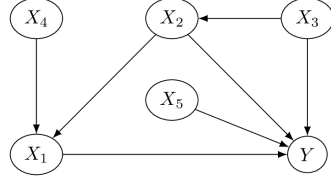
If the number of estimators  $K$  is very small, then there is an inferential price to pay for estimating distributional uncertainty in terms of power. This is reflected in the degrees of freedom of the  $t$ -distribution.

## 4. Simulation Study

In this section, we evaluate the performance of the proposed method via a simulation study. The marginal coverages of calibrated confidence intervals and the lengths of calibrated confidence intervals are evaluated on simulated data sets generated by random perturbation models. In this simulation, we emulate the situation where a data scientist uses linear regression with an adjustment set to estimate a causal effect.

**Setup.** The unperturbed distribution of  $D = (X, Y)$  with covariates  $X \in \mathbb{R}^5$  and response  $Y \in \mathbb{R}$  is generated from the following structural causal model (Bollen, 1989; Pearl, 2009):





$$\begin{aligned}
 \epsilon, \epsilon_1, \epsilon_2, X_3, X_4, X_5 &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \\
 X_2 &\leftarrow X_3 + \epsilon_2, \\
 X_1 &\leftarrow 0.5X_2 + X_4 + \epsilon_1, \\
 Y &\leftarrow X_1 + 0.5X_2 + X_3 + X_5 + \epsilon
 \end{aligned}$$

The goal is to estimate the direct causal effect of  $X_1$  on  $Y$ , which in this setup corresponds to the regression coefficient of  $X_1$  in a regression of  $Y$  on the set  $S = (X_1, X_2)$ . Practitioners often conduct such regressions for different choices of sets  $S$  to evaluate the overall stability of the procedure (Leamer, 1983; Oster, 2019).

In this example, the structural causal model can be used to construct multiple valid estimators. We look at the case where the data analyst considers  $K = 6$  different adjustment sets which all include the confounding variable  $X_2$ . In this case,  $K = 6$  different regression-adjusted estimators estimate the same quantity, the direct causal effect of  $X_1$  on  $Y$ , under the unperturbed distribution. We consider following adjustment sets;  $\{X_1, X_2, X_3\}$ ,  $\{X_1, X_2, X_5\}$ ,  $\{X_1, X_2, X_3, X_4\}$ ,  $\{X_1, X_2, X_3, X_5\}$ ,  $\{X_1, X_2, X_4, X_5\}$ ,  $\{X_1, X_2, X_3, X_4, X_5\}$ .

We now want to model a random shift between the target and the sampling distribution. We generate randomly perturbed data sets in two ways. First, we adopt the random perturbation model described in Lemma 1. We partition the support of the joint distribution of  $X$  and  $Y$  into  $m^{p+1}$  equal probability bins and perturb the probability of each bin with i.i.d. random weights  $Z \cdot W$  where  $W \sim \text{Gamma}(1, 1)$  and  $Z \sim \text{Ber}(1/m^p)$ . For sufficiently large  $m$ , this procedure can be seen as randomly selecting  $m$  bins out of  $m^{p+1}$  bins and perturbing the probability of each selected bin with i.i.d random weights  $W \sim \text{Gamma}(1, 1)$ . In our simulations, we generate  $n$  i.i.d. data points  $D_1, \dots, D_n$  from this randomly perturbed distribution. The strength of the perturbation is given as  $\delta^2 \approx 1 + 2 \cdot n/m$ . Secondly, we employ the random perturbation model described in Example 2 in the Appendix. Here, we sample  $m$  data points from the original distribution and let randomly perturbed distribution be the empirical distribution of  $m$  samples. The strength of the perturbation is given as  $\delta^2 \approx 1 + n/m$ .

Our method is carried out for sample sizes  $n = 200, 500, 1000$  and for  $m = 200, 500, 1000$ , which determines the strength of the perturbation, each with  $N = 1000$  replicates. In each replicate, we generate  $n$  samples from the randomly perturbed distribution, obtain  $K = 6$  different regression-adjusted estimators from the perturbed data set, and construct a calibrated  $(1 - \alpha)$  confidence interval using the inverse-variance weighted estimator according to Algorithm 1. We then evaluate the marginal coverage and length of the calibrated confidence interval and non-calibrated confidence intervals for each regression-adjusted estimator. While the direct estimation of  $\hat{\delta}^2$  is not required in our calibrated confidence intervals, we also include simulation results on the accuracy of  $\hat{\delta}^2$  in Appendix D.

#### 4.1 The Marginal Coverages of Calibrated Confidence Intervals

The marginal coverages of calibrated confidence intervals and non-calibrated confidence intervals are given in Figure 2. We see that calibrated confidence intervals have much improved coverage compared to non-calibrated confidence intervals, especially when  $n$  is large and  $m$  is small as the variance due to distributional perturbations dominates the marginal variance. In Appendix D, we additionally look at the case where the data analyst considers

$K = 8$  different adjustment sets including two additional sets  $\{X_1, X_2\}$  and  $\{X_1, X_2, X_4\}$ . In this case, some estimators are highly correlated, meaning that intuitively they are not distinct sources of evidence. This results in slight under-coverage, which highlights our advice that ideally one should use uncorrelated estimators to calibrate inference.

## 4.2 The Lengths of Calibrated Confidence Intervals

The boxplots of lengths of calibrated confidence intervals and non-calibrated confidence intervals are given in Figure 3. Figure 3 indicates that, perhaps surprisingly, calibrated confidence intervals can have even smaller lengths than non-calibrated confidence intervals, despite accounting for both distributional uncertainty and sampling uncertainty. This is due to inverse-variance weighting, which reduces the variance of the final estimate in comparison to each of the individual estimators. Note that the proportion of outliers marked as circles in boxplots is typically less than 5% for each boxplot. The distribution of the lengths of calibrated confidence intervals has a heavier tail than that of non-calibrated ones, as the former follows the square root of a scaled chi-square distribution with  $K - 1$  degrees of freedom.

## 5. Real-World Data Analysis

Ultimately, the goal of our procedure is to increase stability and trustworthiness of decision-making. In this section, we demonstrate that our method can improve stability on a real data set. We will see that even in situations without distributional perturbations, the proposed method can increase stability of decision-making. The data set (Cortez and Silva, 2008) was collected by using school reports and questionnaires to estimate final grades of students in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features. It is available at the UCI machine learning repository (Dheeru and Graff, 2017). We adopt 20 covariates in the data set. The response  $Y$  is the final year grade in Portuguese language. There are 649 students in total.

The goal is to determine the relative importance of  $L = 7$  selected binary covariates: 1) parents' cohabitation status, 2) whether the student received extra educational support from the school, 3) whether the student received family educational support, 4) whether the student is in a relationship, 5) whether the student had extra paid classes within the course subject, 6) whether the student's mother had secondary or higher education, and 7) whether the student's father had secondary or higher education. The relative importance is determined by the rank order of the covariates' effect sizes in a linear regression.

In the simulation setup, we aim to emulate a situation where as baseline the analyst has several reasonable choices to estimate a certain target quantity, and makes these decisions randomly. On the other hand, as a comparison, the analyst aggregates the estimators and conducts uncertainty quantification as proposed above. Ideally, the different estimators are driven by scientific background knowledge, as in the previous section. Here, for illustration purposes, we investigate the extreme case where background knowledge is very limited, that is, the statistician does not have strong preferences regarding which covariates to include in the regression.

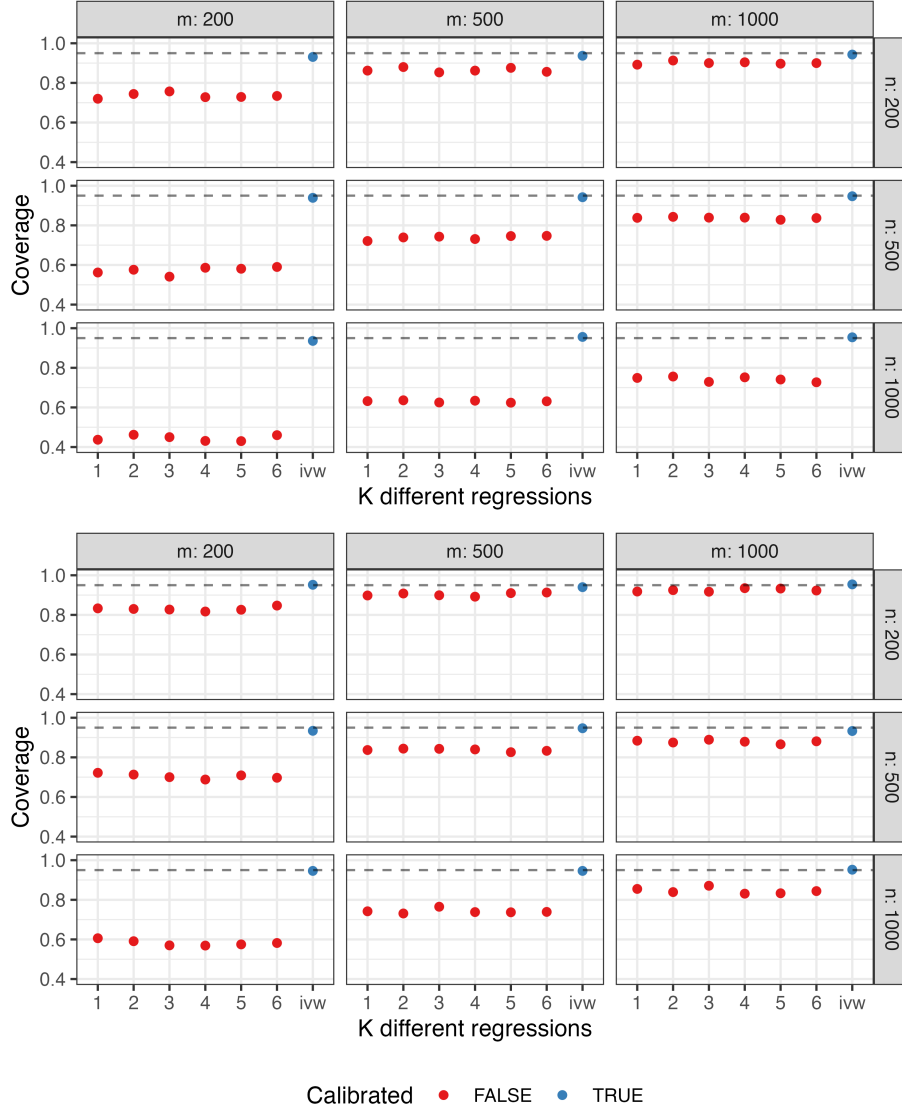


Figure 2: Marginal coverages of calibrated confidence intervals: The panel above shows the results under the perturbation model in Lemma 1 and the panel below shows the results under the perturbation model in Example 2 in the Appendix. Marginal coverages of non-calibrated confidence intervals for each regression-adjusted estimator and calibrated confidence intervals for the inverse-variance weighted estimator are presented for  $m = 200, 500, 1000$  and  $n = 200, 500, 1000$ . The strength of the perturbation is given as  $\delta^2 \approx 1 + n/m$ . The dashed lines indicate the nominal coverage 0.95.

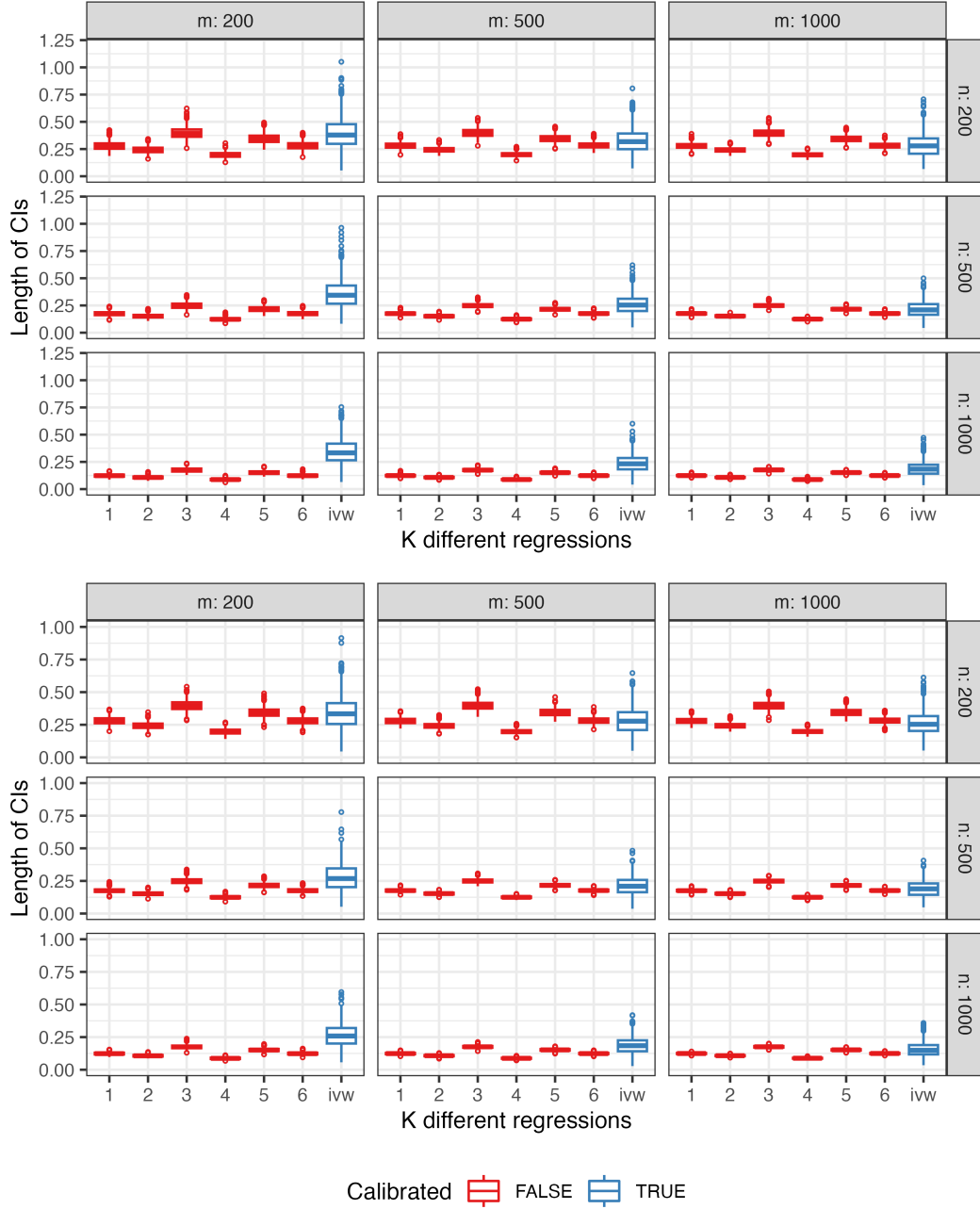


Figure 3: Lengths of calibrated confidence intervals: The panel above shows the results under the perturbation model in Lemma 1 and the panel below shows the results under the perturbation model in Example 2 in the Appendix. Boxplots of lengths of  $N = 1000$  non-calibrated confidence intervals for each regression-adjusted estimator and calibrated confidence intervals for the inverse-variance weighted estimator are presented for  $m = 200, 500, 1000$  and  $n = 200, 500, 1000$ .

	$\ell$	1	2	3	4	5	6	7
Method 1 (Non-Calibrated, $K = 10$ )	0.102	0.203	0.407	0.648	0.817	0.898	1.000	
Method 2 (Calibrated, $K = 10$ )	0.210	0.296	0.449	0.658	0.828	0.912	1.000	

	$\ell$	1	2	3	4	5	6	7
Method 1 (Non-Calibrated, $K = 20$ )	0.090	0.203	0.417	0.659	0.817	0.893	1.000	
Method 2 (Calibrated, $K = 20$ )	0.235	0.313	0.445	0.679	0.845	0.912	1.000	

Table 1: The stability of the ranking: The table above shows results with  $K = 10$  sets of covariates and the table below shows results with  $K = 20$  sets of covariates. Mean over  $N = 500$  iterations of the computed set similarity measure between  $S_{1,\ell}$  and  $S_{2,\ell}$  for each  $\ell = 1, \dots, 7$  is provided for each method.

Suppose we are given multiple sets of covariates, all containing the 7 binary covariates of our interest. We consider the following two methods. In method 1, a statistician randomly chooses one of the sets of covariates, performs a linear regression, and ranks the effect sizes of 7 covariates. In method 2, a statistician employs our proposed method. In particular, they perform linear regressions with multiple sets of covariates and for each covariate, calculate an inverse-variance weighted estimator and its effect size in consideration of distributional perturbations as described in Section 3. Then, they rank these effect sizes. Note that we use the additional constraint  $\hat{\delta} = \max(\hat{\delta}, 1)$  in our implementation.

We evaluate the two methods' stability in ranking effect sizes. To evaluate method  $i$ , we randomly split the data set into two, perform method  $i$  on each split, and compare the rankings resulting from each split. To measure the stability of the ranking, we compute the set similarity measure between  $S_{1,\ell} = \{\text{Top } \ell \text{ covariates by the effect size on split 1}\}$  and  $S_{2,\ell} = \{\text{Top } \ell \text{ covariates by the effect size on split 2}\}$  for each  $\ell = 1, \dots, L = 7$  as  $|S_{1,\ell} \cap S_{2,\ell}|/L$ . We repeat this procedure  $N = 500$  times and record the average set similarity measure. In each replicate, we randomly generate  $K = 10, 20$  sets of covariates that include the 7 covariates of our interest. The results can be found in Table 1. Overall, we see our method (Method 2) improves the stability of the ranking, notably outperforming Method 1 for  $\ell = 1, 2, 3$ . Note that the method 1 gives slightly worse results than random guessing for small  $\ell$ . One possible explanation is that sample splitting introduces small negative correlations between splits: If a regression coefficient is close to zero on the entire data set and on one split by chance the coefficient is large, then the coefficient is expected to be small on the other split.

Additionally, we compare lengths of calibrated and non-calibrated confidence intervals for each selected binary covariate using the full data set. From the results provided in Figure 4, one can see that our method is not so conservative given that we are adjusting confidence intervals with a scaling factor  $\hat{\delta}$ . Moreover, the variance of the length of calibrated confidence intervals tends to decrease as we increase the number of sets of covariates from  $K = 10$  to  $K = 20$ . For a more detailed look at the distribution of  $\hat{\delta}$ , the histograms of the confidence interval lengths are provided in Appendix E.

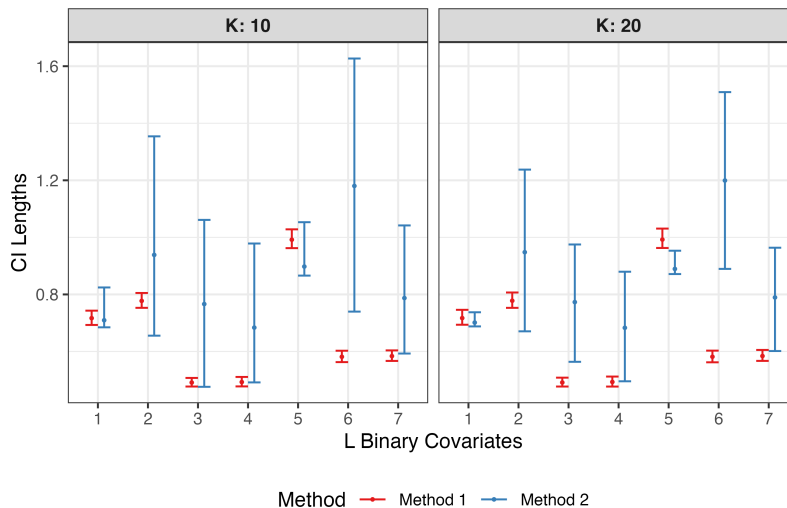


Figure 4: The lengths of calibrated and non-calibrated confidence intervals: Mean, 2.5% and 97.5% quantiles of the lengths of calibrated (Method 2) and non-calibrated (Method 1) confidence intervals for each selected binary covariate over  $N = 1000$  iterations are provided for  $K = 10$  and  $K = 20$ .

## 6. Discussion

In practice, data analysts often compute not just one estimator but multiple estimators for a single target quantity. For example, in causal inference, practitioners often consider multiple strategies to estimate the treatment effect. They could compute multiple regression-adjusted estimators for different choices of adjustment sets (see Example 1 and Section 4). If they believe that the treatment effect is homogeneous, they can derive several reasonable estimators for different subgroups of observations.

Often, it is recommended to study the estimator-to-estimator variability between sensible choices of estimators. If the estimator-to-estimator variability is high, then the analyst might have reason to not trust the estimates. In these cases, such stability investigations may be more informative than traditional  $p$ -values or confidence regions. This warrants an investigation of the theoretical properties of this practice. Does this practice have any guarantees and if so, which? Can we integrate this type of stability analysis into statistical inference?

We study a variant of this procedure from a distributional perspective. The data analyst may have access to multiple estimators, each purportedly estimating the same quantity, as justified by scientific background knowledge. In this context, estimator-to-estimator variability can be leveraged to scale confidence intervals. We show that these scaled confidence intervals account for both sampling uncertainty and distributional uncertainty within an isotropic perturbation model. Such uncertainty quantification seems desirable, especially in settings where the sampling uncertainty is of similar or lower order than other types of uncertainty.

This isotropic perturbation model is motivated by empirical phenomena in Figure 1. It assumes that the distribution shift is a superposition of many small random distributional changes.

The calibration procedure is not meant to replace existing methods that address confounding or selection bias via bias corrections, regression adjustment, or weighting procedures. Instead, our procedure can be used in conjunction with these methods as a second step that “calibrates” the confidence intervals.

The isotropic perturbation model is a strong assumption, but it is a weaker assumption than assuming that the data is i.i.d. from the target distribution, which is commonly made. Thus, the proposed calibration procedure works under strictly less assumptions on the data generating process than the most common inferential strategy. Instead of relying on i.i.d. sampling from  $\mathbb{P}$ , inference in the proposed model is based on a symmetry assumption and on scientific background knowledge for finding multiple reasonable estimators.

Of course, in practice perturbations might affect parts of the distribution differently. In such cases the proposed method can potentially have over-coverage or under-coverage. Looking forward, it would be desirable to extend the isotropic perturbation model (which has only one single parameter  $\delta$ ) to more flexible models that depend on multiple parameters. Such perturbation models would allow training different uncertainty models for different parts of the distribution, potentially leading to more realistic and flexible uncertainty quantification than existing approaches.

Furthermore, while our method currently operates with a single data set, a promising extension involves exploring scenarios with multiple perturbed data sets. When having access to multiple perturbed data sets, we can model the different data sets arising from the perturbed data generating distributions. Some first discussions about using and modeling multiple perturbed data sets can be found in Rothenhäusler and Bühlmann (2023) and Bansak et al. (2024).

A companion R package, `calinf`, is available at <https://github.com/rothenhaeusler/calinf>. Our package allows to draw data under the distributional uncertainty model and calibrate inference in generalized linear models. We provide an example of calibrated inference where the data analyst computes regression-adjusted estimators for different choices of adjustment sets. If multiple estimators are not available, it is also possible to estimate  $\delta$  using other types of scientific background knowledge. On the GitHub page, we discuss an example where the data analyst has background knowledge of population parameters.

## Acknowledgments

We thank the AE, three anonymous reviewers, Bin Yu, Peter Bühlmann, Guido Imbens, Xiao-Li Meng, Kevin Guo, Suyash Gupta, Ying Jin, and James Yang for helpful feedback and inspiring discussions. We are grateful for the support of the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and the Dieter Schwarz Foundation.

## Appendix

In Section A, we discuss additional properties of the isotropic perturbation model. Section B contains the proofs. Section C discusses how to form robust confidence intervals if the data analyst trusts one of the estimators  $\hat{\theta}^k$  more than others. Section D presents additional simulation results.

### Appendix A. Properties of The Isotropic Perturbation Model

Recall that conditionally on  $\xi$ , the data  $(D_i)_{1 \leq i \leq n}$  are drawn i.i.d. from the perturbed distribution  $\mathbb{P}^\xi(D = \bullet)$ , where  $\xi$  is an unobserved random variable. Note that an estimator  $\hat{\theta} = \hat{\theta}(D_1, \dots, D_n)$  for some parameter  $\theta^0(\mathbb{P})$  now has two sources of uncertainty: the uncertainty due to sampling and the uncertainty due to the random perturbation.

$$\hat{\theta} - \theta^0(\mathbb{P}) = \underbrace{\hat{\theta} - \theta(\mathbb{P}^\xi)}_{\text{variation due to sampling}} + \underbrace{\theta(\mathbb{P}^\xi) - \theta^0(\mathbb{P})}_{\text{variation due to random perturbation}}$$

We refer to the second component as *distributional uncertainty*. In this section we will study such distributional perturbation models in more detail. In Section A.1, we provide additional insights into the motivation behind the random distributional perturbation model. In Section A.2, we will show that under a strong symmetry assumption, there exists only one class of perturbation models that is characterized by a one-dimensional parameter  $\delta_{\text{dist}}$ . In Section A.3, we will sketch an extension of the random perturbation model that allows different parts of the distribution to be affected by different perturbations.

#### A.1 Additional Insights into the Isotropic Perturbation Model

Here we present additional insights into the weight-based distributional perturbation model. We draw inspiration from real-world examples presented in Figure 1 to construct the random perturbation model. A priori, there may be several mathematical random perturbation models leading to Figure 1. To simplify the discussion, in the following we will ignore sampling uncertainty. First, we will show that constant variance inflation implies random weights that are (almost) uncorrelated for disjoint events. Then, in the discrete setting, we will show that random weights imply constant variance inflation.

First, we study models that give rise to the constant variance inflation observed in Figure 1. To be precise, as working assumption we assume that for all square-integrable functions  $\psi(D)$  under  $\mathbb{P}$ ,

$$\text{Var}_P(\mathbb{E}^\xi[\psi(D)]) = \delta_{\text{dist}}^2 \text{Var}_{\mathbb{P}}(\psi(D)), \quad (15)$$

for some variance inflation factor  $\delta_{\text{dist}}^2$ . As before,  $\mathbb{P}$  refers to the unperturbed distribution of  $D$  and  $P$  refers to the marginal distribution of the perturbation and the observed data,



$(\xi, D_1, \dots, D_n)$ . Using equation (15), for all square-integrable functions  $\psi(D)$ ,  $\psi'(D)$ ,

$$\begin{aligned} & 2\text{Cov}_P(\mathbb{E}^\xi[\psi(D)], \mathbb{E}^\xi[\psi'(D)]) \\ &= \text{Var}_P(\mathbb{E}^\xi[\psi(D)] + \mathbb{E}^\xi[\psi'(D)]) - \text{Var}_P(\mathbb{E}^\xi[\psi(D)]) - \text{Var}_P(\mathbb{E}^\xi[\psi'(D)]) \\ &= \delta_{\text{dist}}^2 (\text{Var}_{\mathbb{P}}(\psi(D) + \psi'(D)) - \text{Var}_{\mathbb{P}}(\psi(D)) - \text{Var}_{\mathbb{P}}(\psi'(D))) \\ &= 2\delta_{\text{dist}}^2 \text{Cov}_{\mathbb{P}}(\psi(D), \psi'(D)). \end{aligned}$$

Thus, for disjoint  $D$ -measurable events  $A$  and  $B$  with  $\mathbb{P}(A) = \mathbb{P}(B) = 1/K$ ,

$$\text{Cov}_P(\mathbb{P}^\xi[A], \mathbb{P}^\xi[B]) = \text{Cov}_P(\mathbb{E}^\xi[1_A], \mathbb{E}^\xi[1_B]) = \delta_{\text{dist}}^2 \text{Cov}_{\mathbb{P}}(1_A, 1_B) = -\frac{\delta_{\text{dist}}^2}{K^2},$$

$$\text{Var}_P(\mathbb{P}^\xi[A]) = \text{Var}_P(\mathbb{E}^\xi[1_A]) = \delta_{\text{dist}}^2 \text{Var}_{\mathbb{P}}(1_A) = \delta_{\text{dist}}^2 \frac{1}{K} \left(1 - \frac{1}{K}\right).$$

Thus,  $\mathbb{P}^\xi[A]$  and  $\mathbb{P}^\xi[B]$  have the same variance and are marginally uncorrelated (ignoring lower order terms). Moreover, the right hand sides depend only on  $\delta_{\text{dist}}^2$  and  $K$ . This inspires us to construct a random perturbation model by initially partitioning the probability space into  $K$  disjoint bins with equal probability and then adjusting the probability of each partition with random weights constructed by positive i.i.d. random variables. As discussed in Section 2.2, empirical means are asymptotically Gaussian as the partitioning becomes finer, no matter how exactly the probability space was partitioned.

We will now go in the reverse direction. We will show that random re-weighting implies equation (15) in a simple discrete model. We will consider the simple example of a discrete uniform distribution  $\mathbb{P}(D = k) = \frac{1}{K}$  for  $k = 1, \dots, K$ . Let  $W_1, \dots, W_K$  be i.i.d. positive random variables with finite variance. We define the randomly perturbed distribution  $\mathbb{P}^\xi$  by setting

$$\mathbb{P}^\xi(D = k) = \frac{\xi_k}{K}, \tag{16}$$

where

$$\xi_k := \frac{W_k}{\sum_{k=1}^K W_k / K}.$$

Note that since  $W_1, \dots, W_K$  are positive i.i.d. random variables, the random perturbations  $\xi_1, \dots, \xi_K$  are exchangeable non-negative random variables that sum to  $\sum \xi_k = K$ . Then for  $1 \leq k_1 \neq k_2 \leq K$ , we have

$$\begin{aligned} \text{Cov}_P(\mathbb{P}^\xi[D = k_1], \mathbb{P}^\xi[D = k_2]) &= -\frac{\delta_{\text{dist}}^2}{K^2} \\ \text{Var}_P(\mathbb{P}^\xi[D = k_1]) &= \delta_{\text{dist}}^2 \frac{1}{K} \left(1 - \frac{1}{K}\right) \end{aligned}$$

where  $\delta_{\text{dist}}^2 := \frac{1}{K-1} \text{Var}(\xi_1)$ . We used that since  $\text{Var}(\sum_k \xi_k) = \text{Var}(K) = 0$ , we have  $\frac{1}{K-1} \text{Var}(\xi_1) = -\text{Cov}(\xi_1, \xi_2)$ . Moreover, for any function  $\psi : \{1, \dots, K\} \rightarrow \mathbb{R}$ ,

$$\begin{aligned} \text{Var}(\mathbb{E}^\xi[\psi(D)]) &= \text{Var}(\xi_1) \sum_k \frac{\psi(k)^2}{K^2} + \text{Cov}(\xi_1, \xi_2) \sum_{k_1 \neq k_2} \frac{\psi(k_1)\psi(k_2)}{K^2} \\ &= \delta_{\text{dist}}^2 \left( (K-1) \sum_k \frac{\psi(k)^2}{K^2} - \sum_{k_1 \neq k_2} \frac{\psi(k_1)\psi(k_2)}{K^2} \right) \\ &= \delta_{\text{dist}}^2 \left( \frac{1}{K} \sum_k \psi(k)^2 - \frac{1}{K^2} \left( \sum_k \psi(k) \right)^2 \right) \\ &= \delta_{\text{dist}}^2 \text{Var}_{\mathbb{P}}(\psi(D)). \end{aligned}$$

Thus, the random re-weighting model (16) implies equation (15).

## A.2 Uniqueness of Distributional Perturbation Model

We see that under the distributional perturbation model introduced earlier, the variance of the perturbation is proportional to the variance in the unperturbed distribution. This raises the question whether there are other “symmetric” random perturbation schemes that do not satisfy the variation inflation property in equation (15). The following result gives a negative answer to this question. We will see that under a symmetry assumption, there exists only one type of perturbation model, which is equivalent to the one in equation (15). Roughly speaking, the symmetry assumption states that two events that have equal probability under  $\mathbb{P}$  are perturbed in a similar fashion. In the following, we write  $\mathbb{Q}$  for the marginal distribution of  $(D, \xi)$ , where first the perturbation  $\xi$  is drawn and then  $D \sim \mathbb{P}^\xi$ . The proof of the following result can be found in Section B.5.

**Theorem 9 (Characterization of isotropic perturbation models)** *Let  $(D, \xi) \sim \mathbb{Q}$  and assume that there exists a function  $h(\bullet)$  such that  $h(D)$  is uniformly distributed on  $[0, 1]$ . Let  $\mathbb{P}^\xi = \mathbb{Q}(\bullet|\xi)$  and let  $\mathbb{P}$  denote the marginal distribution of  $D$  under  $\mathbb{Q}$ . Assume that for any  $D$ -measurable events  $A$  and  $B$  with  $\mathbb{P}(A) = \mathbb{P}(B)$ ,*

$$\text{Var}(\mathbb{P}^\xi(A)) = \text{Var}(\mathbb{P}^\xi(B)). \quad (17)$$

*Furthermore, assume that for every sequence of  $D$ -measurable events  $A_j$  with  $\mathbb{P}(A_j) \rightarrow 0$ ,*

$$\text{Var}(\mathbb{P}^\xi(A_j)) \rightarrow 0.$$

*Then for any  $\psi(D) \in L^2(\mathbb{P})$*

$$\text{Var}(\mathbb{E}^\xi[\psi(D)]) = \delta_{\text{dist}}^2 \text{Var}_{\mathbb{P}}(\psi(D)), \quad (18)$$

*for some fixed  $\delta_{\text{dist}} \geq 0$ .*

The assumption that  $h(D)$  exists is satisfied for any probability space that includes a continuous random variable. Thus, this is a regularity assumption that makes sure that the

probability space is “rich enough”. Let us discuss what this result means for the behaviour of empirical means. Let  $D_1, \dots, D_n$  be i.i.d. drawn from  $\mathbb{P}^\xi$ . Then, for all square-integrable functions  $\psi(D) \in L^2(\mathbb{P})$  marginally across sampling uncertainty and distributional uncertainty we have

$$\begin{aligned} \text{Var}_P \left( \frac{1}{n} \sum_{i=1}^n \psi(D_i) \right) &= \left( \frac{1}{n} + \delta_{\text{dist}}^2 - \frac{\delta_{\text{dist}}^2}{n} \right) \text{Var}_{\mathbb{P}}(\psi(D)) \\ &= \frac{\delta^2}{n} \text{Var}_{\mathbb{P}}(\psi(D)), \end{aligned}$$

with  $\delta^2 = 1 + n\delta_{\text{dist}}^2 - \delta_{\text{dist}}^2$ . Since  $\mathbb{P}^\xi = \mathbb{Q}(\bullet|\xi)$  we also have  $E_P[\frac{1}{n} \sum_{i=1}^n \psi(D_i)] = \mathbb{E}[\psi(D)]$ .

There are two major assumptions in this theorem. The first assumption says that two events that have the same probability are perturbed in the same fashion. This can be seen as a symmetry assumption. The second assumption says that events that have a small probability are only perturbed by a small amount. This can be seen as a regularity assumption.

Then, up to a one-dimensional parameter  $\delta$ , the variance of functions is uniquely determined. This means that using strong symmetry we have reduced the problem of estimating an infinite-dimensional perturbation model to a one-dimensional quantity  $\delta$ . Note that the statement in Theorem 9 is slightly weaker than Lemma 1, since it is only a statement about variances and not about the asymptotic distribution of  $\frac{1}{n} \sum_{i=1}^n \psi(D_i)$ .

In practice, some researchers might object to the symmetry assumption in equation (17). It turns out that the perturbation model can be generalized. In the following section, we will give a brief outlook of how perturbation models can be used to perturb different parts of a distribution differently.

### A.3 Beyond Isotropic Distributional Perturbations

The discussion in Section A.2 shows that under a strong symmetry assumption, up to an unknown scale factor  $\delta$ , there exists only one type of perturbation model. However, in practice there might be a situation where one does not expect a perturbation to affect all parts of the distribution in the same way. Consider  $D = (X, Y)$ . For example, one might expect that the distribution of  $X$  is perturbed between settings but that the measurement error is invariant. This may lead one to want to model a situation where  $p(x)$  is perturbed but  $p(y|x)$  is not perturbed. Under appropriate regularity conditions on  $\psi$  we have

$$\begin{aligned} \mathbb{E}^\xi[\psi(X, Y)] - \mathbb{E}[\psi(X, Y)] &= \mathbb{E}^\xi[\mathbb{E}[\psi(X, Y)|X]] - \mathbb{E}[\mathbb{E}[\psi(X, Y)|X]] \\ &\stackrel{d}{\approx} \mathcal{N}(0, \delta_{\text{dist}}^2 \text{Var}_{\mathbb{P}}(\mathbb{E}[\psi(X, Y)|X])). \end{aligned}$$

If  $\delta_{\text{dist}}$  is known or can be estimated, this allows us to adjust variance and confidence intervals to account for uncertainty both due to sampling and distributional perturbations, similarly as in Section 3.

## Appendix B. Proofs

### B.1 Auxiliary Results and Proof of Lemma 1

**Notation:** We write  $\mathbb{P}$  for the target distribution and  $\mathbb{P}^\xi$  for the randomly perturbed distribution from which we draw  $n$  i.i.d. data samples  $(D_i)_{i=1,\dots,n}$ . In both examples  $\xi$  can be seen as a random variable that encodes the perturbations. The expectation of  $f(D_1, \dots, D_n)$  over the joint distribution of  $(\xi, D_1, \dots, D_n)$  can be written as  $E_\xi[\mathbb{E}^\xi(f(D_1, \dots, D_n))]$  where  $E_\xi$  means we take the expectation over  $\xi$  and  $\mathbb{E}^\xi$  means that we take the expectation over  $(D_1, \dots, D_n)$ , conditionally on  $\xi$ .

#### B.1.1 AUXILIARY RESULTS

Let us first state an auxiliary lemma that will turn out helpful for proving Lemma 1.

**Lemma 10** *Let the assumptions of Lemma 1 hold. For the sequence of random variables  $\xi = \xi(n)$ , for any bounded  $\psi(\bullet)$  we have that*

$$\mathbb{E}^\xi[\psi(D)] - \mathbb{E}[\psi(D)] \stackrel{d}{=} \gamma_n \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))} \cdot Z + o_p(\gamma_n),$$

where  $Z$  follows a standard normal distribution and  $\gamma_n^2 = \frac{\text{Var}(W)}{m(n)E[W]^2}$ . Here we write  $m(n)$  to make it explicit that  $m$  grows with  $n$ .

**Proof** Let  $\phi = \psi \circ h$ . Without loss of generality, assume that  $\mathbb{E}[\phi(U)] = 0$ . Note that

$$\sqrt{m}(\mathbb{E}^\xi[\phi(U)] - \mathbb{E}[\phi(U)]) = \frac{\sqrt{m} \sum_{k=1}^m \int_{x \in I_k} \phi(x) dx \cdot (W_k - E[W])}{\sum_{k=1}^m W_k / m}.$$

Let

$$Y_{m,k} := \sqrt{m} \int_{x \in I_k} \phi(x) dx \cdot (W_k - E[W]).$$

First, note that

$$E[Y_{m,k}] = 0 \tag{19}$$

for all  $k$ . As the second step, we want to show that

$$\sum_{k=1}^m E[Y_{m,k}^2] = \text{Var}(W) \cdot m \sum_{k=1}^m \left( \int_{x \in I_k} \phi(x) dx \right)^2 \rightarrow \text{Var}(W) \cdot \text{Var}_{\mathbb{P}}(\phi(U)). \tag{20}$$

For any  $f \in L^2([0, 1])$ , define  $\Pi_m(f)$  as

$$\Pi_m(f)(x) = \sum_{k=1}^m \left( m \int_{x \in I_k} f(x) dx \right) \cdot I(x \in I_k).$$

Then, we have

$$\left| m \sum_{k=1}^m \left( \int_{x \in I_k} \phi(x) dx \right)^2 - \text{Var}_{\mathbb{P}}(\phi(U)) \right| = \|\phi - \Pi_m(\phi)\|_2^2 \rightarrow 0.$$

as  $m$  goes to infinity. This is because any bounded function can be approximated by a sequence of step functions of the form  $\sum_{k=1}^m b_k I(x \in I_k)$ . Next we will show that for any  $\epsilon > 0$ ,

$$g_m(\epsilon) = \sum_{k=1}^m E[Y_{m,k}^2; |Y_{m,k}| \geq \epsilon] \rightarrow 0. \quad (21)$$

This is implied by the dominated convergence theorem as

$$\begin{aligned} & \sum_{k=1}^m E[Y_{m,k}^2; |Y_{m,k}| \geq \epsilon] \\ & \leq \sum_{k=1}^m \left( \int_{x \in I_k} \phi^2(x) dx \right) E[(W_k - E[W])^2 I(\|\phi\|_\infty |W_k - E[W]|/\sqrt{m} \geq \epsilon)] \\ & = \|\phi\|_2^2 E[(W - E[W])^2 I(\|\phi\|_\infty |W - E[W]|/\sqrt{m} \geq \epsilon)] \rightarrow 0. \end{aligned}$$

Combining equations (19), (20), and (21), we can apply Lindeberg's CLT. With Slutsky's theorem, we have

$$\sqrt{m}(\mathbb{E}^\xi[\phi(U)] - \mathbb{E}[\phi(U)]) = \frac{\sum_{k=1}^m Y_{m,k}}{\sum_{k=1}^m W_k/m} \xrightarrow{d} \mathcal{N}(0, \text{Var}(W) \text{Var}_{\mathbb{P}}(\phi(U))/E[W]^2).$$

This completes the proof. ■

**Lemma 11** *Let the assumptions of Lemma 1 hold. Assume that for a sequence of random variables  $\xi = \xi(n)$  there exists a sequence  $\gamma_n$  with limit  $\delta^2 = \lim_n (1 + n\gamma_n^2) < \infty$  such that for any bounded  $\psi(\bullet)$  we have*

$$\mathbb{E}^\xi[\psi(D)] - \mathbb{E}[\psi(D)] \stackrel{d}{=} \gamma_n \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))} \cdot Z + o_p(\gamma_n), \quad (22)$$

where  $Z$  follows a standard normal distribution. Then, for any bounded  $\psi(\bullet)$ , it holds that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\psi(D_i^n) - \mathbb{E}[\psi(D)]) \xrightarrow{d} \mathcal{N}(0, \delta^2 \text{Var}_{\mathbb{P}}(\psi(D))).$$

**Proof** In the proof, we suppress the dependence of  $\xi$  on  $n$ . We want to show that for any  $x$ ,

$$E_\xi \left[ \mathbb{P}^\xi \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n (\psi(D_i^n) - \mathbb{E}[\psi(D)]) \leq x \cdot \sqrt{\delta^2 \text{Var}_{\mathbb{P}}(\psi(D))} \right) \right] = \Phi(x) + o(1),$$

where  $\Phi$  is the cdf of a standard Gaussian random variable. Let us define

$$Y_n = x \cdot \delta \cdot \frac{\sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}}{\sqrt{\text{Var}_{\mathbb{P}^\xi}(\psi(D))}} - \frac{\sqrt{n}(\mathbb{E}^\xi[\psi(D)] - \mathbb{E}[\psi(D)])}{\sqrt{\text{Var}_{\mathbb{P}^\xi}(\psi(D))}},$$

where  $\text{Var}_{\mathbb{P}^\xi}(\psi(D))$  denotes the variance of  $\psi(D)$  where  $D \sim \mathbb{P}^\xi$ . Then,

$$E_\xi \left[ \mathbb{P}^\xi \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n (\psi(D_i^n) - \mathbb{E}^\xi[\psi(D)]) + \sqrt{n}(\mathbb{E}^\xi[\psi(D)] - \mathbb{E}[\psi(D)]) \leq x \cdot \sqrt{\delta^2 \text{Var}_{\mathbb{P}}(\psi(D))} \right) \right]$$

$$= E_\xi \left[ \mathbb{P}^\xi \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\psi(D_i^n) - \mathbb{E}^\xi[\psi(D)]}{\sqrt{\text{Var}_{\mathbb{P}^\xi}(\psi(D))}} \leq Y_n \right) \right]. \quad (23)$$

We define  $g_n(y; \xi)$  as

$$g_n(y; \xi) = \mathbb{P}^\xi \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\psi(D_i^n) - \mathbb{E}^\xi[\psi(D)]}{\sqrt{\text{Var}_{\mathbb{P}^\xi}(\psi(D))}} \leq y \right).$$

By Berry–Esseen, it holds that

$$\sup_y \left| \mathbb{P}^\xi \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\psi(D_i) - \mathbb{E}^\xi[\psi(D)]}{\sqrt{\text{Var}_{\mathbb{P}^\xi}(\psi(D))}} \leq y \right) - \Phi(y) \right| \leq \frac{C \mathbb{E}^\xi |\psi(D)|^3}{(\mathbb{E}^\xi |\psi(D)|^2)^{3/2} \sqrt{n}},$$

for all  $n$ . Invoking equation (22) for  $\psi^2$  and  $\psi^3$ , we have that  $\mathbb{E}^\xi |\psi(D)|^3 / (\mathbb{E}^\xi |\psi(D)|^2)^{3/2}$  converges in probability to  $\mathbb{E} |\psi(D)|^3 / (\mathbb{E} |\psi(D)|^2)^{3/2} < \infty$  as  $n \rightarrow \infty$ . Then the right-hand side of the above inequality converges in probability to 0 as  $n \rightarrow \infty$ , which implies that

$$\sup_y |g_n(y; \xi) - \Phi(y)| \xrightarrow{P} 0.$$

Using this result,

$$\begin{aligned} (23) &= E_\xi[g_n(Y_n)] = E_\xi[g_n(Y_n)] - E_\xi[\Phi(Y_n)] + E_\xi[\Phi(Y_n)] \\ &\leq E_\xi[\sup_y |g_n(y) - \Phi(y)|] + E_\xi[\Phi(Y_n)] \\ &= E_\xi[\Phi(Y_n)] + o(1). \end{aligned}$$

Here, we used the dominated convergence theorem. Using equation (22),  $\text{Var}_{\mathbb{P}^\xi}(\psi(D)) \xrightarrow{P} \text{Var}_{\mathbb{P}}(\psi(D))$ . Then, we have

$$Y_n \xrightarrow{d} \delta x - \sqrt{\delta^2 - 1} Z,$$

where  $Z$  is a standard Gaussian random variable. Since  $\Phi$  is bounded and continuous, by Portmanteau Lemma, we get

$$\lim_{n \rightarrow \infty} E_\xi[\Phi(Y_n)] = E[\Phi(\delta x - \sqrt{\delta^2 - 1} Z)] = \Phi(x).$$

This completes the proof. ■

### B.1.2 PROOF OF LEMMA 1

Now let us show that the Lemma 1 holds.

**Proof** Without loss of generality for notational simplicity we restrict ourselves to the case  $l = 1$ , i.e.  $\psi : \mathcal{D} \mapsto \mathbb{R}$ . As before we write  $\phi(U) = \psi \circ h(U)$ . For any  $\psi \in L^2(\mathbb{P})$  and

for any  $\epsilon > 0$ , there exists a bounded function  $\psi^B$  such that  $\mathbb{E}[\psi(D)] = \mathbb{E}[\psi^B(D)]$  and  $\|\psi - \psi^B\|_{L^2(\mathbb{P})} < \epsilon$ . Note that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\psi(D_i^n) - \mathbb{E}[\psi(D)]) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\psi(D_i^n) - \psi^B(D_i^n)) \quad (\text{a})$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^n (\psi^B(D_i^n) - \mathbb{E}[\psi^B(D)]). \quad (\text{b})$$

Without loss of generality, let's assume that  $\mathbb{E}[\psi(D)] = 0$ . Note that

$$(\text{a}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( (\psi - \psi^B)(D_i^n) - \mathbb{E}^\xi[(\psi - \psi^B)(D)] \right) + \sqrt{n} \mathbb{E}^\xi[(\psi - \psi^B)(D)].$$

The marginal variance of its first part is

$$\begin{aligned} E[\text{Var}(\frac{1}{\sqrt{n}} \sum_{i=1}^n ((\psi - \psi^B)(D_i^n) - \mathbb{E}^\xi[(\psi - \psi^B)(D)])) | \xi] &\leq E_\xi[\mathbb{E}^\xi[(\psi - \psi^B)^2(D)]] \\ &= \mathbb{E}[(\psi - \psi^B)^2(D)] < \epsilon^2. \end{aligned}$$

Let's look at the second part of (a). Recall that we write  $\phi(U) = \psi \circ h(U)$ . Note that for any  $\phi \in L^2([0, 1])$  such that  $\mathbb{E}[\phi(U)] = 0$ ,

$$\sqrt{m}(\mathbb{E}^\xi[\phi(U)]) = \frac{\sqrt{m} \sum_{k=1}^m \int_{x \in I_k} \phi(x) dx \cdot (W_k - E[W])}{\sum_{k=1}^m W_k / m}.$$

With  $\phi = (\psi - \psi^B) \circ h$ , the variance of the numerator is bounded as

$$\begin{aligned} \text{Var}(W) \sum_{k=1}^m m \left( \int_{x \in I_k} \phi(x) dx \right)^2 &\leq \text{Var}(W) \sum_{k=1}^m \int_{x \in I_k} \phi^2(x) dx \\ &= \text{Var}(W) \mathbb{E}[\phi^2(U)] \\ &< \text{Var}(W) \cdot \epsilon^2, \end{aligned}$$

where the first inequality holds by Jensen's inequality with  $m \int_{x \in I_k} dx = 1$ . Therefore,

$$\sqrt{n} \mathbb{E}^\xi[(\psi - \psi^B)(D)] = \sqrt{r} \cdot \frac{\sqrt{m} \sum_{k=1}^m \int_{x \in I_k} \phi(x) dx \cdot (W_k - E[W])}{E[W]} + s_n$$

where  $s_n$  is  $o_p(1)$ . Combining results, we have that  $E[(\text{a}) - s_n] = 0$  and

$$\text{Var}_P((\text{a}) - s_n) \leq C \cdot \epsilon^2$$

for some constant  $C$ . Now we want to show that for any  $x$ ,

$$\lim_{n \rightarrow \infty} P \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{(\psi(D_i^n) - \mathbb{E}[\psi(D)])}{\delta \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}} \leq x \right) = \Phi(x)$$

where  $\Phi(x)$  is the cdf of a standard Gaussian random variable. Note that

$$\begin{aligned}
 & P\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{(\psi(D_i^n) - \mathbb{E}[\psi(D)])}{\delta \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}} \leq x\right) \\
 & \leq P\left(\frac{(b)}{\delta \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}} \leq x + 2\eta\right) + P\left(\frac{|(a)|}{\delta \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}} > 2\eta\right) \\
 & \leq P\left(\frac{(b)}{\delta \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}} \leq x + 2\eta\right) + P\left(\frac{|(a) - s_n|}{\delta \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}} > \eta\right) + P\left(\frac{|s_n|}{\delta \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}} > \eta\right) \\
 & \leq P\left(\frac{(b)}{\delta \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}} \leq x + 2\eta\right) + \frac{C \cdot \epsilon^2}{\eta^2 \delta^2 \text{Var}_{\mathbb{P}}(\psi(D))} + P\left(\frac{|s_n|}{\delta \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}} > \eta\right),
 \end{aligned}$$

where the last inequality holds by Chebyshev's inequality. With Lemma 10 and Lemma 11, we have that

$$(b) \xrightarrow{d} \delta N(0, \text{Var}_{\mathbb{P}}(\psi^B(D))).$$

Note that

$$\left(\sqrt{\text{Var}_{\mathbb{P}}(\psi(D))} - \sqrt{\text{Var}_{\mathbb{P}}(\psi^B(D))}\right)^2 \leq \mathbb{E}[(\psi - \psi^B)^2(D)] \leq \epsilon^2,$$

and thus

$$1 - \epsilon \cdot \frac{1}{\sqrt{\text{Var}_{\mathbb{P}}(\psi^B(D))}} \leq \frac{\sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}}{\sqrt{\text{Var}_{\mathbb{P}}(\psi^B(D))}} \leq 1 + \epsilon \cdot \frac{1}{\sqrt{\text{Var}_{\mathbb{P}}(\psi^B(D))}}.$$

Then, we get that

$$\begin{aligned}
 & \limsup_{n \rightarrow \infty} P\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{(\psi(D_i^n) - \mathbb{E}[\psi(D)])}{\delta \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}} \leq x\right) - \Phi(x) \\
 & \leq \Phi\left(\left(1 + \epsilon \cdot \frac{1}{\sqrt{\text{Var}_{\mathbb{P}}(\psi^B(D))}}\right)(x + 2\eta)\right) - \Phi(x) + \frac{C \cdot \epsilon^2}{\eta^2 \delta^2 \text{Var}_{\mathbb{P}}(\psi(D))}.
 \end{aligned}$$

Similarly, we can show that

$$\begin{aligned}
 & \liminf_{n \rightarrow \infty} P\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{(\psi(D_i^n) - \mathbb{E}[\psi(D)])}{\delta \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}} \leq x\right) - \Phi(x) \\
 & \geq \Phi\left(\left(1 - \epsilon \cdot \frac{1}{\sqrt{\text{Var}_{\mathbb{P}}(\psi^B(D))}}\right)(x - 2\eta)\right) - \Phi(x) - \frac{C \cdot \epsilon^2}{\eta^2 \delta^2 \text{Var}_{\mathbb{P}}(\psi(D))}.
 \end{aligned}$$

Note that results hold for arbitrary  $\eta > 0$  and  $\epsilon > 0$ . Let  $\eta = \sqrt{\epsilon}$ . Then for any  $x$ ,

$$\lim_{n \rightarrow \infty} P\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{(\psi(D_i^n) - \mathbb{E}[\psi(D)])}{\delta \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))}} \leq x\right) - \Phi(x) = 0.$$

This completes the proof. ■



## B.2 Examples of Variance Inflation Induced by Non-i.i.d. sampling

In the following examples we discuss how Assumption 1 with  $\delta \neq 1$  arises in non-standard sampling settings. For simplicity, we start with an artificial example: sampling with replacement from an unknown subpopulation.

**Example 2 (Sampling with replacement from an unknown subpopulation)** *Assume that  $D'_1, \dots, D'_m$  drawn i.i.d. from  $\mathbb{P}$ . Set  $\xi = (D'_1, \dots, D'_m)$ . We define the randomly perturbed distribution  $\mathbb{P}^\xi$  as the empirical measure*

$$\mathbb{P}^\xi(D \in \bullet) = \frac{1}{m} \sum_{i=1}^m 1_{D'_i \in \bullet}.$$

*Let  $n \rightarrow \infty$  and assume that  $m(n)$  is a sequence of integers such that  $\frac{n}{m(n)}$  converges to some limit  $r \in (0, \infty)$ . Conditionally on  $\xi$ , let  $(D_1^n, \dots, D_n^n)$  be i.i.d. draws from  $\mathbb{P}^\xi$ . Then equation (7) holds for any  $\psi(\bullet)$  with finite second moment with*

$$\delta^2 = 1 + r.$$

**Proof** Suppose that  $D'_1, \dots, D'_m$  are drawn from  $\mathbb{P}$  for some sequence  $m = m(n)$ . Let  $\mathbb{P}^\xi$  denote the empirical measure of  $D'_1, \dots, D'_m$ . Then by the CLT, for any  $\psi(\bullet)$  with finite second moment,

$$\begin{aligned} \mathbb{E}^\xi[\psi(D)] - \mathbb{E}[\psi(D)] &= \frac{1}{m} \sum_{i=1}^m \psi(D'_i) - \mathbb{E}[\psi(D)] \\ &\stackrel{d}{=} \gamma_n \sqrt{\text{Var}_{\mathbb{P}}(\psi(D))} \cdot Z + o_p(\gamma_n) \end{aligned}$$

where  $Z$  follows a standard normal distribution and  $\gamma_n^2 = 1/m(n)$ . By applying Lemma 11 and following the proof of Lemma 1, for any  $\psi(\bullet)$  with finite second moment, we get

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\psi(D_i^n) - \mathbb{E}[\psi(D)]) \xrightarrow{d} \mathcal{N}(0, \delta^2 \text{Var}_{\mathbb{P}}(\psi(D))),$$

where  $\delta^2 = 1 + r$ . ■

Sampling with replacement from a finite population might seem a bit artificial. The next example shows that a similar conclusion holds if we sample clusters, where units in a single cluster are highly correlated, and units between clusters are independent. If the cluster structure is known, one can use clustered standard errors. However, in general the dependence structure might be unknown.

**Example 3 (Sampling clusters with unobserved membership)** *Here, we consider a setting where some observations are associated, but where the overall dependence structure is unknown. This is similar to the previous setting, but there are no ties in the data set. Consider  $\mathbb{P}$  a probability distribution with positive density over a compact subset of  $\mathbb{R}^p$ .*

Draw i.i.d. observations  $D'_1, \dots, D'_m$  from  $\mathbb{P}$ . Set  $\xi = (D'_1, \dots, D'_m)$ . Conditionally on  $\xi$ , let  $(D_1^n, \dots, D_n^n)$  be i.i.d. draws from  $\mathbb{P}^\xi$ , where

$$\mathbb{P}^\xi(D \in \bullet) = \frac{1}{m} \sum_{i=1}^m \mathbb{P}(D \in \bullet \mid \|D'_i - D\|_2 \leq \epsilon_n),$$

where  $\epsilon_n > 0$  is a deterministic sequence with  $\epsilon_n = o(1/\sqrt{n})$ . Furthermore, let  $n \rightarrow \infty$  and assume that  $m = m(n)$  is a sequence of integers such that  $\frac{n}{m(n)}$  converges to some limit  $r \in (0, \infty)$ . Then, equation (7) holds for any bounded Lipschitz continuous  $\psi(\bullet)$  with

$$\delta^2 = 1 + r.$$

**Proof** Using Lipschitz continuity and  $\epsilon_n = o(1/\sqrt{n})$ , we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\psi(D_i^n) - \mathbb{E}[\psi(D)]) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\psi(D_i'') - \mathbb{E}[\psi(D)]) + o_p(1),$$

where the  $D_i''$  are drawn with replacement from  $D'_1, \dots, D'_m$ . We can now invoke Example 2.  $\blacksquare$

### B.3 Proof of Proposition 4

**Proof** Note that

$$\begin{aligned} \widehat{\text{Var}}_{\mathbb{P}}(\phi^k(D)) &= \underbrace{\frac{1}{n} \sum_{i=1}^n \left( \hat{\phi}^k(D_i) - \phi^k(D_i) - \frac{1}{n} \sum_{i=1}^n \hat{\phi}^k(D_i) + \frac{1}{n} \sum_{i=1}^n \phi^k(D_i) \right)^2}_{(i)} \\ &\quad + \underbrace{\frac{2}{n} \sum_{i=1}^n \left( \hat{\phi}^k(D_i) - \phi^k(D_i) - \frac{1}{n} \sum_{i=1}^n \hat{\phi}^k(D_i) + \frac{1}{n} \sum_{i=1}^n \phi^k(D_i) \right) \left( \phi^k(D_i) - \frac{1}{n} \sum_{i=1}^n \phi^k(D_i) \right)}_{(ii)} \\ &\quad + \underbrace{\frac{1}{n} \sum_{i=1}^n \left( \phi^k(D_i) - \frac{1}{n} \sum_{i=1}^n \phi^k(D_i) \right)^2}_{(iii)}. \end{aligned}$$

As the  $\phi^k$  has finite fourth moments, we can use Lemma 1 to obtain (iii) =  $\text{Var}_{\mathbb{P}}(\phi^k(D)) + o_p(1)$ . Then by Cauchy-Schwartz inequality and Jensen's inequality,

$$\begin{aligned} (i) &\leq \frac{2}{n} \sum_{i=1}^n \left( \hat{\phi}^k(D_i) - \phi^k(D_i) \right)^2 + 2 \left( \frac{1}{n} \sum_{i=1}^n \left( \hat{\phi}^k(D_i) - \phi^k(D_i) \right) \right)^2 \\ &\leq \frac{4}{n} \sum_{i=1}^n \left( \hat{\phi}^k(D_i) - \phi^k(D_i) \right)^2. \end{aligned}$$

Since our influence function estimators are consistent, (i) =  $o_p(1)$ . Then again by Cauchy-Schwartz inequality, (ii) =  $o_p(1)$ . Combining results, we get

$$\widehat{\text{Var}_{\mathbb{P}}(\phi^k(D))} = \text{Var}_{\mathbb{P}}(\phi^k(D)) + o_p(1).$$

This completes the proof. ■

#### B.4 Proof of Theorem 6

**Proof** By Assumption 1, 2 and Lemma 1,

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}^1 - \theta^0) \\ \vdots \\ \sqrt{n}(\hat{\theta}^K - \theta^0) \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} \phi^1(D_i) \\ \vdots \\ \phi^K(D_i) \end{pmatrix} + o_p(1) = \delta \mathbf{Z} + o_p(1),$$

where  $\mathbf{Z} = (Z_1, \dots, Z_K)^\top \sim \mathcal{N}(\mathbf{0}, \text{diag}(\text{Var}(\phi^1), \dots, \text{Var}(\phi^K)))$ . As  $n \rightarrow \infty$ , using that  $\sum \alpha_k = 1$  and  $\hat{\alpha}_k = \alpha_k + o_p(1)$ ,

$$\sqrt{n}(\hat{\theta}^W - \theta^0) = \sqrt{n} \sum_{k=1}^K \hat{\alpha}_k(\hat{\theta}^k - \theta^0) + o_p(1) = \delta \sum_{k=1}^K \alpha_k Z_k + o_p(1) \xrightarrow{d} \delta \mathcal{N}(0, \alpha), \quad (24)$$

where  $\alpha = \frac{1}{\sum_{k=1}^K \frac{1}{\text{Var}_{\mathbb{P}}(\phi^k(D))}}$ . By a similar calculation, we have that

$$n\hat{\sigma}_{bet}^2 = \delta^2 \sum_{k=1}^K \alpha_k (Z_k - \sum_{j=1}^K \alpha_j Z_j)^2 + o_p(1).$$

Thus,

$$n\hat{\sigma}_{bet}^2 = \delta^2 \alpha L_K + o_p(1), \quad (25)$$

where

$$L_K = \frac{1}{\alpha} \sum_{k=1}^K \alpha_k (Z_k - \sum_{j=1}^K \alpha_j Z_j)^2.$$

We will now show that

$$L_K \sim \chi^2(K-1) \quad \perp \quad \frac{1}{\sqrt{\alpha}} \sum_{j=1}^K \alpha_j Z_j \sim \mathcal{N}(0, 1). \quad (26)$$

First, note that

$$L_K = \sum_{k=1}^K \left( Z_k \frac{\sqrt{\alpha_k}}{\sqrt{\alpha}} - \frac{\sqrt{\alpha_k}}{\sqrt{\alpha}} \sum_{j=1}^K \alpha_j Z_j \right)^2.$$

With this definition,

$$\begin{aligned}
 L_K &= \sum_{k=1}^K \left( \frac{\sqrt{\alpha_k}}{\sqrt{\alpha}} Z_k - \frac{\sqrt{\alpha_k}}{\sqrt{\alpha}} \sum_{j=1}^K \alpha_j Z_j \right)^2 \\
 &= \sum_{k=1}^K \left( \frac{\sqrt{\alpha_k}}{\sqrt{\alpha}} Z_k - \sqrt{\alpha_k} \sum_{j=1}^K \sqrt{\alpha_j} \frac{\sqrt{\alpha_j}}{\sqrt{\alpha}} Z_j \right)^2 \\
 &= \sum_{k=1}^K (\tilde{Z}_k - w_k \sum_{j=1}^K w_j \tilde{Z}_j)^2,
 \end{aligned}$$

where  $\tilde{Z}_k := Z_k \sqrt{\alpha_k} / \sqrt{\alpha} = Z_k / \sqrt{\text{Var}_{\mathbb{P}}(\phi^k)}$  are i.i.d. standard normal and  $w_k := \sqrt{\alpha_k}$ . Please note that  $\sum_k w_k^2 = 1$ . Thus, we can write this equation

$$L_K = \|\tilde{Z} - w(w \cdot \tilde{Z})\|_2^2 = \|(\text{Id} - \Pi)\tilde{Z}\|_2^2,$$

where  $\Pi$  projects on the one-dimensional subspace spanned by  $w$ . Let  $b_1, \dots, b_{K-1}$  be an orthonormal basis of the span of  $\Pi$ . Then, by rotational invariance of the  $\ell_2$  norm,

$$L_K = \|(\text{Id} - \Pi)\tilde{Z}\|_2^2 = \sum_{k=1}^{K-1} (b_k \cdot \tilde{Z})^2.$$

Furthermore, since the  $b_k$  are orthogonal to each other  $b_k \cdot \tilde{Z}$  are independent standard Gaussians. Thus,  $L_K$  follows a  $\chi^2(K-1)$  distribution. Furthermore, since the  $b_k$  are orthogonal to  $w$ ,  $L_K$  is independent of

$$\sum_{k=1}^K w_k \tilde{Z}_k.$$

Furthermore, by definition

$$\frac{1}{\sqrt{\alpha}} \sum_{k=1}^K \alpha_k Z_k = \sum_{k=1}^K w_k \tilde{Z}_k \sim \mathcal{N}(0, 1)$$

and thus  $L_K$  is independent of  $\frac{1}{\sqrt{\alpha}} \sum_{k=1}^K \alpha_k Z_k$ . Therefore, (26) holds. Using (26) with (25) and (24), we get

$$\frac{\hat{\theta}^W - \theta^0}{\hat{\sigma}_{bet}/\sqrt{K-1}} = \frac{\delta \sum_{k=1}^K \alpha_k Z_k}{\sqrt{\alpha} \delta \sqrt{L_K/(K-1)}} + o_P(1) = \frac{\frac{\sum_k \alpha_k Z_k}{\sqrt{\alpha}}}{\sqrt{L_K/(K-1)}} + o_P(1) = T_{K-1} + o_P(1),$$

where  $T_{K-1}$  is a t-distributed random variable with  $K-1$  degrees of freedom. This completes the proof.  $\blacksquare$

### B.5 Proof of Theorem 9

**Proof** In this proof, if not specified otherwise, all variances and covariances are meant with respect to  $\mathbb{Q}$ , that is marginally over both the variation in  $D$  and  $\xi$ . We will directly work with  $U = h(D)$ . Define

$$f(x) = \text{Var}(\mathbb{P}^\xi(U \in [0, x])).$$

Let  $A$  and  $B$  be two disjoint subsets of  $[0, 1]$ . Define  $a = \mathbb{P}(U \in A)$  and  $b = \mathbb{P}(U \in B)$ . Then,

$$\begin{aligned} f(a+b) &= \text{Var}(\mathbb{P}^\xi(U \in A \cup B)) \\ &= \text{Var}(\mathbb{P}^\xi(U \in A)) + \text{Var}(\mathbb{P}^\xi(U \in B)) + 2\text{Cov}(\mathbb{P}^\xi(U \in A), \mathbb{P}^\xi(U \in B)) \\ &= f(a) + f(b) + 2\text{Cov}(\mathbb{P}^\xi(U \in A), \mathbb{P}^\xi(U \in B)). \end{aligned}$$

Thus, for any two disjoint sets  $A$  and  $B$ ,

$$\text{Cov}(\mathbb{P}^\xi(U \in A), \mathbb{P}^\xi(U \in B)) = \frac{f(a+b) - f(a) - f(b)}{2}.$$

Define

$$g(a, b) = \frac{f(a+b) - f(a) - f(b)}{2}. \quad (27)$$

Let us first show that  $f$  is continuous. Let  $a_n \rightarrow a$ ,  $a_n \geq a$ . Then,

$$f(a_n) - f(a) = f(a) + 2g(a_n - a, a) + f(a_n - a) - f(a) = 2g(a_n - a, a) + f(a_n - a).$$

By Cauchy-Schwartz,

$$g(a_n - a, a) \leq \sqrt{f(a_n - a)f(a)}.$$

By assumption,  $f(a_n - a) \rightarrow 0$ . Thus,  $f(a_n) \rightarrow f(a)$ . The case  $a_n \rightarrow a$ ,  $a_n \leq a$  can be treated analogously. Thus,  $f(\bullet)$  and  $g(\bullet, \bullet)$  are continuous.

Partition the probability space into disjoint  $D$ -measurable events  $A_i$ ,  $i = 1, \dots, n$  with  $P(U \in A_i) = 1/n$ . Then,

$$0 = \text{Var}(\mathbb{P}^\xi(\cup A_i) - \mathbb{P}(\cup A_i)) = nf(1/n) + n(n-1)g(1/n, 1/n).$$

Thus,

$$g(1/n, 1/n) = -1/(n-1)f(1/n). \quad (28)$$

We will now show that  $f(x) = x(1-x)\delta_{\text{dist}}^2$  for  $x = 1/2^k$ , where  $\delta_{\text{dist}}^2 = 4f(1/2)$ . This will show that up to the constant  $\delta_{\text{dist}}^2 = 4f(1/2)$ ,  $f$  and  $g$  are uniquely defined. First, we will show this equality for  $x = 1/4$ .

$$f(1/2) = f(1/4) + f(1/4) + 2g(1/4, 1/4) = 2f(1/4) - 2/3f(1/4).$$

Thus,

$$f(1/2) = 4/3f(1/4).$$

Rearranging,

$$f(1/4) = 1/4(1 - 1/4)4f(1/2) = x(1-x)\delta_{\text{dist}}^2$$

for  $x = 1/4$ . Induction step: Assume that

$$f(x) = x(1-x)\delta_{\text{dist}}^2$$

for  $x = 1/2^k$ . Now we want to show that

$$f(x/2) = x/2(1-x/2)\delta_{\text{dist}}^2.$$

To this end, using (27) and (28),

$$f(x) = f(x/2) + f(x/2) - 2/(2/x - 1)f(x/2).$$

Thus,

$$f(x) = (2 - 2x/(2 - x))f(x/2) = (4 - 2x - 2x)/(2 - x)f(x/2) = (4 - 4x)/(2 - x)f(x/2).$$

By induction assumption,

$$f(x/2) = (2 - x)/(4 - 4x)x(1 - x)\delta_{\text{dist}}^2 = x/2(1 - x/2)\delta_{\text{dist}}^2.$$

Thus, by induction for all  $x = 1/2^k$

$$f(x) = x(1 - x)\delta_{\text{dist}}^2.$$

Now we want to show that for any  $k$  and  $j \leq 2^k$  and  $x = j/2^k$

$$f(x) = x(1 - x)\delta_{\text{dist}}^2.$$

For any  $k$  and  $j$  with  $1 \leq j \leq 2^k$ , using the definition of  $f$  and (28),

$$\begin{aligned} f(j/2^k) &= jf(1/2^k) - j(j-1)/(2^k-1)f(1/2^k) = (j2^k - j^2)/(2^k-1)f(1/2^k) \\ &= (j2^k - j^2)/(2^k-1)1/2^k(1-1/2^k)\delta_{\text{dist}}^2 = (j2^k - j^2)1/2^k1/2^k\delta_{\text{dist}}^2 = j/2^k(1-j/2^k)\delta_{\text{dist}}^2. \end{aligned}$$

Thus, for all  $k$  and  $j \leq 2^k$ , and  $x = j/2^k$ ,

$$f(x) = x(1 - x)\delta_{\text{dist}}^2.$$

Using continuity of  $f$ , for all  $x \in [0, 1]$

$$f(x) = x(1 - x)\delta_{\text{dist}}^2.$$

We will now derive an explicit formula for  $g$ . For any  $k$  and  $j, j'$  with  $j + j' \leq 2^k$ ,

$$\begin{aligned} g(j/2^k, j'/2^k) &= jj'g(1/2^k, 1/2^k) = -jj'/(2^k-1)f(1/2^k) = -jj'/(2^k-1)1/2^k(1-1/2^k)\delta_{\text{dist}}^2 \\ &= -jj'1/2^k1/2^k\delta_{\text{dist}}^2 = -j/2^kj'/2^k\delta_{\text{dist}}^2. \end{aligned}$$

By continuity, for all  $x \geq 0, y \geq 0$  with  $x + y \leq 1$ ,

$$g(x, y) = -xy\delta_{\text{dist}}^2.$$

Now assume that for some  $D$ -measurable disjoint sets  $A_i$  and some constants  $y_i$ ,

$$\psi(D) = \sum 1_{A_i} y_i.$$

Then,

$$\text{Var}(\mathbb{E}^\xi[\psi(D)] - \mathbb{E}[\psi(D)]) = \sum_i y_i^2 f(P(A_i)) + \sum_{i \neq j} y_i y_j g(\mathbb{P}(A_i), \mathbb{P}(A_j)).$$

To simplify, let's write  $p_i = P(A_i)$ . Using explicit formulas for  $f$  and  $g$ ,

$$\text{Var}(\mathbb{E}^\xi[\psi(D)] - \mathbb{E}[\psi(D)]) = \sum_i \delta_{\text{dist}}^2 p_i (1 - p_i) - \sum_{i \neq j} \delta_{\text{dist}}^2 y_i y_j p_i p_j. \quad (29)$$

On the other hand,

$$\delta_{\text{dist}}^2 \text{Var}_{\mathbb{P}}(\psi(D)) = \delta_{\text{dist}}^2 \left( \sum_i p_i (1 - p_i) y_i^2 + \sum_{i \neq j} \text{Cov}(1_{A_i}, 1_{A_j}) y_i y_j \right).$$

As the sets are disjoint,  $\text{Cov}(1_{A_i}, 1_{A_j}) = -p_i p_j$ . Thus,

$$\delta_{\text{dist}}^2 \text{Var}_{\mathbb{P}}(\psi(D)) = \delta_{\text{dist}}^2 \left( \sum_i p_i (1 - p_i) y_i^2 + \sum_{i \neq j} -p_i p_j y_i y_j \right). \quad (30)$$

Combining equation (29) with equation (30),

$$\text{Var}(\mathbb{E}^\xi[\psi(D)] - \mathbb{E}[\psi(D)]) = \delta_{\text{dist}}^2 \text{Var}_{\mathbb{P}}(\psi(D)).$$

By measure-theoretic induction, this result is extended to any  $\psi(D) \in L^2(\mathbb{P})$ . ■

## B.6 Asymptotic Behaviour of $M$ -estimators

### B.6.1 PROOF OF LEMMA 2

**Proof** The proof follows Van der Vaart (2000), Theorem 5.14 with  $m_\theta(D) = -L(\theta, D)$ .

Fix some  $\theta$  and let  $U_\ell \downarrow \theta$  be a decreasing sequence of open balls around  $\theta$  of diameter converging to zero. Write  $m_U(D)$  for  $\sup_{\theta \in U} m_\theta(D)$ . The sequence  $m_{U_\ell}$  is decreasing and greater than  $m_\theta$  for every  $\ell$ . As  $\theta \rightarrow m_\theta(D)$  is continuous, by monotone convergence theorem, we have  $\mathbb{E}[m_{U_\ell}(D)] \downarrow \mathbb{E}[m_\theta(D)]$ .

For  $\theta \neq \theta^0$ , we have  $\mathbb{E}[m_\theta(D)] < \mathbb{E}[m_{\theta^0}(D)]$ . Combine this with the preceding paragraph to see that for every  $\theta \neq \theta^0$  there exists an open ball  $U_\theta$  around  $\theta$  with  $\mathbb{E}[m_{U_\theta}(D)] < \mathbb{E}[m_{\theta^0}(D)]$ . For any given  $\epsilon > 0$ , let the set  $B = \{\theta \in \Omega : \|\theta - \theta^0\|_2 \geq \epsilon\}$ . The set  $B$  is compact and is covered by the balls  $\{U_\theta : \theta \in B\}$ . Let  $U_{\theta_1}, \dots, U_{\theta_p}$  be a finite sub-cover. Then,

$$\begin{aligned} \sup_{\theta \in B} \frac{1}{n} \sum_{i=1}^n m_\theta(D_i^n) &\leq \sup_{j=1, \dots, p} \frac{1}{n} \sum_{i=1}^n m_{U_{\theta_j}}(D_i^n) \\ &= \sup_{j=1, \dots, p} \mathbb{E}[m_{U_{\theta_j}}(D)] + o_p(1) < \mathbb{E}[m_{\theta^0}(D)] + o_p(1), \end{aligned} \quad (31)$$

where for the equality we apply Lemma 1 with  $\psi(D) = m_{U_{\theta_j}}(D)$  for all  $j = 1, \dots, p$ .

If  $\hat{\theta} \in B$ , then

$$\sup_{\theta \in B} \frac{1}{n} \sum_{i=1}^n m_{\theta}(D_i^n) \geq \frac{1}{n} \sum_{i=1}^n m_{\hat{\theta}}(D_i^n) \geq \frac{1}{n} \sum_{i=1}^n m_{\theta^0}(D_i^n) - o_p(1),$$

where the last inequality comes from the definition of  $\hat{\theta}$ . Using Lemma 1 with  $\psi(D) = m_{\theta^0}(D)$ , we have

$$\frac{1}{n} \sum_{i=1}^n m_{\theta^0}(D_i^n) - o_p(1) = \mathbb{E}[m_{\theta^0}(D)] - o_p(1).$$

Therefore,

$$\{\hat{\theta} \in B\} \subset \left\{ \sup_{\theta \in B} \frac{1}{n} \sum_{i=1}^n m_{\theta}(D_i^n) \geq \mathbb{E}[m_{\theta^0}(D)] - o_p(1) \right\}.$$

By the equation (31), the probability of the event on the right hand side converges to zero as  $n \rightarrow \infty$ . This completes the proof.  $\blacksquare$

### B.6.2 PROOF OF LEMMA 3

**Proof** The proof follows Van der Vaart (2000), Theorem 5.41 with  $\Psi_n(\theta) = \frac{1}{n} \sum_{i=1}^n \partial_{\theta} L(\theta, D_i^n)$  and  $\dot{\Psi}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \partial_{\theta}^2 L(\theta, D_i^n)$ . By Taylor's theorem there exists a (random vector)  $\tilde{\theta}$  on the line segment between  $\theta^0$  and  $\hat{\theta}$  such that

$$0 = \Psi_n(\hat{\theta}) = \Psi_n(\theta^0) + \dot{\Psi}_n(\theta^0)(\hat{\theta} - \theta^0) + \frac{1}{2}(\hat{\theta} - \theta^0)^{\top} \ddot{\Psi}_n(\tilde{\theta})(\hat{\theta} - \theta^0).$$

By Lemma 1 with  $\psi(D) = \partial_{\theta}^2 L(\theta^0, D)$ , we have

$$\dot{\Psi}_n(\theta^0) = \frac{1}{n} \sum_{i=1}^n \partial_{\theta}^2 L(\theta^0, D_i^n) = \mathbb{E}[\partial_{\theta}^2 L(\theta^0, D)] + o_P(1). \quad (32)$$

By assumption, there exists a ball  $B$  around  $\theta^0$  such that  $\theta \rightarrow \partial_{\theta}^3 L(\theta, D)$  is dominated by a fixed function  $h(\cdot)$  for every  $\theta \in B$ . The probability of the event  $\{\hat{\theta} \in B\}$  tends to 1. On this event

$$\|\ddot{\Psi}_n(\tilde{\theta})\| = \left\| \frac{1}{n} \sum_{i=1}^n \partial_{\theta}^3 L(\tilde{\theta}, D_i^n) \right\| \leq \frac{1}{n} \sum_{i=1}^n h(D_i^n).$$

Using Lemma 1 with  $\psi(D) = h(D)$ , we get

$$\|\ddot{\Psi}_n(\tilde{\theta})\| \leq \frac{1}{n} \sum_{i=1}^n h(D_i^n) = O_P(1). \quad (33)$$

Combining (32) and (33) gives us

$$\begin{aligned} -\Psi_n(\theta^0) &= \left( \mathbb{E}[\partial_{\theta}^2 L(\theta^0, D)] + o_P(1) + \frac{1}{2}(\hat{\theta} - \theta^0)^{\top} O_P(1) \right) (\hat{\theta} - \theta^0) \\ &= (\mathbb{E}[\partial_{\theta}^2 L(\theta^0, D)] + o_P(1)) (\hat{\theta} - \theta^0). \end{aligned}$$



The probability that the matrix  $\mathbb{E}[\partial_\theta^2 L(\theta^0, D)] + o_P(1)$  is invertible tends to 1. Multiplying the preceding equation by  $\sqrt{n}$  and applying  $(\mathbb{E}[\partial_\theta^2 L(\theta^0, D)] + o_P(1))^{-1}$  left and right complete the proof. ■

## Appendix C. Robust Calibrated Inference

In some cases, the data analyst may trust one of the estimators  $\hat{\theta}^k$  more than others. For example, the data analyst may be convinced that  $\theta^1 = \theta^0$  but may not be sure whether  $\theta^k = \theta^0$  for  $k \geq 2$ . In this case, the data analyst may report the confidence interval for  $\theta^0$  using  $\hat{\theta}^1$  instead of  $\hat{\theta}^W$  with  $\delta$  estimated by looking at the between-estimator variance of the remaining  $K - 1$  estimators. Now we present how to build asymptotically valid confidence intervals in such cases.

**Theorem 12** (*Asymptotic validity of calibrated confidence interval*). *Suppose Assumption 1 holds for  $k = 1, \dots, K$  and the influence functions  $\phi^1(D), \dots, \phi^K(D)$  are uncorrelated. Suppose  $\theta^1 = \theta^0$  but  $\theta^k$  may not be  $\theta^0$  for  $k \geq 2$ . Furthermore assume that we have consistent estimates of the variances of influence functions such that  $\widehat{\text{Var}}_{\mathbb{P}}(\phi^k(D)) = \text{Var}_{\mathbb{P}}(\phi^k(D)) + o_P(1)$  for  $k = 1, \dots, K$ . Let  $\hat{\theta}^W = \sum_{k=2}^K \hat{\alpha}_k \hat{\theta}^k$  be the inverse-variance weighted estimator of  $K - 1$  estimators where the weights are*

$$\hat{\alpha}_k = \frac{\frac{1}{\widehat{\text{Var}}_{\mathbb{P}}(\phi^k(D))}}{\sum_{j=2}^K \frac{1}{\widehat{\text{Var}}_{\mathbb{P}}(\phi^j(D))}}.$$

Let  $\hat{\sigma}_{bet}$  be the weighted between-estimator variance of  $K - 1$  estimators defined as

$$\hat{\sigma}_{bet}^2 = \sum_{k=2}^K \hat{\alpha}_k (\hat{\theta}^k - \hat{\theta}^W)^2.$$

Then for any  $\alpha \in (0, 1)$ , it holds that as  $n \rightarrow \infty$ ,

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left( \theta^0 \in \left[ \hat{\theta}^1 \pm t_{K-2, 1-\alpha/2} \cdot \sqrt{\sum_{j=2}^K \frac{\widehat{\text{Var}}_{\mathbb{P}}(\phi^j(D))}{\widehat{\text{Var}}_{\mathbb{P}}(\phi^j(D))} \frac{\hat{\sigma}_{bet}}{\sqrt{K-2}}} \right] \right) \geq 1 - \alpha,$$

where  $t_{K-2, 1-\alpha/2}$  is the  $1-\alpha/2$  quantile of the  $t$  distribution with  $K-2$  degrees of freedom. To be clear, here we marginalize over both the randomness due to sampling and the randomness due to the distributional perturbation.

The resulting confidence intervals are expected to be conservative. Firstly, we lose one degree of freedom of the  $t$ -distribution. Secondly, we get an overcoverage if  $\theta^k \neq \theta^0$  for  $k \geq 2$ .

**Proof** If  $\theta^k \neq \theta^0$  for some  $k \geq 2$ , then by asymptotic linearity  $\hat{\sigma}_{bet}^2$  converges to some  $\tau^2 > 0$ . As in the proof of Theorem 6, we get  $\hat{\theta}^1 - \theta^0 = \mathcal{N}(0, \delta^2 \text{Var}_{\mathbb{P}}(\phi^1)/n) + o_P(1/\sqrt{n})$ . Since the variance estimates are consistent,

$$P \left( \theta^0 \in \left[ \hat{\theta}^1 \pm t_{K-2, 1-\alpha/2} \cdot \sqrt{\sum_{j=2}^K \frac{\widehat{\text{Var}_{\mathbb{P}}(\phi^j(D))}}{\widehat{\text{Var}_{\mathbb{P}}(\phi^j(D))}} \frac{\hat{\sigma}_{bet}}{\sqrt{K-2}}} \right] \right) \rightarrow 1.$$

Now let us consider the case  $\theta^0 = \theta^1 = \dots = \theta^K$ . From the proof of Theorem 6, we know that

$$\frac{\sqrt{n}(\hat{\theta}^1 - \theta^0)}{\sqrt{\widehat{\text{Var}_{\mathbb{P}}(\phi^1(D))}}} \stackrel{d}{=} \delta Z + o_p(1),$$

where  $Z \sim N(0, 1)$ . Moreover,

$$n\hat{\sigma}_{bet}^2 \stackrel{d}{=} \delta^2 \frac{1}{\sum_{j=2}^K \frac{1}{\widehat{\text{Var}_{\mathbb{P}}(\phi^j(D))}}} \cdot L_{K-1} + o_p(1),$$

where  $L_{K-1}$  follows the chi-square distribution with  $K-2$  degrees of freedom. Note that  $Z$  and  $L_{K-1}$  are independent. Then, we get

$$\frac{\frac{\hat{\theta}^1 - \theta^0}{\sqrt{\widehat{\text{Var}_{\mathbb{P}}(\phi_1)}}}{\hat{\sigma}_{bet} \sqrt{\sum_{j=2}^K \frac{1}{\widehat{\text{Var}_{\mathbb{P}}(\phi_k)}} / \sqrt{K-2}}} \stackrel{d}{\rightarrow} \frac{Z}{\sqrt{L_{K-1}/(K-2)}}.$$

Thus, we get

$$\lim_{n \rightarrow \infty} P \left( \frac{\frac{\hat{\theta}^1 - \theta^0}{\sqrt{\widehat{\text{Var}_{\mathbb{P}}(\phi_1)}}}{\hat{\sigma}_{bet} \sqrt{\sum_{j=2}^K \frac{1}{\widehat{\text{Var}_{\mathbb{P}}(\phi_k)}} / \sqrt{K-2}}} \leq x \right) = P(t_{K-2} \leq x),$$

where  $t_{K-2}$  is a  $t$ -distributed random variable with  $K-2$  degrees of freedom. This completes the proof.  $\blacksquare$

## Appendix D. Additional Simulation Results

In this section, we include additional simulation results.

**Accuracy of  $\hat{\delta}^2$ .** The estimation accuracy of  $\hat{\delta}$  compared to the ground truth  $\delta$  is illustrated in Figure 5. Recall that the distribution of  $\hat{\delta}^2$  follows  $\delta^2 \cdot \frac{\chi^2(K-1)}{K-1}$  as  $n, m \rightarrow \infty$ .

**Marginal Coverages of Calibrated Confidence Intervals with Highly Correlated Estimators.** Instead of using  $K = 6$  adjustment sets in the main text, we consider the following  $K = 8$  adjustment sets;  $\{X_1, X_2\}$ ,  $\{X_1, X_2, X_3\}$ ,  $\{X_1, X_2, X_4\}$ ,  $\{X_1, X_2, X_5\}$ ,  $\{X_1, X_2, X_3, X_4\}$ ,  $\{X_1, X_2, X_3, X_5\}$ ,  $\{X_1, X_2, X_4, X_5\}$ ,  $\{X_1, X_2, X_3, X_4, X_5\}$ . The results are depicted in Figure 6. In this case, some estimators are highly correlated, resulting in slight under-coverage of calibrated confidence intervals.

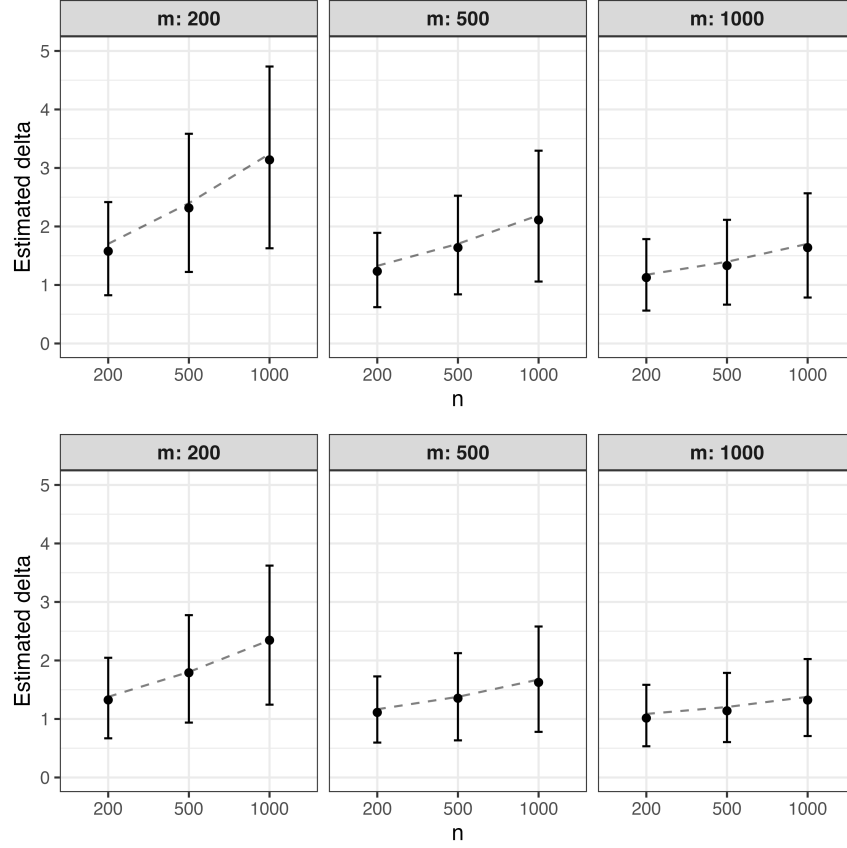


Figure 5: Accuracy of  $\hat{\delta}$ : The panel above shows the results under the perturbation model in Lemma 1 and the panel below shows the results under the perturbation model in Example 2 in the Appendix. Mean, 5%, and 95% quantiles of the estimated  $\hat{\delta}$  for each  $m = 200, 500, 1000$  and  $n = 200, 500, 1000$  are provided. The dashed lines indicate the true values of  $\delta$ .

## Appendix E. Additional Data Analysis Results

In this section, we present additional results from real-world data analysis. Below are the figures showing the histograms of the lengths of confidence interval from Section 5.

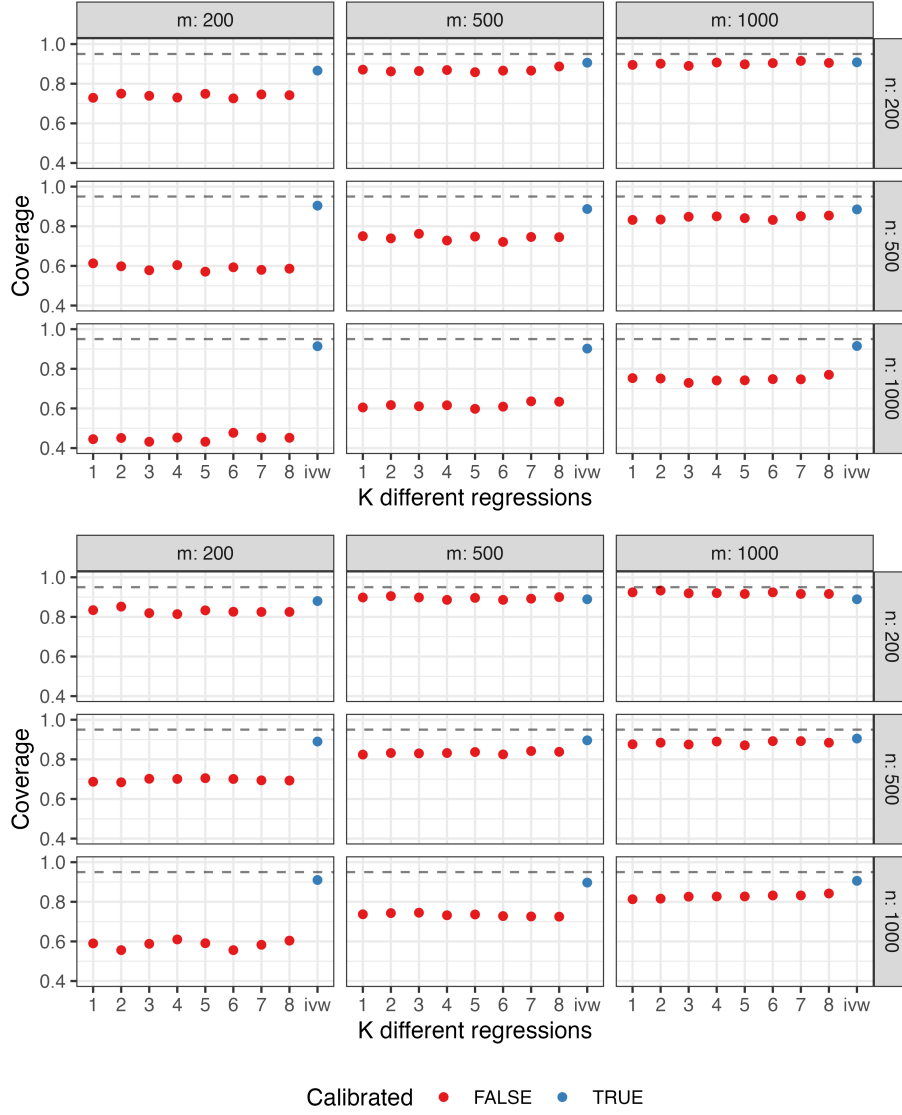


Figure 6: Marginal coverages of calibrated confidence intervals with  $K = 8$  different adjustment sets: The panel above shows the results under the perturbation model in Lemma 1 and the panel below shows the results under the perturbation model in Example 2 in the Appendix. Marginal coverages of non-calibrated confidence intervals for each regression-adjusted estimator and calibrated confidence intervals for the inverse-variance weighted estimator are presented for  $m = 200, 500, 1000$  and  $n = 200, 500, 1000$ . The dashed lines indicate the nominal coverage 0.95.

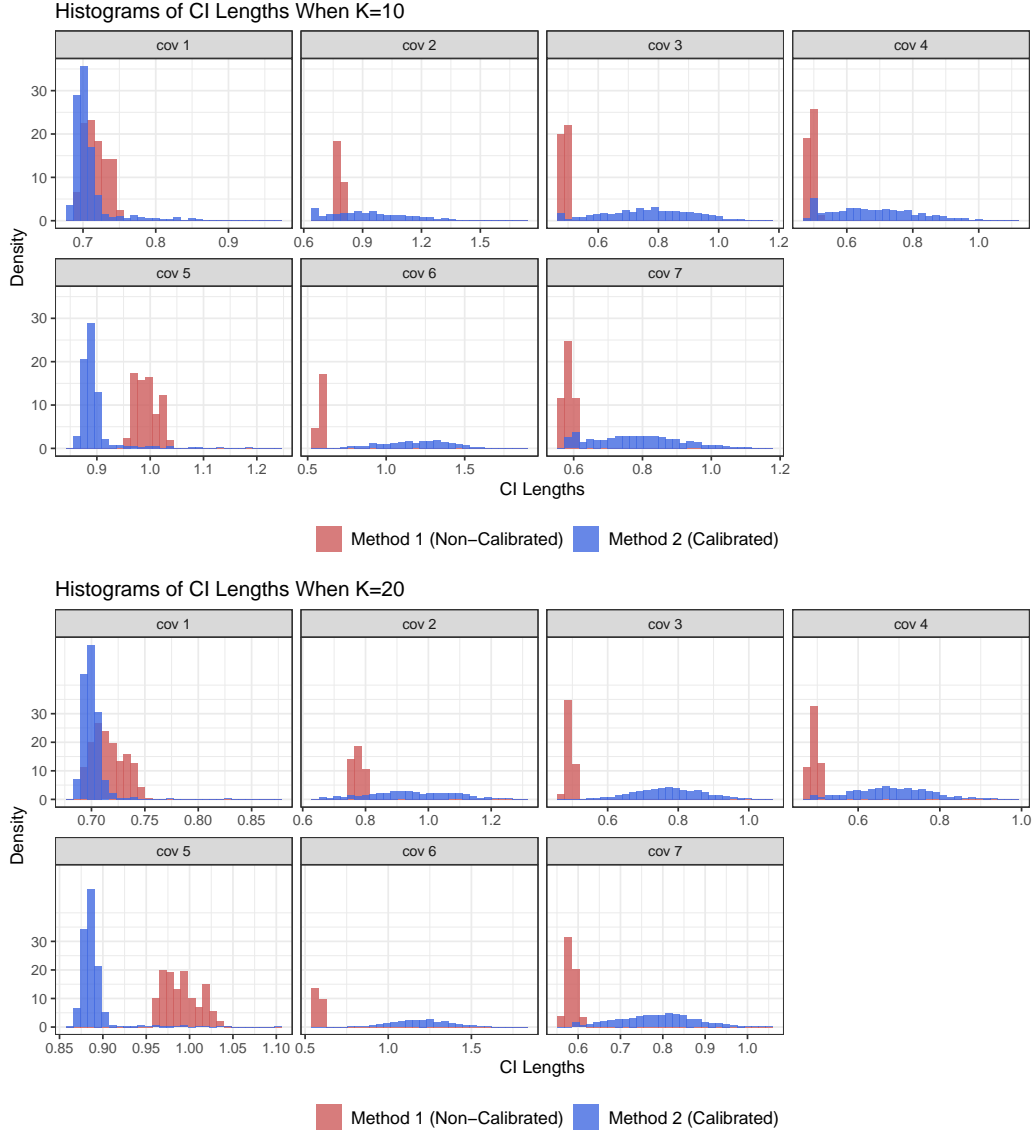


Figure 7: The histograms of the lengths of calibrated and non-calibrated confidence intervals for each selected binary covariate, based on  $N = 1000$  iterations, are provided for  $K = 10$  and  $K = 20$ .

## References

- K. Bansak, E. Paulson, and D. Rothenhäusler. Learning under random distributional shifts. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 238:3943–3951, 2024.
- A. Ben-Tal, D. Den Hertog, A. De Waegenare, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.
- K.A. Bollen. *Structural Equations with latent variables*. John Wiley & Sons, 1989.
- G. E. P. Box. Sampling and bayes’ inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A*, 143(4):383–404, 1980.
- L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- P. Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.
- P. Chiappori, S. Orefice, and C. Quintana-Domeque. Fatter attraction: anthropometric and socioeconomic matching on the marriage market. *Journal of Political Economy*, 120(4):659–695, 2012.
- J. Cornfield, W. Haenszel, E. C. Hammond, A. Lilienfeld, M. Shimkin, and E. Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22(1):173–203, 1959.
- P. Cortez and A. Silva. Using data mining to predict secondary school student performance. *Proceedings of 5th Future Business Technology Conference*, pages 5–12, 2008.
- N. K. Denzin. *The research act: A theoretical introduction to sociological methods*. Aldine Publishing Company, 1970.
- D. Dheeru and C. Graff. UCI machine learning repository, 2017. URL <http://archive.ics.uci.edu/ml>.
- T. Drautzburg. Why are recessions so hard to predict? random shocks and business cycles. Technical report, Federal Reserve Bank of Philadelphia, 2019.
- J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.
- R. M. Dudley. *Real Analysis and Probability*. CRC Press, 2018.
- D. Freedman. Statistical models and shoe leather. *Sociological Methodology*, pages 291–313, 1991.
- C. Heinze-Deml and N. Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

- P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.
- G. Imbens and D. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.
- B. Karmakar, B. French, and D. S. Small. Integrating the evidence from evidence factors in observational studies. *Biometrika*, 106(2):353–367, 2019.
- R. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.
- E. Leamer. Let’s take the con out of econometrics. *The American Economic Review*, 73(1): 31–43, 1983.
- X.L. Meng. Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, 12(2):685–726, 2018.
- M. Munafò and G. Smith. Repeating experiments is not enough. *Nature*, 553(7689):399–401, 2018.
- E. Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.
- C. J. Patel, B. Burford, and J. P. A. Ioannidis. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9):1046–1058, 2015.
- J. Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2nd edition, 2009.
- J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society, Series B*, 78(5):947–1012, 2016.
- N. Pfister, E. G. Williams, J. Peters, R. Aebersold, and P. Bühlmann. Stabilizing variable selection and regression. *The Annals of Applied Statistics*, 15(3):1220–1246, 2021.
- J. Prochazka, K. Parilakova, P. Rudolf, V. Bruk, R. Jungwirthova, S. Fejtova, R. Masaryk, and M. Vaculik. Pain as social glue: A preregistered direct replication of experiment 2 of bastian et al. (2014). *Psychological Science*, 33(3):463–473, 2022.
- M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Causal transfer in machine learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.
- P. Rosenbaum. The role of a second control group in an observational study. *Statistical Science*, 2(3):292–306, 1987.
- P. Rosenbaum. Evidence factors in observational studies. *Biometrika*, 97(2):333–345, 2010.

- P. Rosenbaum. *Replication and Evidence Factors in Observational Studies*. CRC Press, 2021.
- P. Rosenbaum and D. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.
- D. Rothenhäusler and P. Bühlmann. Distributionally robust and generalizable inference. *Statistical Science*, 38(4):527–542, 2023.
- D. Rothenhäusler, C. Heinze, J. Peters, and N. Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 1513–1521, 2015.
- D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.
- B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J.M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–1262, 2012.
- A. Skene, J. Shaw, and T. Lee. Bayesian modelling and sensitivity analysis. *Journal of the Royal Statistical Society*, 35(2):281–288, 1986.
- N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016.
- A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.
- A.W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000.
- T. VanderWeele and P. Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.
- R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447, 1974.
- B. Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013.
- B. Yu and K. Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, 2020.
- K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827, 2013.