Statistical Science
2023, Vol. 38, No. 4, 527–542
https://doi.org/10.1214/23-STS902

Distributionally Robust and Generalizable Inference

Dominik Rothenhäusler and Peter Bühlmann

Abstract. We discuss recently developed methods that quantify the stability and generalizability of statistical findings under distributional changes. In many practical problems, the data is not drawn i.i.d. from the target population. For example, unobserved sampling bias, batch effects, or unknown associations might inflate the variance compared to i.i.d. sampling. For reliable statistical inference, it is thus necessary to account for these types of variation. We discuss and review two methods that allow to quantify distribution stability based on a single dataset. The first method computes the sensitivity of a parameter under worst-case distributional perturbations to understand which types of shift pose a threat to external validity. The second method treats distributional shifts as random which allows to assess average robustness (instead of worst-case). Based on a stability analysis of multiple estimators on a single dataset, it integrates both sampling and distributional uncertainty into a single confidence interval.

Key words and phrases: Distributional robustness, external validity, generalizability, stability, uncertainty quantification.

1. INTRODUCTION

Uncertainty quantification and inference in terms of confidence statements in complex models has been a core topic in statistics over many decades. In the last 10 years, substantial progress has been made for high-dimensional and complex models, and we will briefly review these developments in Section 1.2. The main focus of this paper is different though, namely about generalizability and external validity of statistical findings and its corresponding inference. In ordinary language: if a statistical result is significant in a study (i.e., a dataset), to what extent can it be expected to be significant in another study which is similar but not exactly of the same nature as the original one? This question and corresponding solutions can be mathematically formalized, and we will describe them in Sections 2–4. Such generalizability and external validity of statistical inference is often of major interest in the context of empirical studies in, for example, medicine, public health, or economics [60, 67].

To judge the generalizability and trustworthiness of a statistical result, it is crucial to investigate the fragility of the analysis. Yu and Kumbier [70] discuss different types of perturbations that can be injected in the analysis process. If multiple datasets are available, one can adjust inference to account for the fact that the target population is different from the population at hand. As an example, [15] consider the problem of transporting inferences from multiple randomized trials to a new target population: the new population, which is not among the observed multiple datasets and potentially of slightly or moderately different nature is the one for which we want to generalize to.

Having statistical inference tools which are externally valid for somewhat different populations than the ones in the data is a crucial component for improving replicability of statistical (and scientific) results. The famous article by John Ioannidis [35] on the replicability crisis mentions major issues about biases from reporting and different protocols. Distributionally robust statistical procedures for confidence statements can be useful for addressing an aspect of the replication problem, without explicitly aiming to understand its possibly very diverse set of underlying reasons.

1.1 Internal and External Validity

We consider the setting where the data are realizations from either a single data-generating distribution P' or a set of data-generating distributions $\{P'_e; e \in \mathcal{E}\}$ where e is an index for a subpopulation and $\mathcal{E}$ is the space of observed sup-populations in the data. We typically assume

Dominik Rothenhäusler is Assistant Professor, Department of Statistics, Stanford University, Stanford, California 94305-4020, USA (e-mail: rdominik@stanford.edu). Peter Bühlmann is Professor, Seminar for Statistics, ETH Zürich, Switzerland (e-mail: buhlmann@stat.math.ethz.ch).

that the data are i.i.d. or independent realizations depending on fixed covariates from these distribution(s) P' (or P'_e); but the framework also includes sampling from a stochastic process or structured sampling in mixed effects models.

An inferential statistical statement for a parameter $\theta(\cdot)$ is called *internally valid* if it is statistically valid (or correct) for $\theta(P')$ or $\theta(P'_e)$ for some $e \in \mathcal{E}$. Note that the parameter is a functional of the distribution P, P_e or P'. Thus, the parameter of interest $\theta(\cdot)$ is a functional of a data generating distribution from which the observed data arises. On the other hand, *external validity* is concerned about a parameter $\theta(P)$, where $P \neq P'$ or $P \neq P'_e$ for all $e \in \mathcal{E}$. Thus, external validity is about a parameter of a distribution which has not been seen in the data, for example a regression parameter in new data which has a different data generating distribution than the one generating the observed (training) data.

There is a fast growing literature on the theme of external validity, including distributional robustness [62], domain adaptation and transfer learning [49], and transportability [52]. We will present a brief summarizing view of them in Section 2. On the other hand, there is very little work on distributionally robust confidence statements. We review here some of the work from the latter topic [25, 36] and provide an overarching perspective of the state-of-the-art.

1.2 Internal Sampling Stability

Stability is an important concept to obtain higher degree of replicability. The easiest version is internal sampling stability and is often implemented via subsampling or bootstrapping the observed data [45, 43, 69, 9, 32, 70]. Inspecting and improving sampling stability is particularly useful for complex models and corresponding procedures: we mention here as some examples uncertainty assessment in high-dimensional models [71, 63, 18, 64, 48].

Other forms of stability can be even used for external validity, and this is discussed in Section 2.

1.2.1 *Post-selection inference*. Since uncertainty quantification is difficult and often fragile in complex models, post-selection inference procedures became rather popular [7, 40, 38]. They are reliable and provide good internal replicability for the discovery of a particular selected hypothesis. However, if some data-driven model selection with, for example, the Lasso is performed [40, 38], the entire procedure becomes often unstable and leads to a very bad degree of replicability. The reason for it is as follows: the Lasso would typically pick a different set of selected variables on another dataset (or a subsampled one) and hence, the inference after selection will also focus on a different parameter and its hypothesis: it is as much not replicable as the difference among the selected models from the Lasso. This point is often not made very explicit and things are expected to worsen when it comes to external validity.

2. EXTERNAL VALIDITY OF POINT ESTIMATION: DISTRIBUTIONAL ROBUSTNESS, DOMAIN ADAPTATION AND CAUSALITY

External validity and corresponding (point) estimation strategies have been developed from different perspectives, all of them aiming to address the issue when the external (new) data has a different distribution than the original internal (training) data. In the following, we give a high-level description of the topic.

2.1 Robust Methods

Protection against small-to-medium unknown perturbations can be achieved with robust methods. For large perturbations, these procedures become conservative. There is an important distinction between "classical" and distributional robustness.

In the former "classical" case, the goal is to estimate a parameter of the unperturbed reference (or target) distribution when the (training) data is contaminated and often interpreted as realizations of a mixture of the reference and contamination distribution. There is only internal data, and the contaminations are among the observed samples. The methodology proceeds by data-driven down-weighting of outliers (contaminated data points), giving them less weight than $1/n$ with n denoting the total (internal) sample size. See, for example, [30, 26].

In distributional robustness the aim is to predict well under adversarial perturbations in the external test data and the parameter of interest is with respect to a perturbed adversarial distribution. Here, the training dataset is internal and clean, while the contaminations or perturbations are not among the observed training data. This scenario is often relevant in modern machine learning. In this conceptual description, distributional robustness arises from up-weighting certain data points (giving them more weight than $1/n$) in order to achieve good performance on test data. For example, in regression one would aim to estimate a function $f(\cdot)$ which optimizes

$$\operatorname*{argmin}_{f(\cdot)\in\mathcal{F}} \sup_{P;d(P,P')\leq\rho} \mathbb{E}_P\big[\big(Y - f(X)\big)^2\big],$$

where $\mathcal{F}$ is a suitable class of functions, P' is the internal training distribution, P is the external distribution, $d(\cdot,\cdot)$ a metric between probability distributions, ρ a certain positive number, Y a univariate response, and X the vector of covariates (and $(Y, X) \sim P$) [5, 8].

2.2 Domain Adaptation and Re-Weighting

Domain adaptation methods can cope with large distributional shift and perturbations from the training data

distribution P' to a target (test) distribution P which generates new data. This can be achieved by re-weighting which takes the distributional change into account [41, 39].

For example, one might be interested in

$$\begin{aligned}&\operatorname*{argmin}_{f(\cdot)\in\mathcal{F}} \mathbb{E}_P\big[(Y-f(X))^2\big]\\&\quad=\operatorname*{argmin}_{f(\cdot)\in\mathcal{F}} \mathbb{E}_{P'}\big[(Y-f(X))^2 w(X,Y)\big],\end{aligned}$$

Here, $w(X,Y)=\frac{dP}{dP'}(X,Y)$ is the Radon–Nikodym derivative. The results above motivate weighted empirical risk minimization:

$$\hat{f}=\operatorname*{argmin}_{f(\cdot)\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\hat{w}(X_i,Y_i)\big(Y_i-f(X_i)\big)^2,$$

for some estimate $\hat{w}(\bullet)$ of $w(\bullet)$. There is often an assumption that restricts the shift in a particular way. For example, if $w(\bullet)$ only depends on X, we are in the popular setting of *covariate shift* [56], cf. There is an underlying assumption about some overlap between the training and target (or test) distribution which then enables adapting to a different domain which may be far away in terms of a probabilistic distance.

In a different line of work, one tries to learn invariant representations of the features [49, 3]. This is based on the idea that if a representation of the data is invariant between the training distribution P' and target distribution P, then feeding these representations into a prediction algorithm might exhibit improved generalizability compared to feeding the untransformed data into a prediction algorithm.

The empirical success of such domain adaptation methods and algorithms is primarily documented in the field of computer vision [24, 23, 53]: indeed, it is remarkable that even though $d(P,P')$ is large for a metric $d(\cdot,\cdot)$, it is possible to accurately learn from P' some aspects about P.

2.3 Invariance and Causality

Another framework for achieving external validity is to learn causal representations which are able to generalize well outside the internal data. This includes invariance of feature representations [27], or of residuals and learning some causal structures [54, 57, 42, 28, 10, 59]. With such *structural* approaches and models, no overlap assumption between P' and P is required but they typically rely on multiple sources or environments for learning the invariances. Unlike distributional robustness but in the same vein as domain adaptation, these methods allow for large distributional shifts and interventions between the internal and external data-generating distributions.

2.4 The Role of Multiple Sources or Environments

Internal training data which is grouped according to different sources or groups under different environments, denoted above by $e\in\mathcal{E}$ with $\mathcal{E}$ being the space of observed environments, provides useful information for external generalization. The main reason is that internal sources of heterogeneity can be used to model distributional shifts, invariances or infer causal structure. Such multisource/environment information has been exploited from a theory and practical point of view: for optimizing worst environment risk [44, 11, 61], for domain adaptation [23, 3, 27, 12], and for causal regularization aiming to obtain invariant residuals [54, 28, 59, 2].

We also note that instrumental variables regression is related to multienvironment problems [1, 33, 34]. If the instruments are discrete, they can be thought as encoding different environments, but IV regression also covers continuous forms of heterogeneity. A main and strong assumption is the so-called validity of such instruments: under such strong conditions, the invariant structure is equal to the causal structure, and a causal model is also externally valid under arbitrarily strong perturbations of the covariates.

3. DISTRIBUTIONALLY ROBUST UNCERTAINTY QUANTIFICATION

Considerations of external validity not only affect (point) estimation strategies, but should also affect how we report uncertainty. If data from multiple environments are available, one can conduct some type of meta-analysis. For example, partial conjunction tests [29, 6, 66] allow to conduct valid inference in situations where a few of the datasets are perturbed. Such analysis provides internal validity among the different environments only. It cannot go beyond the internal multienvironment data. If only one dataset is available, there exist much fewer methods that account for distributional uncertainty. Existing methods either:

- employ worst-case bounds between the distribution of P' and P; or
- assume that the probabilities of events change randomly between P' and P.

As an example of the first approach, assume that we know that $\mathrm{D}_{\mathrm{KL}}(P\|P')\le\delta$, where $\mathrm{D}_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence, and that $X_i\overset{\text{i.i.d.}}{\sim}P'$, with $P'=\mathcal{N}(\mu,1)$. We aim to construct a confidence interval $I=I(X_1,\ldots,X_n,\delta)$ which is uniformly valid over the Kullback–Leibler ball, that is,

$$\inf_{P:\mathrm{D}_{\mathrm{KL}}(P\|P')\le\delta}\mathbb{P}\big[\mathbb{E}_P[X]\in I\big]=1-\alpha. \tag{1}$$

Using some algebra, we get that

$$I=\frac{1}{n}\sum_{i=1}^{n}X_i\pm\left(\frac{z_{1-\alpha/2}}{\sqrt{n}}+\sqrt{2\delta}\right)$$

satisfies equation (1), where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard Gaussian random variable. This robust confidence interval is similar in spirit to robust versions of the probability ratio test [31] in the sense that one needs to prespecify the strength of perturbations δ. Note that this confidence interval has a component that does not converge to zero as $n \to \infty$.

Another approach is given by sensitivity analysis in causal inference, which investigates the stability of a statistical finding under (potential) unobserved confounding [14, 58]. Today this is a field of active research [19, 72, 13, 68, 21, 37]. Such sensitivity analysis is of the following nature. First, the tools are usually specific to the estimation strategy, and thus have to be used on a case-by-case basis. Second, when considering worst-case distributional perturbations, very small shifts can already change results substantially. Since most sensitivity analyses are measuring worst-case stability, reported "instabilities" often occur due to the overly conservative worst-case analysis.

In the following, we describe how these issues can potentially be addressed by a different type of sensitivity or stability analysis which is takes a "directional worst-case" view point.

3.1 Towards General-Purpose Tools for Stability Analysis

Often, practitioners are interested in *sign stability* of a one-dimensional statistical parameter. To be more specific, one might want to infer whether $\operatorname{sign}(\theta(P)) = \operatorname{sign}(\theta(P')) \approx \operatorname{sign}(\hat{\theta})$ for a reasonable set of perturbed distributions $P \in \mathcal{P}$. The motivation behind sign stability is that the parameter might correspond to whether or not a medication has a positive effect. We can quantify the sign stability by estimating

$$s = \exp\Big(-\inf_{P} \mathrm{D}_{\mathrm{KL}}(P \| P') \quad \text{such that } \operatorname{sign}\big(\theta(P')\big) \neq \operatorname{sign}\big(\theta(P)\big)\Big). \tag{2}$$

In example above, sign stability captures whether a medication that has a beneficial effect on the observed population might be harmful under distribution shift.

In analogy to the p-value, if s is close to zero, the sign of $\theta(\cdot)$ is very stable under distributional changes. On the other hand, if s is close to one, the sign is highly unstable under distributional changes.

Let's consider an example. Assume we are interested in estimating the mean $\theta(P) = \mathbb{E}_P[X]$. Donsker and Varadhan [20] showed that if the moment generating function of X is finite, then

$$s = \inf_{\lambda} \mathbb{E}_{P'}\big[e^{\lambda X}\big],$$

where P' is from the data generating distribution and hence can be inferred from observed data.

In fact, for i.i.d. observations $X_i \sim P'$ we can use the following plug-in estimator of the distributional stability measure s:

$$\hat{s} = \inf_{\lambda} \frac{1}{n} \sum_{i=1}^{n} e^{\lambda X_i}.$$

Consistency guarantees for this estimator of stability are given in [25]. For other estimands than the expected value, such as parameters in generalized linear models or estimands defined via moment equations, estimating s is more involved since Donsker–Varadhan does not apply directly.

In practice, one can use simple linear approximations to estimate s. If the estimand $\theta(\cdot)$ is differentiable as a functional on the distribution space, by definition

$$\theta(P) - \theta\big(P'\big) = \mathbb{E}_P\big[\phi_{P'}(D)\big] + o\big(d_K\big(P, P'\big)\big) \tag{3}$$

for some metric $d_K(\cdot,\cdot)$ such as the Kolmogorov metric and a function $\phi_{P'}(D)$ with $\mathbb{E}_{P'}[\phi_{P'}(D)] = 0$. Then Donsker–Varadhan suggests using the estimator

$$\hat{s} = \inf_{\lambda} \frac{1}{n} \sum_{i=1}^{n} e^{\lambda(\hat{\theta} + \hat{\phi}(D_i))}, \tag{4}$$

where $\hat{\phi}$ is an estimate of the influence function $\phi_{P'}(\cdot)$ and $\hat{\theta}$ is an estimate of $\theta(P')$. For example, if $(X, Y) \in \mathbb{R}^{p+1}$ and $\theta(P)$ is the kth component of the regression vector, that is $\theta(P) = \beta_k(P)$, where

$$\beta(P) = \arg\min_{\beta} \mathbb{E}_P\big[(Y - X\beta)^2\big],$$

then one can estimate $\phi_{P'}(D_i)$ via

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

where $\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\sum_{j=1}^{n}(Y_j - X_j\beta)^2$. As before, the data $(X_i, Y_i)_{i=1,\ldots,n}$ is drawn i.i.d. from P'. This strategy allows to estimate s for common estimands such as parameters of generalized linear models or parameters defined via moment equations. Algorithms with consistency guarantees are given in [25].

These s-values can then be compared to benchmarks. As an example, in [17], the authors compute benchmarks for different national surveys (e.g., ANES and CES). To be more concrete, they estimate benchmarks $\hat{b}_{P,P'} = e^{-\hat{\mathrm{D}}_{\mathrm{KL}}(P\|P')}$ for P and P' corresponding to different national surveys. They report that the empirical mean of the estimated $\hat{b}$, averaged over multiple pairs of surveys is 0.86. This puts distributional stability measures into context. For example, if the s-value of a statistical result in a similar application is larger than 0.86, then this is an indication that the result might not generalize; in the sense that a distribution shift of the size that is observed between different national surveys is large enough to change

the sign of the result. Note that for simplicity we have ignored statistical uncertainty quantification. Details on how to compute confidence intervals for s-values can be found in [25].

Of course, distributional stability measures are context dependent. In [17], it is argued that one has to choose a smaller threshold when generalizing from national surveys to samples from Amazon Mechanical Turk (MTurk).

Under arbitrary distribution shifts, one will usually be able to change the sign of statistical parameters under very small shifts. Thus, to make the tools more useful in practice, it is important to restrict the class of considered distribution shifts.

3.1.1 *Beyond omni-directional shifts.* In principle, any statistical finding breaks down under arbitrary distributional shifts. Thus, a practitioner might be interested in learning under what circumstances (that is, under which type of distribution shifts) a result breaks. As an example, maybe a correlation between two variables is nearly invariant across populations with different socioeconomic status, but highly variable across different age groups.

In the following, we will discuss how targeted sensitivity or stability analysis can be formalized. Related to the definition of s in (2), for a random variable E which is observed in the data,

$$s_E = \exp\Big(-\inf_{P:P[\cdot|E]=P'[\cdot|E]} \mathrm{D}_{\mathrm{KL}}(P\|P') \quad \text{such that } \operatorname{sign}(\theta(P)) \neq \operatorname{sign}(\theta(P'))\Big). \tag{5}$$

Formally, s_E is a deterministic value that lies in $[0, 1]$; the constraint $P[\cdot|E] = P'[\cdot|E]$ is meant to hold almost surely w.r.t. E. In words, we investigate how much a shift in the marginal distribution of E can affect the parameter, while keeping the conditional distribution $P[\cdot|E]$ invariant.

Estimation of this stability parameter relies on a variation of Donsker–Varadhan's lemma [20]. Using a similar approximation as for equation (4), s_E can be estimated via

$$\hat{s}_E = \inf_{\lambda} \frac{1}{n} \sum_{i=1}^{n} e^{\lambda(\hat{\theta}+\hat{Q}(E_i))}, \tag{6}$$

where $\hat{Q}(e)$ is an estimate of $Q(e) = \mathbb{E}_{P'}[\phi_{P'}(D)|E = e]$. A formal justification of this estimator is given in [25].

Let's return to the case of linear regression with $X = E$, with X being potentially multidimensional. This means we are considering a distribution shift in the covariates while keeping $Y|X$ constant. For $\theta(P) = \beta_k(P)$, a short calculation shows that

$$Q(X_i) = \mathbb{E}_{P'}\big[X^\intercal X\big]^{-1}_{k,\bullet} X_i^\intercal \big(\mathbb{E}_{P'}[Y|X = X_i] - X_i\beta(P')\big).$$

Based on this observation one can form a plug-in estimator of $Q(x)$ and use equation (6) to estimate the directional stability coefficient s_E. Note that if the model is well-specified, $\mathbb{E}_{P'}[Y|X = x] - x\beta(P') = 0$ and thus the stability value s_X is zero as long as $\theta(P') \neq 0$. Thus, for the specific choice of $E = X$, the directional stability coefficient s_E captures whether the model is well-specified.

3.1.2 *Real-world example.* We demonstrate the usage of the stability measure s_E in (5) on the life-cycle savings data [4]. The dataset contains measurements of the ratio between personal savings divided by disposable income (savings ratio—sr), the percentage of population under 15 (pop15), and the percentage of population over 75 (pop75), the disposable income (dpi) and the growth rate of the disposable income (ddpi). Under Modigliani's life-cycle savings hypothesis [46], the savings ratio is explained by these four covariates. We want to investigate the robustness of the linear model under various distributional shifts. A package implementing the estimation of the directional stability measure s_E from (5) as in (6) can be obtained from github.com/rothenhaeusler/stability. The following R code fits a linear model and computes the stability of the linear regression coefficient corresponding to pop15, that is, the stability of $\theta(P) = \beta_{\mathrm{pop15}}(P)$ for different choices of E. We consider distribution shift both in single components of X, but also in the outcome Y:

```
> fit <- lm(sr ~ pop15 + pop75 +
    dpi + ddpi, data = LifeCycleSavings)
> stability(fit,param="pop15")
Stability values

    s_sr s_pop15 s_pop75   s_dpi  s_ddpi
   0.863   0.368   0.781   0.687   0.837
```

Here, the names of the different columns correspond to the different choices of E.

These values can be used to compare the relative stability of parameter values under different choices of E, for example, using the smallest $\hat{s}_E$ as the baseline stability value. In addition, these robustness or stability measures can be compared to reference values computed across real-world datasets, as discussed in the previous section. For the population of the United States, based on the census of 2016, we get an estimate of the benchmark value $\hat{b} = 0.17$ for the change in the distribution of pop15. As $\hat{b} = 0.17$ is smaller than $\hat{s}_{\mathrm{pop15}} = 0.368$ (see the output above), we have to be concerned that due to a large shift in the distribution of pop15, the sign of the regression coefficient might change if we were to collect new data from the US. On the other hand, for Mexico the estimate is $\hat{b} = 0.54$ which is larger than $\hat{s}_{\mathrm{pop15}} = 0.368$. This suggests that we do not have to be concerned that a shift in pop15 changes the regression coefficient if we were to collect new data from Mexico. Such calculations allow us to gauge the extent to which a result will generalize.

In addition to s-values, the `R`-function `stability` provides a visualization of the stability of parameters under distributional shifts. More concretely, for different

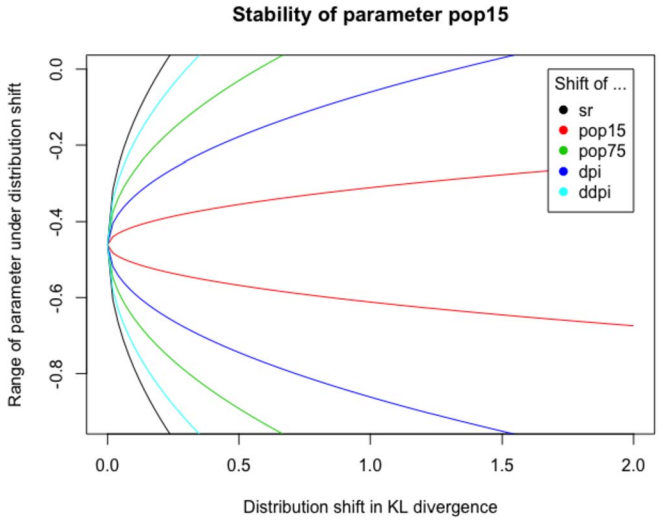

FIG. 1. *Stability of the parameter pop*15 *under various distributional shifts as reported by* `stability()`. *Each coloured pair of lines corresponds to the different individual components of the parameter vector and displays the estimated versions of* (7) *and* (8), *respectively.*

choices of E and an upper bound on the distribution shift x we compute upper and lower bounds for parameter values as follows:

$$y_{\text{upper-bound}} = \sup \theta(P) \quad \text{such that} \tag{7}$$
$$P'[\cdot|E] = P[\cdot|E] \text{ and } \mathrm{D_{KL}}(P\|P') \le x,$$

$$y_{\text{lower-bound}} = \inf \theta(P) \quad \text{such that} \tag{8}$$
$$P'[\cdot|E] = P[\cdot|E] \text{ and } \mathrm{D_{KL}}(P\|P') \le x.$$

In Figure 1, we visualize estimated versions (based on a linear approximation and plug-in) of these upper and lower bounds across x for different choices of the variable E.

This plot allows to derive bounds on parameters based on background knowledge. For example, if the data scientist expects that the distribution of dpi is expected to shift by at most 0.5 in Kullback–Leibler divergence between settings, then one would obtain an (estimated) upper bound of -0.2 for the parameter.

4. CONFIDENCE INTERVALS THAT ACCOUNT FOR BOTH SAMPLING AND DISTRIBUTIONAL UNCERTAINTY

The diagnostic tools discussed in the previous section can be conservative since they still rely on possibly directional worst-case bounds. In practice, we need not be that pessimistic and thus we consider here *average perturbation* effects. Furthermore, we will describe procedures which do not rely on the user's interpretation of stability values but *estimate the amount of perturbations* from data.

For our further developments, we model the perturbation process as random. Intuitively speaking, if the data $D_1, \dots, D_n$ is drawn i.i.d. from the sampling distribution P' which randomly deviates from the target distribution P, we can decompose uncertainty into a sampling component and a distributional component:

$$\theta(P'_n) - \theta(P) = \underbrace{\theta(P'_n) - \theta(P')}_{\text{sampling uncertainty}} + \underbrace{\theta(P') - \theta(P)}_{\text{distributional uncertainty}}.$$

Here, P'_n denotes the empirical measure of $D_1, \dots, D_n$.

Compared to classical statistical inference, we aim to construct confidence intervals for the target $\theta(P)$, instead of the parameter of the sampling distribution $\theta(P')$. However, without any restrictions on the perturbation process, it is impossible to quantify the magnitude of $\theta(P') - \theta(P)$. This raises the question about the distributional perturbation model.

4.1 A Model for Distributional Perturbations

An easy and perhaps natural distributional perturbation model is as follows. For any event, the perturbed distribution will assign slightly different probabilities compared to the target distribution P. To simplify the discussion in the following, we will assume that the sample space $\mathcal{D}$ is discrete with uniform weights on the singletons, that means for all $d \in \mathcal{D}$

$$P[D = d] = \frac{1}{|\mathcal{D}|}.$$

A perturbed distribution can now be formed by drawing exchangeable random variables $\xi_i \ge 0$ $(i = 1, \dots, |\mathcal{D}|)$ with $\sum_i \xi_i = 1$, and setting

$$P^{\xi}[D = d_i] = \xi_i.$$

Although the discussion in this section has focused on the discrete case, a similar argument works in the continuous case, see [36], Section 2.

4.1.1 *Mean and variance of sample means under the perturbation model.* For simplicity, we consider first the sample mean. Conditionally on ξ, the data is drawn i.i.d. from P^{ξ} and we denote the marginal distribution when averaging over the random variables ξ_i $(i = 1, \dots, |\mathcal{D}|)$ by

$$\mathrm{d}P_{\text{marginal}}(d) = \int \mathrm{d}P^{\xi_1,\dots,\xi_{|\mathcal{D}|}}(d)\,\mathrm{d}P(\xi_1, \dots, \xi_{|\mathcal{D}|}).$$

For any function $f(\cdot)$, the marginal expectation of the sample mean, averaging over both sampling and distributional uncertainty, is

$$\mathbb{E}_{\text{marginal}}\left[\frac{1}{n}\sum_{i=1}^{n} f(D_i)\right] = \mathbb{E}_P\big[f(D)\big] \tag{9}$$

and the marginal variance of the sample mean is

$$\mathrm{Var}_{\text{marginal}}\left(\frac{1}{n}\sum_{i=1}^{n} f(D_i)\right) = \frac{\delta^2}{n}\,\mathrm{Var}_P\big(f(D)\big), \tag{10}$$

where the scaling factor δ^2 satisfies

$$\frac{\delta^2}{n} = \underbrace{\frac{1}{n}}_{\text{due to sampling}} + \underbrace{\operatorname{Var}(\xi_1)\frac{n-1}{n}\frac{|\mathcal{D}|^2}{|\mathcal{D}|-1}}_{\text{due to distributional perturbation}}. \tag{11}$$

Note that we write for simplicity the subindex "marginal" instead of P_{marginal}. Under regularity assumptions [36], Section 2, one can show

$$\frac{1}{n}\sum_{i=1}^{n} f(D_i) - \mathbb{E}_P\big[f(D)\big] \approx \mathcal{N}\Big(0, \frac{\delta^2}{n}\operatorname{Var}_P\big(f(D)\big)\Big). \tag{12}$$

Let us give some intuition on how to interpret different values of δ. By equation (11), $\delta^2 \geq 1$. If $\delta = 1$, then $\operatorname{Var}(\xi_i) = 0$ and thus there is no distributional perturbation, that is the data is drawn i.i.d. from $P^\xi = P$. As δ increases, the $f(D_i)$ become increasingly correlated marginally. The variance $\operatorname{Var}(\xi_1)$ is maximized for $\xi_i \in \{0, 1\}$ for all i. In this case, $\operatorname{Var}(\xi_1) = \frac{1}{|\mathcal{D}|} - \frac{1}{|\mathcal{D}|^2}$. Using equation (11) in this most extreme case, we get $\delta^2 = n$. Overall, we have the bound

$$1 \leq \delta^2 \leq n.$$

4.1.2 *A numerical example*. The R package calinf (github.com/rothenhaeusler/calinf) contains functions to generate data from perturbed distributions. Sampling from the distributional perturbation model is slightly more involved than drawing i.i.d. random variables, since we have to choose the strength of the perturbation. We set the state of the distributional perturbation by setting a distributional seed via `distributional_seed`. This step is not optional. On a high level, the distributional seed indicates to the random number generator which observations are drawn from the same perturbed distribution, allowing the random number generator to introduce spurious associations between the variables. Once the distributional seed is set, one can use this to generate perturbed data as follows:

```
d_seed <- distributional_seed(n=1000,
                              delta=5)
x <- drnorm(d_seed)
y <- drnorm(d_seed)
```

The displayed code generates 1000 observations from P^ξ, where P^ξ is a perturbed two-dimensional standard Gaussian distribution. The perturbed data is generated such that equation (10) holds approximately for any square-integrable $f(D)$. For continuous random variables, one cannot directly use the strategy described in Section 4.1. Details on how to sample from perturbed continuous distributions can be found in [36], Section 2. The function `drnorm` generates i.i.d. data from a perturbed Gaussian.

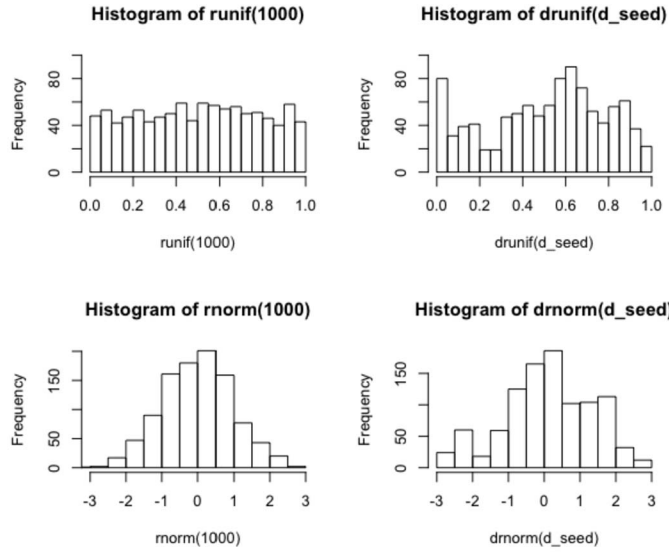

FIG. 2. *Random number generation from the distributional perturbation model. On the upper left, the observations are drawn i.i.d. from the uniform distribution. On the upper right, the observations are drawn from a perturbed uniform distribution. On the bottom left, the observations are drawn from the standard Gaussian distribution. On the lower right, the sample size is* $n = 1000$. *For the distributional perturbations we use* $\delta = 5$.

Analogously, we provide functions to sample from a perturbed binomial distribution (`drbinom`), perturbed uniform distribution (`drunif`), etc. Here, the "dr" in drunif stands for "distributional randomness". An example is shown in Figure 2.

4.2 Estimation Under the Distributional Perturbation Model

In this section, we describe how to do estimation and inference in the distributional perturbation model from Section 4.1.

In a nutshell, as the expectation of sample means is unchanged, estimation under the distributional perturbation model can proceed "as usual": irrespective of the value of δ, one can construct point estimators such as moment-based or maximum-likelihood estimators as if the data were drawn i.i.d. from the target distribution P.

As an example, let us focus on the OLS parameter $\theta(P) = \arg\min \mathbb{E}_P[(Y - X\theta)^2]$. If $\theta(P)$ is unique, it can be rewritten as

$$\theta(P) = \mathbb{E}_P\big[X^\intercal X\big]^{-1}\mathbb{E}_P\big[X^\intercal Y\big].$$

Assume that $D_i = (X_i, Y_i)$, $i = 1, \ldots, n$ are drawn i.i.d. from P^ξ. Due to equation (9), marginally across the sampling and distributional perturbation, we have

$$\mathbb{E}_{\text{marginal}}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^\intercal X_i\right] = \mathbb{E}_P\big[X^\intercal X\big] \quad \text{and}$$

$$\mathbb{E}_{\text{marginal}}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^\intercal Y_i\right] = \mathbb{E}_P\big[X^\intercal Y\big].$$

This motivates the estimator

$$\hat{\theta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^\intercal X_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^\intercal Y_i\right),$$

which is the usual OLS estimator that we would use if the data $(X_i, Y_i)_i$ were drawn i.i.d. from P. Similar ideas can be applied to maximum likelihood estimation and the method of moments to show that estimation can proceed "as usual" [36].

On the other hand, as we will discuss below, the variance of the resulting estimator depends on the (usually unknown) δ. Under the usual and additional minor assumptions, such estimators are asymptotically linear and Gaussian under the distributional perturbation model [36]. For the example with OLS regression, a Taylor expansion shows that

$$\hat{\theta} - \theta(P) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}_P\big[X^\intercal X\big]^{-1} X_i^\intercal \big(Y_i - X_i\theta(P)\big)}_{\phi_P(D_i)} + o_{P_{\text{marginal}}}\left(\frac{\delta}{\sqrt{n}}\right). \tag{13}$$

Thus, up to lower order terms, the difference between the estimator and the target parameter (computed on the unperturbed distribution) is a mean of a function of the data. This is similar to classical expansions in terms of the influence function [65]. The main difference is that the expansion is done in a non-i.i.d. setup that accounts for both sampling uncertainty and distributional uncertainty. Since the estimator is asymptotically linear, we can apply equation (12) to equation (13). Thus, under regularity assumptions [36], one obtains

$$\hat{\theta} - \theta(P) \approx \mathcal{N}\left(0, \frac{\delta^2}{n}\operatorname{Var}_P\big(\phi_P(D)\big)\right),$$

where the distributional approximation is meant to hold w.r.t. P_{marginal} on the left-hand side. In the following, we will see that standard approaches will fail at estimating the correct variance of $\hat{\theta}$.

4.2.1 *Classical statistical inference may drastically underestimate uncertainty.* In the model above, one might be tempted to estimate the variance of statistical quantities as usual. Let's consider the example of the sample mean $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$. In the following, we will see that it is straightforward to estimate $\operatorname{Var}_P(D)$, but that estimation of $\operatorname{Var}_{\text{marginal}}(\overline{D})$ is more challenging.

Estimation of $\operatorname{Var}_P(D)$. If the data were drawn i.i.d. from the target distribution P, one would use the variance estimator $\hat{\sigma}^2$, where $\hat{\sigma}^2$ is the empirical variance

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

Let us now investigate the variance estimator $\hat{\sigma}^2$. We will see that $\mathbb{E}_{\text{marginal}}[\hat{\sigma}^2]$ is close to $\operatorname{Var}_P(D)$. By equation (9), the marginal expectation of $\frac{1}{n}\sum_i D_i^2$ is $\mathbb{E}_P[D^2]$. Similarly, the marginal expectation of $\overline{D}$ is $\mathbb{E}_P[D]$. Using equation (10), the variance of $\overline{D}$ is $\delta^2 \operatorname{Var}_P(D)/n$. Thus, the empirical variance estimate will have expected value

$$\begin{aligned}\mathbb{E}_{\text{marginal}}\big[\hat{\sigma}^2\big] &= \mathbb{E}_P\big[D^2\big] - \left(\mathbb{E}_P[D]^2 + \frac{\delta^2}{n}\operatorname{Var}_P(D)\right)\\ &= \left(1 - \frac{\delta^2}{n}\right)\operatorname{Var}_P(D).\end{aligned}$$

Thus, if δ^2 is small or n is large, then $\mathbb{E}_{\text{marginal}}[\hat{\sigma}^2]$ is close to $\operatorname{Var}_P(D)$; that is the difference is negligible.

This effect can also be easily observed empirically, here illustrated with some commands in `R`. In the following, we draw $n = 1000$ observations in a distributional perturbation model with $\delta = 2$. The estimated variance is relatively close to the variance $\operatorname{Var}_P(D) = 1$, where $P = \mathcal{N}(0, 1)$.

```
> d_seed <- distributional_seed(
                          n     =   1000,
                          delta =   2
                          )
> D <- drnorm(d_seed)
> var(D)
[1] 0.9414752
```

This is good news. However, there are also some bad news. To construct confidence intervals for $\mathbb{E}_P[D]$, we need an estimator of the variance of $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$.

Naive estimation of $\operatorname{Var}_{\text{marginal}}(\overline{D})$. As we will see in the following, the naive estimator $\hat{\sigma}^2_{\text{naive}} = \frac{1}{n}\hat{\sigma}^2$ systematically underestimates the variance of $\overline{D}$, potentially drastically so. Intuitively, this is the case because the data points D_i are positively correlated under P_{marginal}, with unknown correlation. Let us compute the expectation

$$\begin{aligned}\mathbb{E}_{\text{marginal}}\big[\hat{\sigma}^2_{\text{naive}}\big] &= \frac{1}{n}\left(1 - \frac{\delta^2}{n}\right)\operatorname{Var}_P(D)\\ &< \frac{\delta^2}{n}\operatorname{Var}_P(D) \quad \text{equation (11)}\\ &= \operatorname{Var}_{\text{marginal}}(\overline{D}) \quad \text{equation (10)}.\end{aligned}$$

Thus, $\hat{\sigma}^2_{\text{naive}}$ systematically underestimates $\operatorname{Var}_{\text{marginal}}(\overline{D})$. As discussed before, in the most extreme case $\delta^2 = n$ which would make the left-hand side equal to zero. We can also see this effect empirically, illustrated in `R` below.

The naive estimator, computed on the previous example, is

```
> var(D)/n
[1] 0.0009414752
```

On the other hand, the actual variance of $\overline{D}$, marginally across both the distributional perturbation and the sampling process is

```
> simulate_mean <- function(){
> d_seed <- distributional_seed(
                          n     =   1000,
                          delta =   2
                          )
> D <- drnorm(d_seed)
> return(mean(D))
> }
> var(replicate(n=10000,simulate_mean()))
[1] 0.003949516
```

Thus, the naive estimator $\hat{\sigma}^2_{\text{naive}}$ underestimates the variance roughly by a factor of 4 which is to be expected since $\delta^2 = 4$.

Summarizing the discussion in this section, estimation of $\sigma^2 = \text{Var}_P(D)$ can be done as usual, while estimation of $\text{Var}_{\text{marginal}}(\overline{D})$ is more difficult. To be more specific, one can use the empirical variance of $(D_i)_{i=1,\dots,n}$ to estimate $\sigma^2 = \text{Var}_P(D)$. In the more general case of asymptotically linear estimators, one can estimate the variance $\text{Var}_P(\phi_P(D))$ by computing the empirical variance of $(\hat{\phi}(D_i))_{i=1,\dots,n}$, where $\hat{\phi}$ is a plug-in estimate of the influence function [36]. Let us now turn to estimation of $\text{Var}_{\text{marginal}}(\overline{D})$. Since

$$\text{Var}_{\text{marginal}}(\overline{D}) = \frac{\delta^2}{n}\text{Var}_P(D),$$

the main challenge is to estimate δ. In the following two sections, we will discuss two approaches to estimate δ.

4.2.2 *Calibration of uncertainty using triangulation.* In this section, we discuss how a commonly recommended research strategy, called "method triangulation" can be used to estimate distributional uncertainty.

If several estimators of an effect are available, one can use variation of the estimators as a measure of robustness. In the statistics literature, this type of stability analysis has been advocated by Yu and Kumbier [70] as part of the predictability, computability, and stability (PCS) framework. More generally speaking, investigating stability across methods is often referred to as method triangulation [16, 50, 47]. Triangulation is conceptually different from replicability across settings. For example, if the same study is conducted multiple times at different locations, these studies may share similar biases and thus may be consistently incorrect. On the other hand, if different methodologies yield similar conclusions, then the result is less likely to be an artifact. These intuitive arguments can be made precise in the distributional uncertainty framework.

Assume we have access to several estimators $\hat{\theta}_1, \dots, \hat{\theta}_K$ for the same parameter of interest $\theta(P)$. Examples from causal inference include settings where we have:

- multiple instruments,
- multiple adjustment sets, or
- treatment effect homogeneity.

For example, in presence of treatment effect homogeneity, we can estimate average treatment effects on various subpopulations. If there were no distributional uncertainty across the subpopulations, these estimators should agree, at least asymptotically. On the other hand, if there is a lot of distributional uncertainty, these estimators will be very far apart from each other. Thus, we can use the observed variation between estimators as an indication of how much distributional uncertainty is present for the problem at hand. In the following, we will make this more precise.

We assume that the estimators are asymptotically linear, that is,

$$\hat{\theta}_k - \theta_k(P) = \frac{1}{n}\sum_{i=1}^{n}\phi_k(D_i) + o_{P_{\text{marginal}}}\left(\frac{\delta}{\sqrt{n}}\right),$$

for some mean-zero functions ϕ_k, that is $\mathbb{E}_P[\phi_k(D)] = 0$. For the example of ordinary least squares estimation, see also equation (13). This is also justified for maximum likelihood estimators and empirical risk minimization in low-dimensional settings, see [36]. For simplicity, in the following we will assume that $\theta_k(P) = \theta_\ell(P)$ for all k, ℓ. This corresponds to the assumption that if no uncertainty were present (infinite data from the target distribution), all estimators would return the same target quantity. One might have reasons to doubt this assumption. If this assumption holds, the inferential procedure described below will have exact coverage asymptotically. If it is violated, one will generally have overcoverage [36].

Now let us proceed with the estimation of δ. The variation between the different estimation strategies is a measure of the trustworthiness of the result. Considering the squared difference of the estimators yields

$$\begin{aligned} &n(\hat{\theta}_k - \hat{\theta}_\ell)^2 \\ &\quad = n\left(\frac{1}{n}\sum_{i=1}^{n}\phi_k(D_i) - \phi_\ell(D_i)\right)^2 + o_{P_{\text{marginal}}}(\delta^2) \\ &\quad \approx \delta^2\,\text{Var}_P\big(\phi_k(D) - \phi_\ell(D)\big)\chi^2_1. \end{aligned} \tag{14}$$

Here, we used equation (12). Thus, we can form an estimate of δ by setting

$$\hat{\delta}^2 = \frac{1}{K(K-1)}\sum_{k\neq\ell}\frac{n(\hat{\theta}_k - \hat{\theta}_\ell)^2}{\widehat{\text{Var}}_P(\phi_k(D) - \phi_\ell(D))}. \tag{15}$$

One can then use this estimate in conjunction with equation (12) to form 95%-confidence intervals:

$$\hat{\theta} \pm 1.96 \frac{\hat{\delta}\hat{\sigma}}{\sqrt{n}}. \tag{16}$$

Here, $\hat{\sigma}^2$ is the usual variance estimate one would use if the data were drawn i.i.d. from the target distribution. More specifically, one can estimate the influence function ϕ of $\hat{\theta}$ and use plug-in estimate of the variance

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{\phi}(D_i) - \frac{1}{n} \sum_{j=1}^{n} \hat{\phi}(D_i) \right)^2. \tag{17}$$

Under regularity assumptions and for large K, this interval is valid in an asymptotic sense [36]:

$$\mathbb{P}_{\text{marginal}} \left[|\theta(P) - \hat{\theta}| \leq z_{1-\alpha/2} \frac{\hat{\delta}\hat{\sigma}}{\sqrt{n}} \right] \to 1 - \alpha, \tag{18}$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$-quantile of a standard Gaussian random variable. If K is small, then equation (14) suggests replacing $z_{1-\alpha/2}$ with quantiles of a t-distribution with appropriate degrees of freedom [36]. The main takeaway here is that we give coverage guarantees for the unperturbed parameter $\theta(P)$, as opposed to the perturbed parameter $\theta(P^{\xi})$. Furthermore, these guarantees hold marginally, that means across multiple draws of both the distributional and sampling uncertainty.

An important aspect that we have glossed over until now is that for this procedure to work the estimators $\hat{\theta}_k$ have to be sufficiently different. As an extreme example, one cannot use $\hat{\theta}_1 = \hat{\theta}_2 = \cdots = \hat{\theta}_K$. With "sufficiently different" we mean that the influence functions of the estimators have to be different. This is reflected in equation (15). If the influence functions are very similar, the denominator in (15) goes to zero and the procedure becomes increasingly unstable. More details can be found in [36].

One important takeaway from this methodology is that it is not the absolute stability (empirical variation of the estimators) that matters, but relative stability. In equation (15), we divide the variation $(\hat{\theta}_k - \hat{\theta}_\ell)^2$ by the expected variation under i.i.d. sampling $\frac{1}{n}\widehat{\text{Var}}_P(\phi_k(D) - \phi_\ell(D))$. If the actual variation is larger than the expected variation under i.i.d. sampling, we have some indication that there is distributional uncertainty.

4.2.3 *Calibration of uncertainty using knowledge about the superpopulation.* Knowledge about the superpopulation can be leveraged to estimate δ. As an example, the data scientist might know the average age or average income of the target population. Such knowledge can be expressed as moment equations. If the empirical average age is far from the target population average age, then this is an indication that either distribution or sampling uncertainty is high. Thus, we can use such knowledge to construct an estimator of δ. Let us now formalize this idea. As an example, assume that we know

$$\mu = \mathbb{E}_P[X],$$

where $\mu \in \mathbb{R}^K$. In this case, using equation (10), for any fixed k we can construct an unbiased estimate of $\delta^2 \text{Var}_P(X)$:

$$n(\overline{X}_{\bullet k} - \mu_k)^2.$$

Similarly, for each k we can construct an estimator of δ^2 by setting

$$\hat{\delta}_k^2 = \frac{n(\overline{X}_{\bullet k} - \mu_k)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(X_{ik} - \overline{X}_{\bullet k})^2}$$

Note that this is the squared t-test statistic. Even for $n \to \infty$, the variance of $\hat{\delta}_k^2$ does not go to zero. Under the non-i.i.d. sampling model, using equation (12), $\hat{\delta}_k^2/\delta^2$ follows a χ_1^2-distribution asymptotically. Precision can be gained by averaging

$$\hat{\delta}^2 = \frac{1}{K} \sum_{k=1}^{K} \hat{\delta}_k^2.$$

This estimate of δ can then be used to construct confidence intervals as described in equation (16). Under appropriate regularity assumptions, $\hat{\delta} \to \delta$. Thus, this approach will yield asymptotic coverage guarantees as in equation (18), see [36].

4.3 Calibrated Inference in R

In the following, we describe some functions available in the R-package available on GitHub (github.com/rothenhaeusler/calinf) that allow to quantify both sampling and distributional uncertainty. At the center is the approach described in Section 4.2.2. As an example, let us consider the problem of estimating the causal effect of some binary treatment $\text{Tr} \in \{0, 1\}$ on some outcome Y via linear regression, in the presence of some covariates $X_1, \ldots, X_5$. The practitioner might have several reasonable choices for confounder adjustment. Examples of variables that can (but are not necessarily included) in regression adjustment are exogeneous variables that affect the outcome, but not the treatment. Similarly, instrumental variables affect the treatment but are assumed to have no direct effect on the outcome. For valid treatment effect estimation, such adjustment variables can be (but do not have to be) included in a regression. These choices can be specified in a list of formulas:

```
formulas <- list(Y ~ Tr + X1 + X2,
                 Y ~ Tr + X1 + X2 + X3,
                 Y ~ Tr + X1 + X3 + X4,
                 Y ~ Tr + X1 + X2 + X5
                 )
```

In a second step, one can then run a calibrated linear regression:

```
    calm(formulas, data = data,
            target = "Tr")
```

Let us consider a concrete numerical example. We are interested in estimating the causal effect of a binary treatment variable Tr on Y in a structural causal model [51, 55]. Let P be the distribution of $(\mathrm{Tr}, I_1, X_1, X_2, J_1, J_2, Y)$ which is generated as follows:

$$\begin{aligned} X_1 &= \epsilon_1, \\ X_2 &= X_1 + \epsilon_2, \\ I_1 &= \epsilon_3, \\ J_1 &= \epsilon_4, \\ J_2 &= J_1 + \epsilon_5, \\ \mathrm{Tr} &= X_1 + X_2 + I_1 + \epsilon_6, \\ Y &= \mathrm{Tr} + X_1 - X_2 + J_1 + J_2 + \epsilon_7. \end{aligned}$$

Here, $\epsilon \sim \mathcal{N}(0, \mathrm{Id}_7)$. In words, I_1 is an instrument and (J_1, J_2) are variables that affect Y but not the treatment. We are interested in the direct causal effect of Tr on Y, which in this setting can be written as $\theta(P) = \arg\min_\theta \min_\beta \mathbb{E}_P[(Y - \mathrm{Tr}\cdot\theta - Z\beta)^2]$ for some appropriate set of adjustment variables Z. In this setting, there are multiple valid estimation strategies for $\theta(P)$. More precisely, all of the following formulas are valid in the sense that if one had infinite data from P, regression adjustment via these formulas would yield a consistent estimator of $\theta(P)$:

```
formulas <- list(
    Y ~ Tr + X1 + X2,
    Y ~ Tr + X1 + X2 + I_1,
    Y ~ Tr + X1 + X2 + J_1,
    Y ~ Tr + X1 + X2 + J_2,
    Y ~ Tr + X1 + X2 + J_2 + I_1,
    Y ~ Tr + X1 + X2 + J_1 + I_1,
    Y ~ Tr + X1 + X2 + J_1 + J_2,
    Y ~ Tr + X1 + X2 + J_1 + J_2 + I_1
                  )
```

We sample $n = 100$ observations from the random perturbation model with $\delta = 2$. The value δ is not known to the data scientist and thus has to be estimated. Running calibrated linear regression yields the following output:

```
> calm(formulas,df,target="Tr")

Quantification of both distributional
and sampling uncertainty

   Estimate Std. Error Pr(>|z|)
Tr   1.0157     0.0676        0

hat delta = 2.376035
```

As we can see, the estimated scaling factor $\hat{\delta}$ is somewhat close to $\delta = 2$. Estimation of δ is somewhat unstable across draws from the perturbation model. This is due to the fact that estimation of $\hat{\delta}$ has nonnegligible variance even for $n \to \infty$ (see equation (14)). Precision of $\hat{\delta}$ can be improved by adding additional estimators or moment constraints. From a statistical perspective, it is pertinent to investigate the validity of p-values across both sampling uncertainty and distributional uncertainty. To investigate the validity of p-values, we set the direct causal effect in the structural equation model to zero, that is, we set

$$Y = X_1 - X_2 + J_1 + J_2 + \epsilon_7.$$

We then compute naive p-values as reported by

```
lm(Y ~ Tr + X1 + X2)
```

and also compute calibrated p-values via

```
calm(formulas,df,target="Tr").
```

We repeat the two-stage sampling and estimation procedure $N = 1000$ times. The histograms of p-values are depicted in Figure 3.

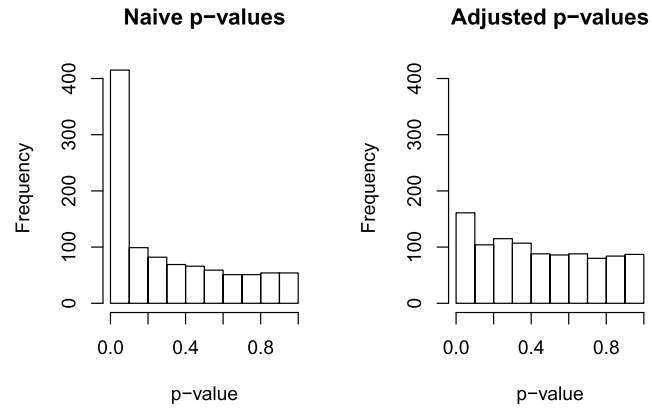

FIG. 3. *Example from Section 4.3. On the left-hand side, we show the histogram of $N = 1000$ naive p-values as reported by* `lm()`. *On the right-hand side, the p-values are computed via* `calm()`, *that is, the p-values are calibrated. The null hypothesis $\theta(P) = 0$ is true. As expected, the naive p-values are not valid for this hypothesis. In fact, more than* 40% *of the naive p-values are smaller than* 0.1. *While not perfect, the distribution of the adjusted p-values is much closer to a uniform distribution.*

The naive p-values are not valid for the hypothesis $\theta(P) = 0$, due to the distributional uncertainty. If there were no distributional uncertainty, the naive p-values would be valid. Intuitively speaking, the naive p-values are based on a variance formula that drastically underestimates uncertainty for the parameter $\theta(P)$. Thus, these p-values are anticonservative. The p-values as reported by calm, while not perfect, follow roughly a uniform distribution.

The R-package `calinf` provides functions also for calibrating inference in generalized linear models. If the outcome Y is binary, one can run calibrated logistic regression:

```
caglm(formulas, family = "binomial",
          data=data, target="Tr")
```

The function `caglm` is a wrapper for `glm`. Thus, one can run any generalized linear model by specifying an appropriate family in `caglm`.

Looking further, the proposed procedure in Section 4.2.2 is not limited to calibrate uncertainty only for generalized linear models. In principle, the proposed approach can be used for any asymptotically linear estimators. In the future, we aim to provide additional functionality that extend beyond these simple use cases.

4.4 Uniqueness of the Distributional Uncertainty Model

The discussion in the previous sections raises the question whether there are other nonadversarial perturbation models that would have led to different asymptotics. In this section, we give a negative answer to this question, within the assumed framework of a randomly perturbed distribution P^ξ.

In the following, for each realization of ξ let P^ξ be a probability measure on $\mathcal{D}$. To be more specific, we assume that $P^\bullet$ is a random probability measure. As an example, P^ξ might be constructed via random re-weighting with potentially nonexchangeable ξ_i's. In the following we will assume that P^ξ is "unbiased", that is, that for every measurable set $A \subseteq \mathcal{D}$ we have $E_\xi[P^\xi[D \in A]] = P[D \in A]$.

When considering distributional perturbation models, arguably there are two assumptions that may seem natural. First, one would like to have that events with probabilities close to zero are only perturbed very little (otherwise, P^ξ would be very different from P). To be more precise, we require that for every sequence of measurable sets $A_1, A_2, \ldots \subseteq \mathcal{D}$ with

$$P(D \in A_j) \to 0 \quad (j \to \infty)$$

we have

$$\mathrm{Var}_\xi\big(P^\xi(D \in A_j)\big) \to 0 \quad (j \to \infty). \tag{19}$$

In addition, we would like to exclude adversarial perturbations that only change a distribution in a very specific way. Mathematically, we model this by an isotropic perturbation, that means that events that have equal probability, are perturbed similarly. This can be seen as a symmetry assumption. To be specific, if $P(D \in A) = P(D \in B)$ then we assume that

$$\mathrm{Var}_\xi\big(P^\xi(D \in A)\big) = \mathrm{Var}_\xi\big(P^\xi(D \in B)\big). \tag{20}$$

THEOREM 1 ([36], Th.2). *Assume that* (19) *and* (20) *holds. Furthermore, assume that there exists a measurable function $u(D)$ such that $u(D) \sim \mathrm{Unif}([0,1])$, for $D \sim P$. Then, there exists $\delta_{\mathrm{dist}} \geq 0$ such that for any square-integrable function $f(D) \in L^2(P)$,*

$$\mathrm{Var}_\xi\big(\mathbb{E}_\xi[f(D)]\big) = \delta_{\mathrm{dist}}^2 \mathrm{Var}_P\big(f(D)\big).$$

The implication of Theorem 1 is as follows. Assume that conditionally on ξ, the data $(D_i)_{i=1,\ldots,n}$ is drawn i.i.d. from the perturbed distribution P^ξ. Then, for all square-integrable functions $f(D) \in L^2(P)$ we have

$$\mathrm{Var}_{\mathrm{marginal}}\left(\frac{1}{n}\sum_{i=1}^n f(D_i)\right) = \left(\frac{1}{n} + \delta_{\mathrm{dist}}^2 - \frac{\delta_{\mathrm{dist}}^2}{n}\right)\mathrm{Var}_P\big(f(D)\big).$$

Ignoring the lower-order term $\frac{\delta_{\mathrm{dist}}^2}{n}$, we can combine uncertainty due to sampling and uncertainty due to the distributional perturbation by setting

$$\delta^2 = 1 + n\delta_{\mathrm{dist}}^2.$$

Then, for all square-integrable functions $f(D) \in L^2(P)$ marginally across both sampling and distributional uncertainty we have

$$\mathrm{Var}_{\mathrm{marginal}}\left(\frac{1}{n}\sum_{i=1}^n f(D_i)\right) \approx \frac{\delta^2}{n}\mathrm{Var}_P\big(f(D)\big).$$

This corresponds to the perturbation model introduced in equation (10).

5. APPLICATION

We apply calibrated inference to a Get-Out-The-Vote field experiment, which investigates whether voter turnout can be increased by social pressure [22]. We study two groups: the "control" group and the "neighbors" group, and we refer to the latter also as the treatment group. The "neighbors" group received a mail with the statement "DO YOUR CIVIC DUTY—VOTE!". The letter lists the voting record of neighbors and threatens to publicize who does and does not vote. The outcome is voter turnout in the August 2006 primary election in Michigan.

The treatment is applied on the household level. On average, there are approximately 2 units per household.

Since units within households are correlated, the data should be analyzed using clustered standard errors. This data generation process can also be seen as a random perturbation model, where units only appear in the dataset if all other units in the household also appear in the dataset. That is, the treatment group is observed from a distribution which is different from the idealized one with i.i.d. sampling for which we want to infer the treatment effect. We want to emulate a scenario where the data is not drawn i.i.d. from the target distribution, with unknown correlations between units. Thus, we drop the household indicator, and hope to recover valid inferential statements by calibrating the p-values.

Since the ground truth is unknown, we re-randomize the treatment variable to simulate a setting where the treatment effect is zero. The covariates and outcomes are left unchanged.

In this setup, one expects the correlation within household units to inflate the variance compared to i.i.d. sampling. In our scenario for illustration, as mentioned above, the household indicator is considered unknown. Thus, we have to infer the variance inflation factor δ from data alone. To estimate δ, we use super-population constraints as described in Section 4.2.3. To form these constraints, we use that for each individual we have records whether they voted in the primary elections in 2000, 2002, and 2004 or the general election in 2000 and 2002. For each of these covariates, as super-population constraints we assume that the covariance between treatment and covariates is zero. Intuitively, if the empirical covariance between treatment and covariates is significantly different from zero under an i.i.d. sampling, there is evidence of positive associations between units.

There are $n = 119{,}999$ households in the dataset that were subject to the treatment or control group. We randomly select $m = 1200 \approx n/100$ households and compute calibrated p-values as well as naive p-values via difference-in-means, assuming that the household identifier is unknown. This process was repeated 10,000 times. The resulting p-values are depicted in Figure 4. The calibrated p-values follow much closer a uniform distribution (which is correct) than the naive p-values. Around 11% of the naive p-values are below 0.05 while only 6.5% of the calibrated p-values are below 0.05. This indicates that the calibration procedure succeeded at capturing the excess variation due to unobserved clustering.

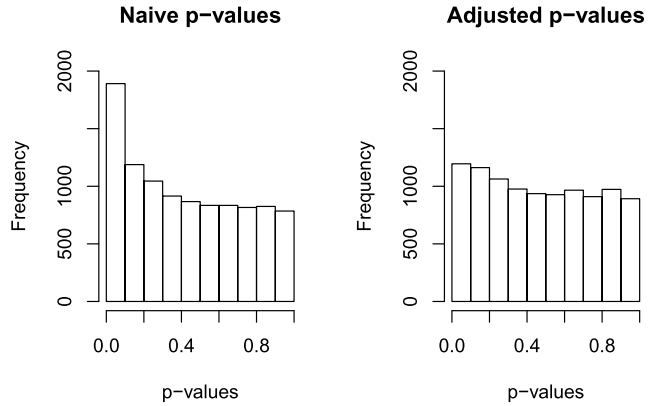

FIG. 4. *Calibrated inference for the Get-Out-The-Vote field experiment* [22]. *On the left-hand side, we show the histogram of* $N = 1000$ *naive* p*-values computed via difference-in-means. On the right-hand side, the* p*-values are calibrated using super-population constraints. The treatment has been re-randomized to guarantee that the null hypothesis* $\theta(P) = 0$ *is true. Thus, the* p*-values should follow a uniform distribution. As some of the units are positively correlated, the naive* p*-values are not valid. Around* 11% *of the naive* p*-values are below* 0.05 *while only* 6.5% *of the adjusted* p*-values are below* 0.05. *The empirical distribution of adjusted* p*-values is much closer to a uniform distribution.*

6. DISCUSSION AND OUTLOOK

We summarize the main points of our exposition and outline how the propagated ideas can potentially be extended to improve replicability and generalizability.

In many practical problems, the data is not drawn i.i.d. from the target population. For example, unobserved sampling bias, confounding, batch effects, or unknown associations can inflate the deviation of the estimator from its target compared to i.i.d. sampling. For reliable statistical inference, it is of paramount importance to account for these additional types of uncertainty. Failure to do so is a major source of lack of replicability of scientific findings in many fields.

We present two approaches to deal with such distribution shifts. In Section 3, we consider a directional notion of distributional stability. In the existing literature, distributional errors are often handled via worst-case bounds. Such bounds can be very conservative and may lead to rather limited information gain as some type of shifts might be more realistic than others. The directional notion of stability allows to probe different perturbations to investigate what type of distribution shift the estimand is most sensitive to. This then leads to a less conservative notion of sensitivity and helps to judge which type of distribution shifts one should be most worried about.

In Section 4, we go beyond worst-case stability. All of the worst-case bounds have in common that some background knowledge of the strength of shifts or confounding is needed to form and interpret these bounds. In contrast, we consider a model that shifts the distribution randomly. This allows to consider average distributional robustness. In addition, it turns out that in such a random perturbation model, it is possible to estimate the size of perturbations by using knowledge in form of moment equations. Such background knowledge can come in the form of having multiple valid estimators for a single target quantity. Based on these estimators, it is possible to form "calibrated" confidence intervals that are valid on average, where we average both over sampling uncertainty and the

distributional perturbation. Procedures to sample from the distributional perturbation model and conduct calibrated inference are implemented in the R-package calinf available at github.com/rothenhaeusler/calinf.

Looking ahead, there are multiple directions that we believe are promising avenues for future research.

When having access to *multiple datasets* or *multisource data*, we can model the different datasets arising from perturbed data generating distributions. In such a context, we point to the following.

Transfer learning under random shifts. In the literature, one often makes a covariate shift assumption, that is, that the conditional distribution of a target Y given a subset of observed attributes stays the same. The distributional perturbation model from Section 4 allows to go beyond this assumption, by allowing for (non)-adversarial shifts even in conditional distributions. We can then formalize optimal transfer learning under random perturbations.

Data fusion across heterogeneous datasets. We can model the differences between multiple datasets as random, as in Section 4. This may lead to straightforward extensions of statistical methodology and optimality results (such as the Cramér-Rao lower bound or semiparametric efficiency bounds) to distributional counterparts.

Multiple testing in the context of distribution shifts. It is well known how to account for multiple testing in the context of sampling uncertainty. Similar issues are at play under multiple distributional perturbations: if we have 100 studies for which the null hypothesis holds but in each of those studies we sample from a randomly perturbed distribution $P' \neq P_{\text{target}}$, it is quite likely that we will get too many false positives since some of the distributions will be strongly perturbed. The more perturbed distributions we look at, the more likely it is that we'll make a false discovery. This suggests that we should account for multiple testing also in distributional stability measures.

Without relying on multiple data sources, we also mention the following.

Nonadversarial confounding. Sensitivity analysis in causal inference investigates the stability of a causal conclusion by taking the worst-case confounded distribution given some restrictions on the strength of confounding. Such bounds are often very conservative. A less pessimistic assumption would be to model unobserved confounding as random (nonadversarial). A random confounding model, perhaps similar to the one in Section 4, potentially opens the door for novel average sensitivity procedures for causal inference.

ACKNOWLEDGMENTS

We thank the Guest Editors and the Editor for the opportunity of presenting our work and the reviewers for constructive comments. The research was partially conducted during D. Rothenhäusler's research stay at the Institute for Mathematical Research at ETH Zürich (FIM).

FUNDING

The research of D. Rothenhäusler was supported by the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

The research of P. Bühlmann was supported by the European Research Council under the Grant Agreement No 786461 (CausalStats—ERC-2017-ADG).

REFERENCES

[1] ANGRIST, J., IMBENS, G. and RUBIN, D. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc*. **91** 444–455.

[2] ARJOVSKY, M., BOTTOU, L., GULRAJANI, I. and LOPEZ-PAZ, D. (2019). Invariant risk minimization. arXiv preprint, arXiv:1907.02893.

[3] BAKTASHMOTLAGH, M., HARANDI, M. T., LOVELL, B. C. and SALZMANN, M. (2013). Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision* 769–776.

[4] BELSLEY, D. A., KUH, E. and WELSCH, R. E. (1980). *Regression Diagnostics*: *Identifying Influential Data and Sources of Collinearity*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York–Chichester–Brisbane. MR0576408

[5] BEN-TAL, A. and NEMIROVSKI, A. (2002). Robust optimization—methodology and applications. *Math. Program*. **92** 453–480. MR1905762 https://doi.org/10.1007/s101070100286

[6] BENJAMINI, Y. and HELLER, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64** 1215–1222. MR2522270 https://doi.org/10.1111/j.1541-0420.2007.00984.x

[7] BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist*. **41** 802–837. MR3099122 https://doi.org/10.1214/12-AOS1077

[8] BERTSIMAS, D., BROWN, D. B. and CARAMANIS, C. (2011). Theory and applications of robust optimization. *SIAM Rev.* **53** 464–501. MR2834084 https://doi.org/10.1137/080734510

[9] BÜHLMANN, P. (2014). Discussion of big Bayes stories and BayesBag. *Statist. Sci*. **29** 91–94. MR3201850 https://doi.org/10.1214/13-STS460

[10] BÜHLMANN, P. (2020). Invariance, causality and robustness: 2018 Neyman Lecture. *Statist. Sci*. **35** 404–426. MR4148216 https://doi.org/10.1214/19-STS721

[11] BÜHLMANN, P. and MEINSHAUSEN, N. (2015). Magging: Maximin aggregation for inhomogeneous large-scale data. *Proc. IEEE* **104** 126–135.

[12] CHEN, Y. and BÜHLMANN, P. (2021). Domain adaptation under structural causal models. *J. Mach. Learn. Res*. **22** Paper No. [261], 80. MR4353040 https://doi.org/10.1007/s11081-020-09512-z

[13] CINELLI, C. and HAZLETT, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **82** 39–67. MR4060976

[14] CORNFIELD, J., HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A. M., SHIMKIN, M. B. and WYNDER, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst*. **22** 173–203.

[15] DAHABREH, I. J., PETITO, L. C., ROBERTSON, S. E., HERNÁN, M. A. and STEINGRIMSSON, J. A. (2020). Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. *Epidemiology* **31** 334–344.

[16] DENZEN, N. (1978). Sociological methods: A sourcebook. New York.
[17] DEVAUX, M. and EGAMI, N. (2022). Quantifying robustness to external validity bias.
[18] DEZEURE, R., BÜHLMANN, P., MEIER, L. and MEINSHAUSEN, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software `hdi`. *Statist. Sci.* **30** 533–558. MR3432840 https://doi.org/10.1214/15-STS527
[19] DING, P. and VANDERWEELE, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27** 368.
[20] DONSKER, M. D. and VARADHAN, S. R. S. (1976). Asymptotic evaluation of certain Markov process expectations for large time. III. *Comm. Pure Appl. Math.* **29** 389–461. MR0428471 https://doi.org/10.1002/cpa.3160290405
[21] DORN, J., GUO, K. and KALLUS, N. (2021). Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. arXiv preprint, arXiv:2112.11449.
[22] GERBER, A. S., GREEN, D. P. and LARIMER, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *Amer. Polit. Sci. Rev.* **102** 33–48.
[23] GONG, B., SHI, Y., SHA, F. and GRAUMAN, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. In 2012 *IEEE Conference on Computer Vision and Pattern Recognition* 2066–2073. IEEE.
[24] GOPALAN, R., LI, R. and CHELLAPPA, R. (2011). Domain adaptation for object recognition: An unsupervised approach. In 2011 *International Conference on Computer Vision* 999–1006. IEEE.
[25] GUPTA, S. and ROTHENHÄUSLER, D. (2021). The s-value: Evaluating stability with respect to distributional shifts. To appear in *Proceedings of the 37th International Conference on Neural Information Processing Systems*.
[26] HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics*: *The Approach Based on Influence Functions*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. MR0829458
[27] HEINZE-DEML, C. and MEINSHAUSEN, N. (2021). Conditional variance penalties and domain shift robustness. *Mach. Learn.* **110** 303–348. MR4207502 https://doi.org/10.1007/s10994-020-05924-1
[28] HEINZE-DEML, C., PETERS, J. and MEINSHAUSEN, N. (2018). Invariant causal prediction for nonlinear models. *J. Causal Inference* **6** Art. No. 20170016, 35. MR4335430 https://doi.org/10.1515/jci-2017-0016
[29] HELLER, R., GOLLAND, Y., MALACH, R. and BENJAMINI, Y. (2007). Conjunction group analysis: An alternative to mixed/random effect analysis. *NeuroImage* **37** 1178–1185.
[30] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732
[31] HUBER, P. J. (1965). A robust version of the probability ratio test. *Ann. Math. Stat.* **36** 1753–1758. MR0185747 https://doi.org/10.1214/aoms/1177699803
[32] HUGGINS, J. H. and MILLER, J. W. (2023). Reproducible model selection using bagged posteriors. *Bayesian Anal.* **18** 79–104. MR4515726 https://doi.org/10.1214/21-ba1301
[33] IMBENS, G. W. (2014). Instrumental variables: An econometrician's perspective. *Statist. Sci.* **29** 323–358. MR3264545 https://doi.org/10.1214/14-STS480
[34] IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics*, *Social*, *and Biomedical Sciences*: *An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751
[35] IOANNIDIS, J. P. A. (2005). Why most published research findings are false. *Chance* **18** 40–47. MR2216666 https://doi.org/10.1080/09332480.2005.10722754
[36] JEONG, Y. and ROTHENHÄUSLER, D. (2022). Calibrated inference: Statistical inference that accounts for both sampling uncertainty and distributional uncertainty. arXiv preprint, arXiv:2202.11886.
[37] JIN, Y., REN, Z. and CANDÈS, E. J. (2023). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proc. Natl. Acad. Sci. USA* **120** Paper No. e2214889120, 13. MR4575282
[38] LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371
[39] LI, S., SONG, S. and HUANG, G. (2017). Prediction reweighting for domain adaption. *IEEE Trans. Neural Netw. Learn. Syst.* **28** 1682–1695. MR3666190 https://doi.org/10.1109/TNNLS.2016.2538282
[40] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. MR3210970 https://doi.org/10.1214/13-AOS1175
[41] LONG, M., WANG, J., DING, G., SUN, J. and YU, P. S. (2014). Transfer joint matching for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1410–1417.
[42] MEINSHAUSEN, N. (2018). Causality from a distributional robustness point of view. In 2018 *IEEE Data Science Workshop* (*DSW*) 6–10. IEEE.
[43] MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523 https://doi.org/10.1111/j.1467-9868.2010.00740.x
[44] MEINSHAUSEN, N. and BÜHLMANN, P. (2015). Maximin effects in inhomogeneous large-scale data. *Ann. Statist.* **43** 1801–1830. MR3357879 https://doi.org/10.1214/15-AOS1325
[45] MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681. MR2750584 https://doi.org/10.1198/jasa.2009.tm08647
[46] MODIGLIANI, F. (1966). The life cycle hypothesis of saving, the demand for wealth and the supply of capital. *Soc. Res.* 160–217.
[47] MUNAFÒ, M. R. and SMITH, G. D. (2018). Repeating experiments is not enough. *Nature* **553** 399–401.
[48] NEYKOV, M., NING, Y., LIU, J. S. and LIU, H. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statist. Sci.* **33** 427–443. MR3843384 https://doi.org/10.1214/18-STS661
[49] PAN, S. J. and YANG, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22** 1345–1359.
[50] PATTON, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Serv. Res.* **34** 1189.
[51] PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161
[52] PEARL, J. and BAREINBOIM, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.
[53] PENG, X., BAI, Q., XIA, X., HUANG, Z., SAENKO, K. and WANG, B. (2019). Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 1406–1415.
[54] PETERS, J., BÜHLMANN, P. and MEINSHAUSEN, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*

78 947–1012. With comments and a rejoinder. MR3557186 https://doi.org/10.1111/rssb.12167

[55] PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2017). *Elements of Causal Inference*: *Foundations and Learning Algorithms*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3822088

[56] QUINONERO-CANDELA, J., SUGIYAMA, M., SCHWAIGHOFER, A. and LAWRENCE, N. D. (2009). *Dataset Shift in Machine Learning*. Mit Press.

[57] ROJAS-CARULLA, M., SCHÖLKOPF, B., TURNER, R. and PETERS, J. (2018). Invariant models for causal transfer learning. *J. Mach. Learn. Res.* **19** Paper No. 36, 34. MR3862443

[58] ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. MR0885915 https://doi.org/10.1093/biomet/74.1.13

[59] ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P. and PETERS, J. (2021). Anchor regression: Heterogeneous data meet causality. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 215–246. MR4250274 https://doi.org/10.1111/rssb.12398

[60] ROTHWELL, P. M. (2005). External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet* **365** 82–93.

[61] SAGAWA, S., KOH, P. W., HASHIMOTO, T. B. and LIANG, P. (2019). Distributionally robust neural networks. In *International Conference on Learning Representations*.

[62] SINHA, A., NAMKOONG, H. and DUCHI, J. (2017). Certifiable distributional robustness with principled adversarial training. arXiv preprint, arXiv:1710.10571, presented at Sixth International Conference on Learning Representations (ICLR 2018).

[63] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

[64] VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12** 1221–1274. With a rejoinder by the authors. MR3724985 https://doi.org/10.1214/17-BA1065

[65] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[66] WANG, J. and OWEN, A. B. (2019). Admissibility in partial conjunction testing. *J. Amer. Statist. Assoc.* **114** 158–168. MR3941245 https://doi.org/10.1080/01621459.2017.1385465

[67] WITTEVEEN, E., WIESKE, L., SOMMERS, J., SPIJKSTRA, J.-J., DE WAARD, M. C., ENDEMAN, H., RIJKENBERG, S., DE RUIJTER, W., SLEESWIJK, M. et al. (2020). Early prediction of intensive care unit–acquired weakness: A multicenter external validation study. *J. Intens. Care Med.* **35** 595–605.

[68] YADLOWSKY, S., NAMKOONG, H., BASU, S., DUCHI, J. and TIAN, L. (2022). Bounds on the conditional and average treatment effect with unobserved confounding factors. *Ann. Statist.* **50** 2587–2615. MR4505372 https://doi.org/10.1214/22-aos2195

[69] YU, B. (2013). Stability. *Bernoulli* **19** 1484–1500. MR3102560 https://doi.org/10.3150/13-BEJSP14

[70] YU, B. and KUMBIER, K. (2020). Veridical data science. *Proc. Natl. Acad. Sci. USA* **117** 3920–3929. MR4075122 https://doi.org/10.1073/pnas.1901326117

[71] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

[72] ZHAO, Q., SMALL, D. S. and BHATTACHARYA, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 735–761. MR3997099