PNAS RESEARCH ARTICLE | STATISTICS OPEN ACCESS

Cross-prediction-powered inference

Tijana Zrnic[a,b] and Emmanuel J. Candès[a,c,1]

Contributed by Emmanuel J. Candès; received December 14, 2023; accepted March 5, 2024; reviewed by Alessandro Rinaldo and Larry Wasserman

While reliable data-driven decision-making hinges on high-quality labeled data, the acquisition of quality labels often involves laborious human annotations or slow and expensive scientific measurements. Machine learning is becoming an appealing alternative as sophisticated predictive techniques are being used to quickly and cheaply produce large amounts of predicted labels; e.g., predicted protein structures are used to supplement experimentally derived structures, predictions of socioeconomic indicators from satellite imagery are used to supplement accurate survey data, and so on. Since predictions are imperfect and potentially biased, this practice brings into question the validity of downstream inferences. We introduce cross-prediction: a method for valid inference powered by machine learning. With a small labeled dataset and a large unlabeled dataset, cross-prediction imputes the missing labels via machine learning and applies a form of debiasing to remedy the prediction inaccuracies. The resulting inferences achieve the desired error probability and are more powerful than those that only leverage the labeled data. Closely related is the recent proposal of prediction-powered inference [A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, T. Zrnic, *Science* 382, 669–674 (2023)], which assumes that a good pretrained model is already available. We show that cross-prediction is consistently more powerful than an adaptation of prediction-powered inference in which a fraction of the labeled data is split off and used to train the model. Finally, we observe that cross-prediction gives more stable conclusions than its competitors; its CIs typically have significantly lower variability.

statistical inference | CIs | machine learning | prediction

Significance

Machine learning is increasingly used as an efficient substitute for traditional data collection when the latter is challenging. For example, predictions of conditions such as poverty, deforestation, and population density based on satellite imagery are used to supplement accurate survey data, which requires significant time and resources to collect. However, predictions are imperfect and potentially biased, calling into question the validity of conclusions drawn from such data. This manuscript introduces a method for valid inference powered by machine learning. The method enables researchers to draw more reliable and accurate conclusions from machine learning predictions.

As data-driven decisions fuel progress across science and technology, ensuring that such decisions are reliable is of critical importance. The reliability of data-driven decision-making rests on having access to high-quality data on one hand, and properly accounting for uncertainty on the other.

One frequently discussed issue is that acquiring high-quality data often involves laborious human labeling, or slow and expensive scientific measurements, or overcoming privacy concerns when human subjects are involved. Machine learning offers a promising alternative: Sophisticated techniques such as generative modeling and deep neural networks are being used to cheaply produce large amounts of data that would otherwise be too expensive or time-consuming to collect. For example, tools to predict protein structure are supporting wide-ranging research in biology (1–4); large language models are being used to generate difficult-to-aggregate information about materials that can be used to fight climate change (5); predictions of socioeconomic and environmental conditions based on satellite imagery are being used for downstream policy decisions (6–9). This increasingly common practice, marked by supplementing high-quality data with machine learning outputs, calls for new principles of uncertainty quantification.

In this work, we study this problem in the semisupervised context, where labels are scarce but features are abundant. For example, precise measurements of environmental conditions are difficult to come by but satellite imagery is abundant. Due to its volume, satellite imagery is routinely used in combination with computer vision algorithms to predict a range of factors on a global scale, including deforestation (10), poverty rates (6), and population densities (11). These predictions provide a compelling substitute for resource-intensive ground-based measurements and surveys. However, it is crucial to acknowledge that, while promising, the predictions are not infallible. Consequently, downstream inferences that uncritically treat them as ground truth will be invalid.

We introduce cross-prediction: a broadly applicable method for semisupervised inference that leverages the power of machine learning while retaining validity. Assume a researcher holds both a small labeled dataset and a large unlabeled dataset, and they seek inference—i.e., a *P*-value or a CI—about a population-level quantity such as the mean outcome or a regression coefficient. Cross-prediction carefully leverages black-box machine learning to impute the missing labels, resulting in both valid and powerful

Author affiliations: [a]Department of Statistics, Stanford University, Stanford, CA 94305; [b]Stanford Data Science, Stanford University, Stanford, CA 94305; and [c]Department of Mathematics, Stanford University, Stanford, CA 94305

Author contributions: T.Z. and E.J.C. designed research; performed research; analyzed data; and wrote the paper.

Reviewers: A.R., The University of Texas at Austin; and L.W., Carnegie Mellon University.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: candes@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2322083121/-/DCSupplemental.

Published April 3, 2024.

Fig. 1. Examples of GEE satellite imagery used in the deforestation analysis of Bullock et al. (13).

inferences. The validity is a result of a particular debiasing step; the power is a result of using sophisticated predictive techniques such as deep learning or random forests. We show that the use of black-box predictions on the unlabeled data can lead to a massive improvement in statistical power compared to using the labeled data alone.

Cross-prediction builds upon the recent proposal of prediction-powered inference (12). Unlike prediction-powered inference, we do not assume that our researcher already has access to a predictive model for imputing the labels. Rather, to apply prediction-powered inference, the researcher would need to use a portion of the labeled data to either train a model from scratch or fine-tune an off-the-shelf model. We show that this leads to a suboptimal solution. Consider the following example studied by Angelopoulos et al. (12). The goal is to form a CI for the fraction of the Amazon rainforest that was lost between 2000 and 2015. A small number of "gold-standard" deforestation labels for certain parcels of land are available, having been collected through field visits (13). In addition, satellite imagery is available for the entire Amazon; see Fig. 1 for Google Earth Engine (GEE) examples used in the deforestation study of Bullock et al. (13). Angelopoulos et al. apply prediction-powered inference after using a portion of the labeled data and a gradient-boosted tree to fine-tune a regression-tree-based predictor of forest cover (14). Our work offers an alternative: We can avoid data splitting and instead apply cross-prediction, still with a gradient-boosted tree, to perform the fine-tuning. By doing so, we significantly reduce the size of the CI, as seen in Fig. 2. This trend will be consistent throughout our experiments: Cross-prediction is more efficient than prediction-powered inference with data splitting. Fig. 2 also shows that cross-prediction outperforms "classical" inference, which forms a CI based on gold-standard labels only and simply ignores the unlabeled data. Additional details about these experiments can be found in the Experiments section.

Another important takeaway from Fig. 2 is that cross-prediction gives more stable inferences: The confidence intervals have lower variability than the intervals computed via baseline approaches. Intuitively, classical inference has higher variability due to the smaller sample size, while prediction-powered inference has higher variability due to the arbitrariness in the data split. We will quantify the stability of cross-prediction in the Experiments section, showcasing its superiority across a range of examples; see Table 4.

Our work is also related to the literature known as semisupervised inference (15). The main difference between existing approaches and our work is that our proposal leverages black-box machine learning methods, allowing for more complicated data modalities (such as high-dimensional imagery) and more sophisticated ways of leveraging the unlabeled data. We elaborate on the relationship to prior work after introducing the formal problem setup.

Problem Setup

We study statistical inference in a semisupervised setting, where collecting high-quality labels is challenging but feature observations are abundant. Formally, we have a dataset consisting of n i.i.d. feature–label pairs, $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \sim \mathbb{P}^n$. In addition, we have a dataset consisting of N unlabeled data points, $\{\tilde{X}_1, \ldots, \tilde{X}_N\} \sim \mathbb{P}_X^N$, where $\mathbb{P}_X$ denotes the marginal distribution over features according to $\mathbb{P}$. We are most interested in the regime where $N \gg n$, as in the case where feature collection is far cheaper than label collection.

Our goal is to perform inference on a property $\theta^*(\mathbb{P})$ of the data-generating distribution $\mathbb{P}$, such as the mean outcome, a quantile of the outcome distribution, or a regression coefficient. Our proposal handles all estimands defined as a solution to an M-estimation problem:

$$\theta^*(\mathbb{P}) = \arg\min_\theta L(\theta), \text{ where } L(\theta) := \mathbb{E}\left[\ell_\theta(X, Y)\right], \qquad [1]$$

for a convex loss function ℓ_θ. Here and throughout, (X, Y) denotes a generic sample from $\mathbb{P}$ independent of everything else. All of the aforementioned estimands can be cast in the form Eq. **1**. For example, if the target of inference is the mean outcome, $\theta^*(\mathbb{P}) = \mathbb{E}[Y]$, then $\theta^*(\mathbb{P})$ minimizes the squared loss:

$$\theta^*(\mathbb{P}) = \arg\min_\theta \mathbb{E}[\ell_\theta(Y)] = \arg\min_\theta \mathbb{E}[(Y-\theta)^2]. \qquad [2]$$

Note that the estimand (and thus the loss) will sometimes only depend on a subset of the features X or only on the outcome Y, as in Eq. **2**. Also note that this manuscript focuses on $\theta^*(\mathbb{P}) \in \mathbb{R}^d$ for a fixed d. Studying high-dimensional settings—for example, understanding what scaling of d is permitted by the theory—is a valuable direction for future work. Below, we write $\theta^* = \theta^*(\mathbb{P})$ for short.

The main question we address is this: How should we leverage the unlabeled data to achieve both valid and powerful inference? Validity alone is an easy target: We can simply dispense with the unlabeled data and find the classical estimator $\hat{\theta}^{\text{class}}$, defined as

$$\hat{\theta}^{\text{class}} = \arg\min_\theta L^{\text{class}}(\theta), \text{ where } L^{\text{class}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i, Y_i). \qquad [3]$$

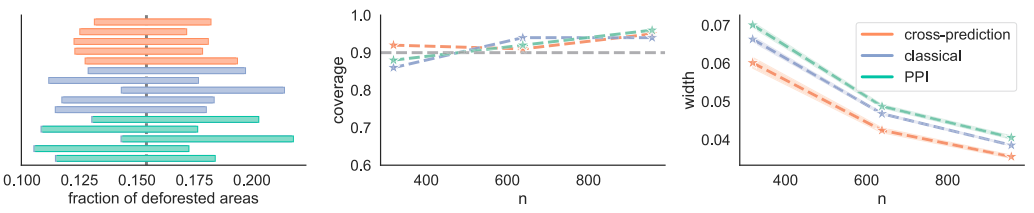

Fig. 2. Estimating the deforestation rate in the Amazon from satellite imagery. *Left*: Example intervals constructed by cross-prediction, classical inference, and prediction-powered inference (PPI), for five random splits into labeled and unlabeled data and a fixed number of gold-standard deforestation labels, $n = 319$. *Middle* and *Right*: Coverage and interval width averaged over 100 random splits into labeled and unlabeled data, for $n \in \{319, 638, 957\}$. The target of inference is the fraction of the Amazon rainforest lost between 2000 and 2015 (gray line in *Left* panel). The target coverage is 90% (gray line in *Middle* panel).

For all standard estimands defined via M-estimation—including means, quantiles, linear regression coefficients—there are off-the-shelf confidence intervals around $\hat{\theta}^{\text{class}}$ that cover θ^* with a desired probability in the large-sample limit, see, e.g., refs. 16 and 17. The classical estimator and the corresponding CIs shall be the main comparison points used to evaluate the performance of cross-prediction.

Related Work

We discuss the relationship between our work and the most closely related technical scholarship.

Semisupervised Inference. Our work falls within the literature known as semisupervised inference (15). Most existing work develops methods specialized to particular estimation problems, such as mean estimation (15, 18), quantile estimation (19), or linear regression (20, 21). One exception is the recent work of Song et al. (22), who also study general M-estimation. Their approach uses a projection-based correction to the classical loss Eq. **3** based on simple statistics from the unlabeled data, such as averages of low-degree polynomials of the features. Unlike existing proposals, our approach is based on imputing the missing labels using black-box machine learning methods, allowing for more complicated data modalities and more intricate ways of leveraging the unlabeled data. For example, it is unclear how to apply existing methods when the features X_i are high-dimensional images. We also note that the semisupervised observation model has been long studied in semisupervised learning (23, 24). However, in this literature, the goal is prediction, rather than inference.

Prediction-Powered Inference. The core idea in this paper is to correct imputed predictions, and this derives from the proposal of prediction-powered inference (12). However, a key assumption in prediction-powered inference is that, in addition to a labeled and an unlabeled dataset, the analyst is given a good pretrained machine learning model. We make no such assumption. To apply the theory of prediction-powered inference, our setting would require using a portion of the labeled data for model training and leaving the rest for inference. In contrast, cross-prediction leverages each labeled data point for both model training and inference, leading to a boost in statistical power. The distinction between having and not having a pretrained model makes a difference even when comparing prediction-powered inference and the classical approach. Angelopoulos et al. (12) do not take into account the data used for model training when comparing the two baselines, because the model is assumed to have been trained before the analysis takes place. This makes sense when considering off-the-shelf models such as AlphaFold. In our comparisons, we do take the training data into account.

Angelopoulos et al. (25) show a central limit theorem for the prediction-powered estimator, allowing for computational and statistical improvements of the original methods for prediction-powered inference. Our inferences will be based on a similar central limit theorem for cross-prediction.

Wang et al. (26) similarly study inferences based on machine learning predictions. They propose using the labeled data to train a predictor of true outcomes from predicted ones, and then applying the predictor to debias the predictions on the unlabeled data. This algorithm does not come with a formal validity guarantee. Motwani and Witten (27) conduct a detailed empirical comparison of the method of Wang et al. and prediction-powered inference.

Theory of Cross-Validation. Cross-prediction is based on a form of cross-fitting. Consequently, our analysis is related to the theoretical studies of cross-validation (28–32). In particular, our theory borrows from the analysis of Bayle et al. (30), who prove a central limit theorem and study inference on the cross-validation test error. Our goal, however, is entirely different; we aim to provide inferential guarantees for an estimand θ^*, as defined in Eq. **1**, in a semisupervised setting.

Semiparametric Inference. Our work is also related to the rich literature on semiparametric inference (33–40), where the goal is to do estimation in the presence of a high-dimensional nuisance parameter. Our debiasing strategy closely resembles doubly robust estimators (41), such as the AIPW estimator (42, 43), and one-step estimators (44). In this literature, the estimand is typically an expected value, such as the average treatment effect. One exception is the work of Jin and Rothenhäusler (45), who study general M-estimators through a semiparametric lens. The use of cross-fitting is common in that literature as well (40, 46, 47). While the technical arguments used in our work bear resemblance to those classically used in semiparametric inference, our motivation is different. Our focus is on showcasing how a theoretically principled use of black-box predictors—neural networks, random forests, and so on—on massive amounts of unlabeled data can boost inference. Since the practice of leveraging unlabeled data through predictions is already prevalent in domains such as remote sensing, our goal is to ground it in statistical theory.

Inference with Missing Data. Semisupervised inference can be seen as a special case of the problem of inference with missing data (48), where missing information about the labels occurs. Our proposed method bears similarities to multiple imputation (49–51) as, at least at a high level, it is based on "averaging out" multiple imputed predictions for the labels. However, our method is substantially different from multiple imputation, most notably due to the fact that it incorporates a particular form of debiasing to mitigate prediction inaccuracies.

Inference under Model Misspecification. Finally, our work relates to a large body of work on inference under model misspecification e.g., refs. 52–55. In particular, we do not assume that our predictions follow any "true" statistical model, and for parameters θ^* defined as a regression solution, we do not assume that the regression model is correct. For example, if θ^* is the solution to a linear regression, we do not assume that the data truly follows a linear model. Like in classical M-estimation, we will show asymptotic normality of our estimator despite the misspecification.

Cross-Prediction

We propose cross-prediction—an estimation technique based on a combination of cross-fitting and prediction. The basic idea is to impute labels for the unlabeled data points, and then remove the bias arising from the inaccuracies in the predictions using the labeled data. We give a step-by-step outline of the construction of the cross-prediction estimator. In the following sections, we will show how to perform inference with this estimator; that is, how to perform hypothesis tests or construct confidence intervals for θ^*.

Cross-Prediction for Mean Estimation. Before discussing the general case, we consider the problem of mean estimation to gain intuition; the object of inference is simply $\theta^* = \mathbb{E}[Y]$.

We begin by partitioning the labeled dataset into K chunks, $I_1 = \{1, \ldots, n/K\}$, $I_2 = \{n/K + 1, \ldots, 2n/K\}$, and so on (we assume for simplicity that n is divisible by K).* Here, K is a user-specified number of folds, e.g., $K = 10$. Then, as in cross-validation, we train a machine learning model K times, each time training on all data except one fold. Let $\mathcal{A}_{\text{train}}$ denote a possibly randomized training algorithm, which takes as input a dataset of arbitrary size and outputs a predictor of labels from features. Then, for each $j \in [K]$, let $f^{(j)}$ be the model obtained by training on all folds but I_j; that is, $f^{(j)} = \mathcal{A}_{\text{train}}(\{(X_i, Y_i)\}_{i \in [n] \setminus I_j})$. We note that $\mathcal{A}_{\text{train}}$ can be quite general; it may or may not treat the training data points symmetrically, and $f^{(j)}$ need not come from a well-defined family of predictors. Rather, $f^{(j)}$ can be any black-box model; e.g., a random forest, a gradient-boosted tree, a neural network, and so on. Moreover, $f^{(j)}$ can be trained from scratch or obtained by fine-tuning an off-the-shelf model. Finally, we use the trained models to impute predictions and compute the cross-prediction estimator, defined as

$$\hat{\theta}^{+} = \frac{1}{KN} \sum_{j=1}^{K} \sum_{i=1}^{N} f^{(j)}(\tilde{X}_i) - \frac{1}{n} \sum_{j=1}^{K} \sum_{i \in I_j} (f^{(j)}(X_i) - Y_i). \quad [4]$$

Intuitively, the first term in Eq. **4** is an empirical approximation of the population mean if we treated the predictions as true labels. The second term in Eq. **4** serves to debias this heuristic: It subtracts an estimate of the bias between the predicted labels and the true labels. We note that the estimator Eq. **4** coincides with the mean estimator of Zhang and Bradic (18) in the special case where $f^{(j)}$ are linear models, that is, $f^{(j)}(x) = x^{\top} \beta_j$ for some β_j. Our analysis applies more broadly, allowing for complex high-dimensional models (e.g., image classifiers).

Observe that the cross-prediction estimator is unbiased, i.e., $\mathbb{E}[\hat{\theta}^{+}] = \theta^*$. Indeed, since $i \in I_j$ is not used to train model $f^{(j)}$, we have $\mathbb{E}[f^{(j)}(\tilde{X}_{i'})] = \mathbb{E}[f^{(j)}(X_i)]$ for all $j \in [K], i \in I_j, i' \in [N]$. Applying this identity yields $\mathbb{E}[\hat{\theta}^{+}] = \mathbb{E}[Y] = \theta^*$.

The classical estimator is of course the sample mean:

$$\hat{\theta}^{\text{class}} = \frac{1}{n} \sum_{i=1}^{n} Y_i, \quad [5]$$

which is also unbiased. Given that both the cross-prediction estimator and the classical estimator are unbiased, it makes sense to ask which one has a lower variance. The main benefit of cross-prediction is that, if the trained models $f^{(j)}$ are reasonably accurate, we expect the variance of the cross-prediction estimator to be lower. To see this, first recall that, typically, $N \gg n$. This means that the first term in $\hat{\theta}^{+}$ should have a vanishing variance due to the magnitude of N. Therefore,

$$\text{Var}(\hat{\theta}^{+}) \approx \text{Var}\left(\frac{1}{n} \sum_{j=1}^{K} \sum_{i \in I_j} (f^{(j)}(X_i) - Y_i) \right).$$

As the sample mean, the remaining term is an average of n terms. However, when the models are accurate, i.e., $f^{(j)}(X_i) \approx Y_i$, we expect $\text{Var}(f^{(j)}(X_i) - Y_i) \ll \text{Var}(Y_i)$.

The closest alternative to the cross-prediction estimator is the prediction-powered estimator (12), that is, its straightforward adaptation to the setup without a pretrained model. As discussed earlier, prediction-powered inference relies on having a pretrained model f. We can reduce our setting to this case by introducing data splitting: We use the first $n_{\text{tr}} \leq n$ data points from the labeled dataset to train a model f and the rest of the labeled data to compute the prediction-powered estimator:

$$\hat{\theta}^{\text{PP}} = \frac{1}{N} \sum_{i=1}^{N} f(\tilde{X}_i) - \frac{1}{n - n_{\text{tr}}} \sum_{i=n_{\text{tr}}+1}^{n} (f(X_i) - Y_i). \quad [6]$$

The prediction-powered estimator is also unbiased: $\mathbb{E}[\hat{\theta}^{\text{PP}}] = \theta^*$. However, this strategy is potentially wasteful because, after f is trained, the training data are thrown away and not subsequently used for estimation. Cross-prediction uses the data more efficiently, by leveraging each data point for both training and estimation.

General Cross-Prediction. To introduce the cross-prediction estimator in full generality, recall that we are considering all estimands of the form Eq. **1**. As in the case of mean estimation, we split the labeled data into K folds and train a predictive model $f^{(j)}$ on all folds but fold $j \in [K]$. The proposed cross-prediction estimator is defined as

$$\hat{\theta}^{+} = \arg\min_{\theta} \; L^{+}(\theta), \text{ where}$$
$$L^{+}(\theta) := \frac{1}{KN} \sum_{j=1}^{K} \sum_{i=1}^{N} \tilde{\ell}_{\theta,i}^{f^{(j)}} - \frac{1}{n} \sum_{j=1}^{K} \sum_{i \in I_j} (\ell_{\theta,i}^{f^{(j)}} - \ell_{\theta,i}). \quad [7]$$

Here, we use the short-hand notation $\tilde{\ell}_{\theta,i}^{f^{(j)}} := \ell_\theta(\tilde{X}_i, f^{(j)}(\tilde{X}_i))$, $\ell_{\theta,i}^{f^{(j)}} := \ell_\theta(X_i, f^{(j)}(X_i))$, and $\ell_{\theta,i} := \ell_\theta(X_i, Y_i)$. The intuition remains the same as before: The first term is an empirical approximation of the population loss if we treated the predictions as true labels, and the second term aims to debias this heuristic. One can verify that the mean estimator in Eq. **4** is indeed a special case of the general estimator in Eq. **7**, by taking ℓ_θ to be the squared loss, as per Eq. **2**.

The cross-prediction estimator optimizes an unbiased objective, since $\mathbb{E}[L^{+}(\theta)] = L(\theta)$. This follows because $\mathbb{E}[\ell_\theta(\tilde{X}_{i'}, f^{(j)}(\tilde{X}_{i'}))] = \mathbb{E}[\ell_\theta(X_i, f^{(j)}(X_i))]$ for all $j \in [K], i \in I_j, i' \in [N]$, seeing that $i \in I_j$ is not used to train model $f^{(j)}$. Furthermore, by the same argument as before, we expect $L^{+}(\theta)$ to have a lower variance than the classical objective in Eq. **3** if N is large and the trained predictors are reasonably accurate. We note that $L^{+}(\theta)$ may not be a convex function in general, but solving for $\hat{\theta}^{+}$ is tractable in many cases of interest. For example, in the case of means and generalized linear models, $L^{+}(\theta)$ is convex.

The prediction-powered estimator is similar to the cross-prediction estimator, but it requires data splitting and does not average over multiple model fits. It is equal to

$$\hat{\theta}^{\text{PP}} = \arg\min_{\theta} \; L^{\text{PP}}(\theta), \text{ where}$$
$$L^{\text{PP}}(\theta) := \frac{1}{N} \sum_{i=1}^{N} \tilde{\ell}_{\theta,i}^{f} - \frac{1}{n - n_{\text{tr}}} \sum_{i=n_{\text{tr}}+1}^{n} (\ell_{\theta,i}^{f} - \ell_{\theta,i}),$$

where, as before, f is trained on the first n_{tr} labeled data points. The fact that cross-prediction averages the results of multiple model fits allows it to achieve more stable inference. Indeed, in our experiments, we will observe that cross-prediction is more stable than prediction-powered inference throughout.

*By removing at most $K - 1$ data points, the size of the labeled dataset can be made divisible by K. Since in our applications K will typically be equal to 10, this truncation has a negligible effect.

Inference for the Mean

We now discuss inference with the cross-prediction estimator. For simplicity, we first look at mean estimation, where $\theta^* = \mathbb{E}[Y]$. We will see that much of the discussion will carry over to general M-estimation problems.

Inference with the cross-prediction estimator in Eq. **4** is difficult because the terms being averaged are all dependent through the labeled data. In contrast, the classical estimator in Eq. **5** averages independent terms, allowing for confidence intervals based on the central limit theorem. The prediction-powered estimator in Eq. **6** is similarly amenable to inference based on the central limit theorem, seeing that all the terms are independent conditional on f. In this section, we show that, under a relatively mild regularity condition, the cross-prediction estimator likewise satisfies a central limit theorem. This will in turn immediately allow constructing CIs and hypothesis tests for θ^*.

The central limit theorem will require that, as the sample size grows, the predictions concentrate sufficiently rapidly around their expectation. Intuitively, one can think of the condition as requiring that the predictions are sufficiently stable. While the stability property is difficult to verify for complex black-box models, we empirically observe that inference based on the resulting central limit theorem nevertheless provides the correct coverage. We observe this across different estimation problems, data modalities, sample sizes, and so on.

Our analysis based on stability is inspired by the work of Bayle et al. (30), who study inference on the cross-validation test error, since the inferential challenges in cross-prediction are similar to those in cross-validation. The ultimate goals of the two analyses are, however, entirely different.

Below we state the stability condition. For all x, we define $\bar{f}(x) := \mathbb{E}[f^{(1)}(x)]$; the "average" model $\bar{f}$ is the predictor we would obtain if we could train many models on independent datasets of size $n - n/K$ and average out their predictions.

Assumption 1. *We say that the predictions are stable if, as $n \to \infty$,*

$$\sqrt{K \operatorname{Var}\left(f^{(1)}(X) - \bar{f}(X) \mid f^{(1)}\right)} \xrightarrow{L^1} 0.$$

Assumption 1 requires that the models $f^{(j)}$ converge to their "average" model $\bar{f}$, but there is no assumption that $\bar{f}$ is by any means well-specified. If the number of folds is fixed (e.g., $K = 10$), as we will typically assume, then Assumption 1 is satisfied if the variance of the difference between the learned predictions $f^{(1)}(X)$ and the average predictions $\bar{f}(X)$ vanishes at any rate, $\operatorname{Var}\left(f^{(1)}(X) - \bar{f}(X) \mid f^{(1)}\right) \xrightarrow{L^1} 0$. We expect that any reasonably stable learning algorithm $\mathcal{A}_{\text{train}}$ should satisfy Assumption 1 (intuitively, any algorithm not too sensitive to perturbing a single data point). Violations of the assumption might arise if the number of folds is allowed to grow, e.g., as in the case of leave-one-out cross-fitting, since then the variance has to tend to zero sufficiently rapidly.

Equipped with Assumption 1, we can now state the central limit theorem for cross-prediction.

Theorem 1 (Cross-prediction CLT for the mean). *Let θ^* be the mean outcome, $\theta^* = \mathbb{E}[Y]$. Suppose that the predictions are stable (Assumption 1). Further, assume that $\frac{n}{N}$ has a limit, and that $\bar{\sigma}^2 = \operatorname{Var}(\bar{f}(X))$ and $\bar{\sigma}^2_\Delta = \operatorname{Var}(\bar{f}(X) - Y)$ have a nonzero limit. Then,*

$$\frac{\sqrt{n}}{\sqrt{\frac{n}{N}\bar{\sigma}^2 + \bar{\sigma}^2_\Delta}} \left(\hat{\theta}^+ - \theta^*\right) \xrightarrow{d} \mathcal{N}(0,1).$$

With this, inference on θ^* is now straightforward as long as we can estimate the asymptotic variance consistently. We will discuss strategies for doing so later on.

Corollary 1 (Cross-prediction inference for the mean). *Let θ^* be the mean outcome, $\theta^* = \mathbb{E}[Y]$. Assume the conditions of Theorem 1, and suppose that we have estimators $\hat{\sigma}^2 \xrightarrow{p} \bar{\sigma}^2$ and $\hat{\sigma}^2_\Delta \xrightarrow{p} \bar{\sigma}^2_\Delta$. Let*

$$\mathcal{C}^+_\alpha = \left(\hat{\theta}^+ \pm z_{1-\alpha/2} \frac{\sqrt{\frac{n}{N}\hat{\sigma}^2 + \hat{\sigma}^2_\Delta}}{\sqrt{n}}\right).$$

Then, $\liminf_{n,N} \mathbb{P}\left(\theta^* \in \mathcal{C}^+_\alpha\right) \geq 1 - \alpha$.

Per standard notation, $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution. Corollary 1 follows by a direct application of Theorem 1, together with Slutsky's theorem.

Inference for General M-Estimation

We generalize the principle introduced in the last section to handle arbitrary M-estimation problems. Indeed, the results presented in this section will strictly subsume the previous results.

As in the case of the mean, we will require that the predictions are "stable" in an appropriate sense. Naturally, the notion of stability will depend on the loss function used to define the M-estimator.

Assumption 2. *With $\bar{f}(\cdot)$ as before, we say that the predictions are stable if for all θ, as $n \to \infty$,*

$$\sqrt{K \operatorname{Var}\left(\nabla \ell_\theta(X, f^{(1)}(X)) - \nabla \ell_\theta(X, \bar{f}(X)) \mid f^{(1)}\right)} \xrightarrow{L^1} 0.$$

Here, $\operatorname{Var}(\cdot \mid f^{(1)})$ denotes the covariance matrix conditional on $f^{(1)}$. Also, for vectors and matrices, by "$\xrightarrow{L^1} 0$" we mean convergence in mean to zero element-wise. Notice that by setting $\ell_\theta(y) = (\theta - y)^2$ to be the squared loss, Assumption 2 reduces to Assumption 1 in the case of mean estimation. As in the case of Assumption 1, Assumption 2 should be interpreted as a stability requirement on $\mathcal{A}_{\text{train}}$. Moreover, there is again no requirement of correct specification of $\bar{f}$.

We will provide two approaches to inference in this section; which one is more appropriate will depend on the inference problem at hand.

One approach will be based on the characterization of θ^* as a zero of the gradient of the expected loss, $\nabla L(\theta^*) = \mathbb{E}[\nabla \ell_{\theta^*}(X, Y)] = 0$, which follows by the convexity of the loss. In particular, we will construct a confidence set for θ^* by finding all θ accepted by a valid test for the null hypothesis that $\nabla L(\theta) = 0$. Since the test is valid and θ^* satisfies the null condition, the true solution θ^* will be excluded with small probability. The hypothesis test for the population gradient $\nabla L(\theta)$ will follow from a central limit theorem for the gradient of the cross-prediction loss,

$$\nabla L^+(\theta) = \frac{1}{KN} \sum_{j=1}^{K} \sum_{i=1}^{N} \nabla \tilde{\ell}^{f^{(j)}}_{\theta,i} - \frac{1}{n} \sum_{j=1}^{K} \sum_{i \in I_j} (\nabla \ell^{f^{(j)}}_{\theta,i} - \nabla \ell_{\theta,i}).$$

The other approach will be based on showing asymptotic normality of the cross-prediction estimator. For this, we build on the proof of asymptotic normality of the prediction-powered estimator (with a pretrained model) (25), which in turn builds on classical asymptotic normality of M-estimators (17). The asymptotic normality will allow forming standard CLT intervals around $\hat{\theta}^+$.

We implicitly assume mild regularity on the losses $\ell_\theta(x,y)$ and $\ell_\theta(x,f^{(j)}(x))$, in particular that they are differentiable and locally Lipschitz around θ^* for all possible $f^{(j)}$ (see definition A.1 in ref. 25). Our second inference approach will require the usual condition that $\hat{\theta}^+$ is consistent, $\hat{\theta}^+ \xrightarrow{p} \theta^*$; this is satisfied quite broadly, e.g., when the parameter space is compact or when $L^+(\theta)$ is convex. The latter holds for all generalized linear models, for example. See refs. 17 and 25 for further discussion.

Theorem 2 states the main technical result of this section, which extends Theorem 1 to general M-estimation problems.

Theorem 2 (Cross-prediction CLT). *Suppose that the predictions are stable (Assumption 2). Further, assume that $\frac{n}{N}$ has a limit, and that $\bar{\Sigma}_\theta = \mathrm{Var}(\nabla \ell^{\bar{f}}_{\theta,i})$ and $\bar{\Sigma}_{\Delta,\theta} = \mathrm{Var}(\nabla \ell^{\bar{f}}_{\theta,i} - \nabla \ell_{\theta,i})$ have a nonzero limit. Denote $\bar{V}_\theta = \frac{n}{N}\bar{\Sigma}_\theta + \bar{\Sigma}_{\Delta,\theta}$. Then,*

$$\sqrt{n}\bar{V}_\theta^{-1/2}\left(\nabla L^+(\theta) - \nabla L(\theta)\right) \xrightarrow{d} \mathcal{N}(0, I).$$

If, additionally, the Hessian $H_{\theta^*} = \nabla^2 L(\theta^*)$ is nonsingular, $\hat{\theta}^+ \xrightarrow{p} \theta^*$, and K is constant, then

$$\sqrt{n}\bar{\Sigma}^{-1/2}\left(\hat{\theta}^+ - \theta^*\right) \xrightarrow{d} \mathcal{N}(0, I),$$

where $\bar{\Sigma} = H_{\theta^*}^{-1}\bar{V}_{\theta^*}H_{\theta^*}^{-1}$.

Theorem 2 immediately yields two methods for computing a confidence set for θ^*, as stated below.

Corollary 2 (Cross-prediction inference). *Suppose that we have estimators $\hat{\Sigma} \xrightarrow{p} \bar{\Sigma}$ and $\hat{V}_\theta \xrightarrow{p} \bar{V}_\theta$, for all θ. Then, assuming the conditions of Theorem 2, for either*

$$\mathcal{C}_\alpha^+ = \left\{\theta : \left\|\hat{V}_\theta^{-1/2}\nabla L^+(\theta)\right\|^2 \le \frac{\chi^2_{d,1-\alpha}}{n}\right\} \text{ or}$$

$$\mathcal{C}_\alpha^+ = \left(\hat{\theta}_i^+ \pm z_{1-\alpha/(2d)}\sqrt{\frac{\hat{\Sigma}_{ii}}{n}}\right)_{i=1}^d,$$

we have $\liminf_{n,N} \mathbb{P}\left(\theta^* \in \mathcal{C}_\alpha^+\right) \ge 1-\alpha$.

Above, $\chi^2_{d,1-\alpha}$ is the $(1-\alpha)$-quantile of the chi-squared distribution with d degrees of freedom; when $d = 1$ (as in the case of mean estimation), $\chi_{d,1-\alpha}$ is equal to $z_{1-\alpha/2}$. Note also that in the case of mean estimation, the two confidence sets are identical and reduce to the set from Corollary 1. In the second confidence set we apply a Bonferroni correction over the d coordinates of the estimand for simplicity and clarity of exposition; we can obtain an asymptotically exact $(1-\alpha)$-confidence set as $\mathcal{C}_\alpha^+ = \left\{\hat{\theta}^+ + v : v^\top\hat{\Sigma}v \le \frac{\chi^2_{d,1-\alpha}}{n}\right\}$.

Next, we apply Theorem 2 and Corollary 2 to concrete problems—quantile estimation, linear regression, and generalized linear models—to get explicit CI constructions.

Example: Quantile Estimation. Assume we are interested in a quantile of Y, $\theta^* = \inf\left\{y : \mathbb{P}(Y \le y) \ge q\right\}$. The quantile can equivalently be written as any minimizer of the pinball loss,

$$\begin{aligned}\theta^* &= \arg\min_\theta \mathbb{E}[\ell_\theta(Y)] \\ &= \arg\min_\theta \mathbb{E}[q(Y-\theta)\mathbf{1}\{Y > \theta\} + (1-q)(\theta - Y)\mathbf{1}\{Y \le \theta\}].\end{aligned}$$

The subgradient of the pinball loss is equal to $\nabla\ell_\theta(y) = -q\mathbf{1}\{y > \theta\} + (1-q)\mathbf{1}\{y \le \theta\} = -q + \mathbf{1}\{y \le \theta\}$. Plugging this expression into the first confidence set from Corollary 2 yields

$$\mathcal{C}_\alpha^+ = \left\{\theta : \left|\tilde{F}^+(\theta) - \Delta^+(\theta) - q\right| \le z_{1-\alpha/2}\frac{\sqrt{\frac{n}{N}\hat{\sigma}^2_\theta + \hat{\sigma}^2_{\Delta,\theta}}}{\sqrt{n}}\right\},$$

where $\tilde{F}^+(\theta) = \frac{1}{KN}\sum_{j=1}^K\sum_{i=1}^N \mathbf{1}\{f^{(j)}(\tilde{X}_i) \le \theta\}$ is the average empirical CDF of the predictions on the unlabeled data, and $\Delta^+(\theta) = \frac{1}{n}\sum_{j=1}^K\sum_{i\in I_j}(\mathbf{1}\{f^{(j)}(X_i) \le \theta\} - \mathbf{1}\{Y_i \le \theta\})$ is the difference between the empirical CDFs of the predictions and true outcomes on the labeled data. The SEs are equal to $\bar{\sigma}^2_\theta = \mathrm{Var}(\mathbf{1}\{\bar{f}(X) \le \theta\})$ and $\bar{\sigma}^2_{\Delta,\theta} = \mathrm{Var}(\mathbf{1}\{\bar{f}(X) \le \theta\} - \mathbf{1}\{Y \le \theta\})$. The confidence set $\mathcal{C}_\alpha^+$ thus consists of all values θ such that the average predicted CDF $\tilde{F}^+(\theta)$, corrected by the bias $\Delta^+(\theta)$, is close to the target level q.

Example: Linear Regression. In linear regression, the target of inference is defined by

$$\theta^* = \arg\min_\theta \mathbb{E}[\ell_\theta(X,Y)] = \arg\min_\theta \frac{1}{2}\mathbb{E}[(Y - X^\top\theta)^2]. \quad \textbf{[8]}$$

In this case, the cross-prediction estimator, equal to the solution to $\nabla L^+(\hat{\theta}^+) = 0$, has a closed-form expression. Letting $\tilde{\mathbb{X}} \in \mathbb{R}^{N\times d}$ (resp. $\mathbb{X} \in \mathbb{R}^{n\times d}$) be the unlabeled (resp. labeled) data matrix, $\mathbb{Y} \in \mathbb{R}^n$ be the vector of labeled outcomes, the solution is given by

$$\hat{\theta}^+ = (\tilde{\mathbb{X}}^\top\tilde{\mathbb{X}})^{-1}\left(\tilde{\mathbb{X}}^\top f_{\mathrm{avg(K)}}(\tilde{\mathbb{X}}) - \frac{N}{n}\cdot\mathbb{X}^\top\left(f_{1:K}(\mathbb{X}) - \mathbb{Y}\right)\right),$$

where $f_{\mathrm{avg(K)}}(\tilde{\mathbb{X}}) = \frac{1}{K}\sum_{j=1}^K f^{(j)}(\tilde{\mathbb{X}})$ is the vector of average predictions on the unlabeled data, and $f_{1:K}(\mathbb{X})$ is the vector of predictions on the labeled data: $f_{1:K}(\mathbb{X}) = (f^{(1)}(X_1),\ldots,f^{(1)}(X_{\frac{n}{K}}),f^{(2)}(X_{\frac{n}{K}+1}),\ldots,f^{(K)}(X_n))$. We see that $\hat{\theta}^+$ resembles the usual least-squares estimator with $f_{\mathrm{avg(K)}}(\tilde{\mathbb{X}})$ as the response, except for the extra debiasing factor, $\frac{N}{n}\cdot\mathbb{X}^\top\left(f_{1:K}(\mathbb{X}) - \mathbb{Y}\right)$, that takes into account the prediction inaccuracies.

Instantiating the relevant terms, Theorem 2 implies that $\hat{\theta}^+$ is asymptotically normal with covariance equal to $\Sigma_{\mathrm{OLS}} = H^{-1}\bar{V}_{\theta^*}H^{-1}$, where $H = \mathbb{E}[XX^\top]$ and $\bar{V}_{\theta^*} = \frac{n}{N}\bar{\Sigma}_{\theta^*} + \bar{\Sigma}_\Delta$, for $\bar{\Sigma}_{\theta^*} = \mathrm{Var}((\bar{f}(X) - X^\top\theta^*)X)$ and $\bar{\Sigma}_\Delta = \mathrm{Var}((\bar{f}(X) - Y)X)$.

For a given coordinate of interest i, a CI for θ_i^* can therefore be obtained as

$$\mathcal{C}_\alpha^+ = \left(\hat{\theta}_i^+ \pm z_{1-\alpha/2}\frac{\sqrt{(\hat{\Sigma}_{\mathrm{OLS}})_{ii}}}{\sqrt{n}}\right),$$

given an estimate $\hat{\Sigma}_{\mathrm{OLS}}$ of Σ_{OLS}.

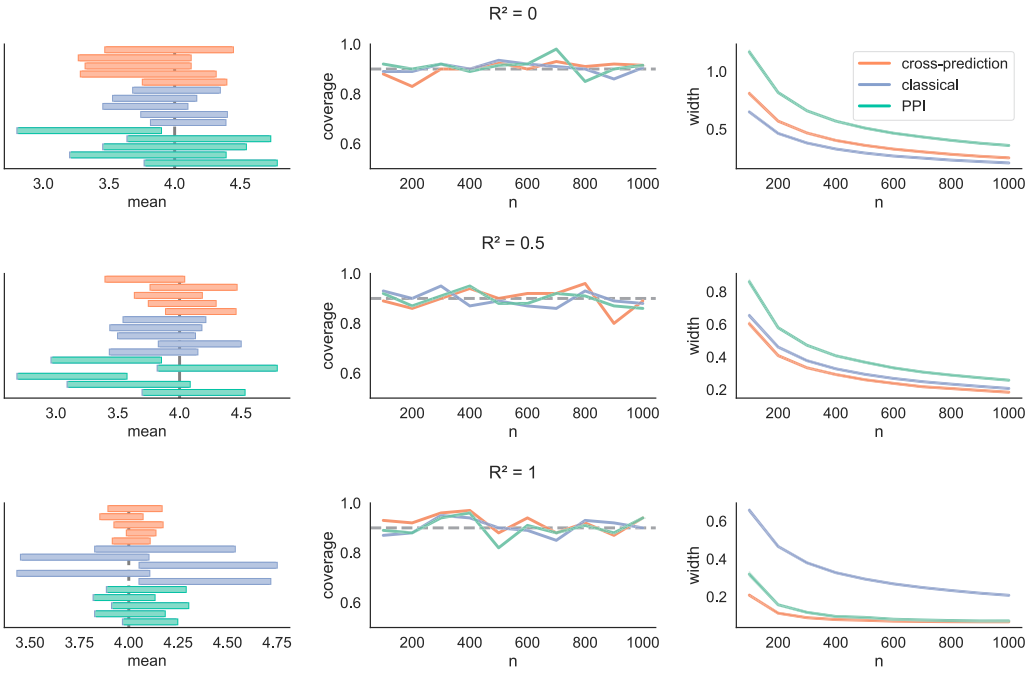

Fig. 3. Mean estimation. Intervals from five randomly chosen trials (*Left*), coverage (*Middle*), and average interval width (*Right*) of cross-prediction, classical inference, and prediction-powered inference (PPI) in a mean estimation problem.

Example: Generalized Linear Models. We can generalize the previous example by considering all generalized linear models (GLMs). In particular, we consider targets of inference given by

$$\begin{aligned}\theta^* &= \arg\min_\theta \mathbb{E}[-\log p_\theta(Y|X)] \\ &= \arg\min_\theta \mathbb{E}[-Y\theta^\top X + \psi(X^\top\theta)], \end{aligned} \quad [9]$$

where $p_\theta(y|x) = \exp(yx^\top\theta - \psi(x^\top\theta))$ is the probability density of the outcome given the features and the log-partition function ψ is convex. The objective Eq. **9** recovers the linear-regression objective Eq. **8** by setting $\psi(s) = \frac{1}{2}s^2$. It captures logistic regression by choosing $\psi(s) = \log(1+e^s)$.

The asymptotic covariance given by Theorem 2 evaluates to $\Sigma_{\text{GLM}} = H_{\theta^*}^{-1}\bar{V}_{\theta^*}H_{\theta^*}^{-1}$, $H_{\theta^*} = \mathbb{E}[\psi''(X^\top\theta^*)XX^\top]$, $\bar{V}_{\theta^*} = \frac{n}{N}\text{Var}((\psi'(X^\top\theta^*) - \bar{f}(X))X) + \text{Var}((\bar{f}(X) - Y)X)$. Therefore, analogously to the OLS case, given an estimate $\hat{\Sigma}_{\text{GLM}}$ of Σ_{GLM}, we can construct a CI for $\hat{\theta}^+$ as

$$\mathcal{C}_\alpha^+ = \left(\hat{\theta}_i^+ \pm z_{1-\alpha/2}\frac{\sqrt{(\hat{\Sigma}_{\text{GLM}})_{ii}}}{\sqrt{n}}\right).$$

Variance Estimation via Bootstrapping

The previous inference results rely on being able to estimate the asymptotic covariance of $\hat{\theta}^+$. We herewith provide an explicit estimation strategy that we will use in our experiments.

Recall that the asymptotic covariance is equal to $\bar{\Sigma} = H_{\theta^*}^{-1}\bar{V}_{\theta^*}H_{\theta^*}^{-1}$, where $\bar{V}_\theta = \frac{n}{N}\bar{\Sigma}_\theta + \bar{\Sigma}_{\Delta,\theta}$, for $\bar{\Sigma}_\theta = \text{Var}(\nabla\ell_{\theta,i}^{\bar{f}})$ and $\bar{\Sigma}_{\Delta,\theta} = \text{Var}(\nabla\ell_{\theta,i}^{\bar{f}} - \nabla\ell_{\theta,i})$. Estimating the Hessian H_θ is easy via plug-in estimation; $\bar{V}_\theta$, on the other hand, depends on the average model $\bar{f}$. If the average model $\bar{f}$ was known, one could compute estimates of $\bar{\Sigma}_\theta$ and $\bar{\Sigma}_{\Delta,\theta}$ by replacing the true covariances with their empirical counterparts. Thus, the challenge is to approximate $\bar{f}$. To achieve this, we apply the bootstrap to simulate multiple model training runs, and at the end, we average the predictions of all the learned models.

In more detail, for each $b \in \{1, 2, \ldots, B\} = [B]$, we sample $n - \frac{n}{K}$ data points uniformly at random from the labeled dataset, and denote the indices of the samples by I_{boot}^b. Then, we use the sampled data points to train a model $f_{\text{boot}}^{(b)}$ using the same model-fitting strategy as for the cross-prediction models $f^{(j)}$. To estimate $\bar{\Sigma}_\theta$, we compute

$$\hat{\Sigma}_\theta = \widehat{\text{Var}}\left(\nabla\ell_\theta(\tilde{X}_i, \bar{f}_{\text{boot}}(\tilde{X}_i)), i \in [N]\right),$$

where $\bar{f}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^B f_{\text{boot}}^{(b)}$ and $\widehat{\text{Var}}$ denotes the empirical covariance. To estimate $\bar{\Sigma}_{\Delta,\theta}$, we compute

Table 1. SD of the lower ($\hat{\sigma}_L$) and upper ($\hat{\sigma}_U$) endpoints of the CIs in the mean estimation problem from Fig. 3, for $n = 100$

	Mean estimation	
	$R^2 = 0$	
Method	$\hat{\sigma}_L$	$\hat{\sigma}_U$
Cross-prediction	0.2694	0.2696
Classical	**0.2124**	**0.2085**
PPI	0.3844	0.3997
	$R^2 = 0.5$	
Method	$\hat{\sigma}_L$	$\hat{\sigma}_U$
Cross-prediction	**0.1769**	0.1897
Classical	0.1908	**0.1885**
PPI	0.2751	0.2684
	$R^2 = 1$	
Method	$\hat{\sigma}_L$	$\hat{\sigma}_U$
Cross-prediction	**0.0591**	**0.0613**
Classical	0.2136	0.2102
PPI	0.1045	0.1061

The minimum value in each column is in bold.

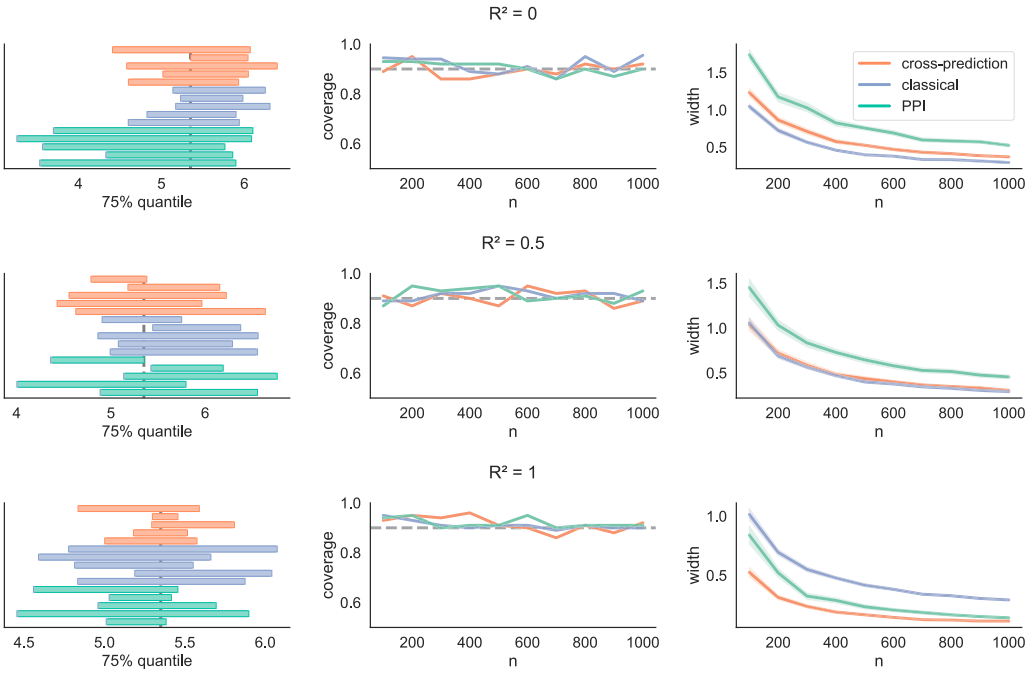

Fig. 4. Quantile estimation. Intervals from five randomly chosen trials (*Left*), coverage (*Middle*), and average interval width (*Right*) of cross-prediction, classical inference, and prediction-powered inference (PPI) in a quantile estimation problem. The target is the 75th percentile.

$$\hat{\Sigma}_{\Delta,\theta} = \widehat{\mathrm{Var}}\left(\left(\nabla \ell_\theta(X_i, f^{(b)}_{\mathrm{boot}}(X_i)) - \nabla \ell_\theta(X_i, Y_i)\right)_{i\in[n]\setminus I^{(b)}_{\mathrm{boot}}, b\in[B]}\right).$$

Finally, we approximate the covariance by $\frac{n}{N}\hat{\Sigma}_\theta + \hat{\Sigma}_{\Delta,\theta}$. In computing $\hat{\Sigma}_{\Delta,\theta}$, we technically do not average out the bootstrapped models because we want to make sure that every point (X_i, Y_i) used to compute the gradient bias is independent of its corresponding model $f^{(b)}_{\mathrm{boot}}$. Intuitively, as per the discussion following Assumption 1, if $\mathcal{A}_{\mathrm{train}}$ is stable we expect $f^{(b)}_{\mathrm{boot}}$ to be a good approximation of the average model $\bar{f}$, which in turn means that the bootstrap covariance estimates should be consistent per conventional wisdom. We show empirically that the covariance estimates lead to valid coverage across a range of applications.

To give one concrete example, consider mean estimation: $\theta^* = \mathbb{E}[Y]$. We compute

$$\hat{\sigma}^2 = \widehat{\mathrm{Var}}\left(\bar{f}_{\mathrm{boot}}(\tilde{X}_i), i \in [N]\right) \text{ and}$$
$$\hat{\sigma}^2_\Delta = \widehat{\mathrm{Var}}\left(f^{(b)}_{\mathrm{boot}}(X_i) - Y_i, i \in [n] \setminus I^{(b)}_{\mathrm{boot}}, b \in [B]\right),$$

and take $\mathcal{C}^+_\alpha = \left(\hat{\theta}^+ \pm z_{1-\alpha/2}\frac{\sqrt{\frac{n}{N}\hat{\sigma}^2 + \hat{\sigma}^2_\Delta}}{\sqrt{n}}\right)$ as the final interval.

Experiments

We evaluate cross-prediction and compare it to baseline approaches on several datasets; the baselines are the classical inference method, which only uses the labeled data, and prediction-powered inference with a data-splitting step in order to train a predictive model. Code for reproducing the experiments is available at: https://github.com/tijana-zrnic/cross-ppi (56).

For each experimental setting, we plot the coverage and CI width estimated over 100 trials for all baselines. We also show the constructed CIs for five randomly chosen trials. Finally, to quantify the stability of inferences, we report the SD of the lower and upper endpoints of the confidence intervals for each method.

We begin with proof-of-concept experiments on synthetic data. Then, we move on to more complex real datasets.

Proof-of-Concept Experiments on Synthetic Data. To build intuition, we begin with simple experiments on synthetic data. The purpose is to confirm what we expect in theory: a) as it gets easier to predict labels from features, cross-prediction, and prediction-powered inference become more powerful and increasingly outperform the classical approach; b) cross-prediction uses the data more efficiently than prediction-powered inference, yielding smaller intervals; c) cross-prediction gives more stable inferences than the baselines when the predictions are useful; d) all three approaches lead to satisfactory coverage.

In all of the following experiments, we have $N = 10{,}000$ unlabeled data points, and we vary the size of the labeled data

Table 2. SD of the lower ($\hat{\sigma}_L$) and upper ($\hat{\sigma}_U$) endpoints of the CIs in the quantile estimation problem from Fig. 4, for $n = 100$

	Quantile estimation $R^2 = 0$	
Method	$\hat{\sigma}_L$	$\hat{\sigma}_U$
Cross-prediction	0.4102	0.3509
Classical	**0.2302**	**0.3024**
PPI	0.5424	0.4614
	$R^2 = 0.5$	
Method	$\hat{\sigma}_L$	$\hat{\sigma}_U$
Cross-prediction	0.3253	**0.3242**
Classical	**0.2569**	0.3305
PPI	0.4141	0.4368
	$R^2 = 1$	
Method	$\hat{\sigma}_L$	$\hat{\sigma}_U$
Cross-prediction	**0.1345**	**0.1545**
Classical	0.2615	0.2806
PPI	0.2151	0.3280

The minimum value in each column is in bold.

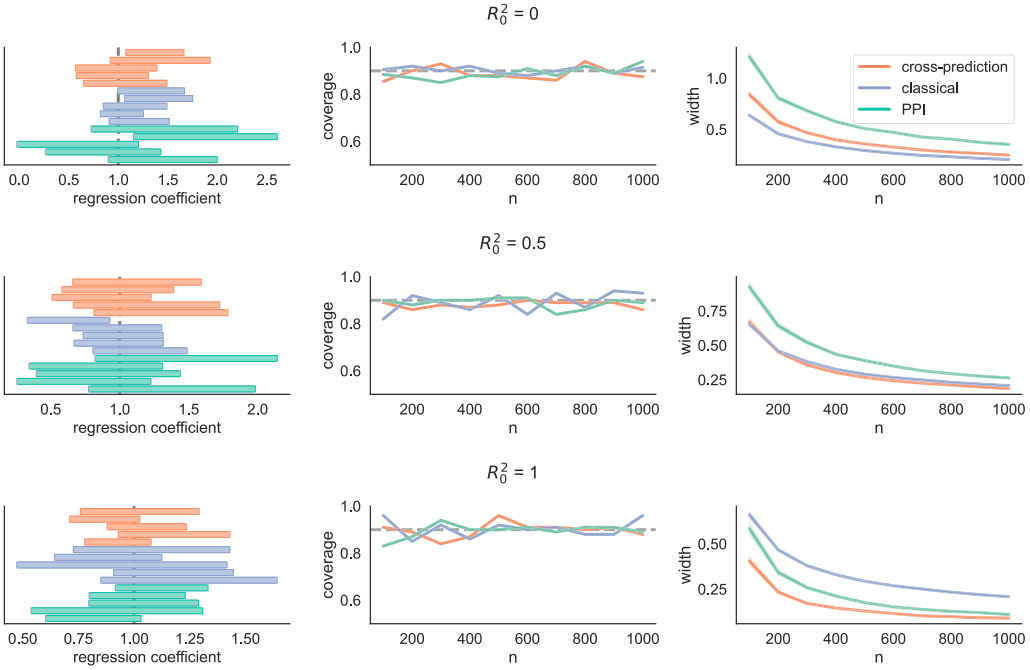

Fig. 5. Linear regression. Intervals from five randomly chosen trials (*Left*), coverage (*Middle*), and average interval width (*Right*) of cross-prediction, classical inference, and prediction-powered inference (PPI) in a linear regression problem.

n between 100 and 1,000, in 100-point increments. We apply cross-prediction with $K = 10$ folds. We estimate the variance using the bootstrap approach described in the last section, with $B = 30$ bootstrap samples. For prediction-powered inference, we use half of the labeled data for model training. To illustrate the point that cross-prediction can be applied with any black-box model, we train gradient-boosted trees via XGBoost (57) to obtain the models $f^{(j)}$. We use the same model-fitting strategy for prediction-powered inference. We fix the target error level to be $\alpha = 0.1$ and average the results over 100 trials.

Mean estimation. For given parameters R^2 and σ_Y^2, the data-generating distribution is defined as $X \sim \mathcal{N}(0, I_2)$, $Y = \mu + X^\top\beta + \xi$, where $\beta_1 = \beta_2 = R\sigma_Y/\sqrt{2}$, and $\xi \sim \mathcal{N}(0, \sigma_Y^2(1 - R^2))$ is independent of X. We fix $\mu = 4, \sigma_Y^2 = 4$ and vary $R^2 = \frac{\mathrm{Var}(X^\top\beta)}{\mathrm{Var}(Y)} \in \{0, 0.5, 1\}$. The idea is to vary the degree to which the outcomes can be explained through the features: When $R^2 = 0$, the outcome is independent of the features and we do not expect predictions to help, while when $R^2 = 1$, the outcome can be perfectly explained through the features and we expect predictions to be helpful. Given that the variance of Y is kept constant regardless of R^2, classical inference has the same distribution of widths across R^2. The target of inference is $\theta^* = \mathbb{E}[Y] = \mu$.

In Fig. 3 we plot the coverage and interval width of cross-prediction, classical inference, and prediction-powered inference, as well as five example intervals. All three methods approximately achieve the target coverage, and cross-prediction dominates prediction-powered inference throughout. Further, we see that the classical approach dominates the alternatives when the features are independent of the outcomes, while the alternatives become more powerful as R^2 increases. To evaluate stability, in Table 1 we report the SD of the lower and upper endpoints of

Table 3. SD of the lower ($\hat{\sigma}_L$) and upper ($\hat{\sigma}_U$) endpoints of the CIs in the linear regression problem from Fig. 5, for $n = 100$

	Linear regression $\mathbf{R_0^2 = 0}$	
Method	$\hat{\sigma}_L$	$\hat{\sigma}_U$
Cross-prediction	0.2801	0.2969
Classical	**0.2091**	**0.2098**
PPI	0.4104	0.4870
	$\mathbf{R_0^2 = 0.5}$	
Method	$\hat{\sigma}_L$	$\hat{\sigma}_U$
Cross-prediction	**0.1875**	**0.2250**
Classical	0.2271	0.2262
PPI	0.2602	0.3326
	$\mathbf{R_0^2 = 1}$	
Method	$\hat{\sigma}_L$	$\hat{\sigma}_U$
Cross-prediction	**0.1102**	**0.1472**
Classical	0.1800	0.1809
PPI	0.1522	0.2530

The minimum value in each column is in bold.

Table 4. SD of the lower ($\hat{\sigma}_L$) and upper ($\hat{\sigma}_U$) endpoints of the CIs in the real-data problems

	Deforestation analysis	
Method	$\hat{\sigma}_L$	$\hat{\sigma}_U$
Cross-prediction	**0.0158**	**0.0182**
Classical	0.0195	0.0232
PPI	0.0200	0.0240
	ACS survey analysis	
Method	$\hat{\sigma}_L$	$\hat{\sigma}_U$
Cross-prediction	**11.2781**	**12.2367**
Classical	14.5346	15.6106
PPI	13.1378	13.8733
	Galaxy analysis	
Method	$\hat{\sigma}_L$	$\hat{\sigma}_U$
Cross-prediction	**0.0029**	**0.0029**
Classical	0.0037	0.0038
PPI	0.0036	0.0037

For each problem, we take $\boldsymbol{n}$ to be the smallest labeled dataset size in the considered range. The minimum value in each column is in bold.

Variable	Label
AGEP	Age
ANC	Ancestry recode
DRIVESP	Number of vehicles calculated from JWRI
FES	Family type and employment status
FPARC	Family presence and age of related children
GRPIP	Gross rent as a percentage of household income past 12 mon...
HISP	Recoded detailed Hispanic origin
JWAP	Time of arrival at work - hour and minute
JWDP	Time of departure for work - hour and minute
JWMNP	Travel time to work
JWRIP	Vehicle occupancy
MV	When moved into this house or apartment

Fig. 6. Subset of the variables available in the ACS PUMS data.

the confidence intervals from Fig. 3, for $n = 100$. We observe that the classical approach is the most stable method when $R^2 = 0$, which makes sense because the predictions can only introduce noise. When $R^2 = 0.5$, cross-prediction and classical inference have a similar degree of stability, while when $R^2 = 1$ cross-prediction is significantly more stable. Moreover, cross-prediction is significantly more stable than prediction-powered inference throughout. These trends hold across different values of n; however, we only include the results for $n = 100$ for simplicity of exposition.

Quantile estimation. We adopt the same data-generating process as for mean estimation. We only change the target of inference θ^* to be the 75th percentile of the outcome distribution.

In Fig. 4 we plot the coverage and interval width of cross-prediction, classical inference, and prediction-powered inference, as well as five example intervals. We observe a qualitatively similar comparison as in the case of mean estimation: All three methods approximately achieve the target coverage, and cross-prediction dominates prediction-powered inference throughout. As before, the classical approach dominates the alternatives when the features are independent of the outcomes, and the alternatives become increasingly powerful as R^2 increases. In Table 2 we evaluate the stability of the methods by reporting the SD of the lower and upper endpoints of the confidence intervals from Fig. 4, for $n = 100$. As before, Table 2 shows that cross-prediction is more stable than prediction-powered inference for all values of R^2, and when $R^2 = 0$ classical inference is the most stable option. When $R^2 = 0.5$, cross-prediction has a slightly more stable upper endpoint than classical inference, while classical inference has a more stable lower endpoint. When $R^2 = 1$, cross-prediction is by far the most stable method. For $R^2 \in \{0, 0.5\}$. Again, these trends are largely consistent across different values of n; however, we only include the results for $n = 100$ for simplicity.

Linear regression. Finally, we look at linear regression. For robustness and interpretability, it is common to include only a subset of the available features in the regression. The process of deciding which variables to include is known as model selection. The variables that are not included in the model may still be predictive of the outcome of interest; we demonstrate that, as such, they can be useful for inference.

The data-generating distribution is defined as follows: We generate $X \sim \mathcal{N}(0, I_3)$, $Y = X^\top \beta + \xi$, where $\beta = (1, 1, R_0\sigma_Y)$ and $\xi \sim \mathcal{N}(0, \sigma_Y^2(1 - R_0^2))$. Again, the idea is to vary how much of the outcome can be explained through prediction versus how much of it is exogenous randomness. We fix $\sigma_Y^2 = 4$ and vary $R_0 \in \{0, 0.5, 1\}$. The target of inference is defined as the least-squares solution when regressing Y on (X_1, X_2), that is, the first coordinate of this solution. In this case, this is simply equal to $\theta^* = \beta_1 = 1$.

In Fig. 5 we plot the coverage and interval width of cross-prediction, classical inference, and prediction-powered inference. When $R_0^2 = 0$, the classical approach outperforms the prediction-based approaches; as R_0^2 grows, meaning more of the randomness of the outcome can be attributed to X_3, the prediction-based approaches dominate the classical one. As before, cross-prediction yields smaller intervals than prediction-powered inference. We remark that, even though the inference problem posits a linear model, the prediction-based approaches still use XGBoost for model training. Like in the previous two examples, we report on the stability of the three methods in Table 3. We again fix $n = 100$ for simplicity. Cross-prediction is far more stable than prediction-powered inference throughout, and it is more stable than classical inference for nonzero values of R_0^2.

Estimating Deforestation from Satellite Imagery. We briefly revisit the problem of deforestation analysis from Fig. 2. As we saw in the figure, cross-prediction gave tighter CIs for the deforestation rate than using gold-standard measurements of deforestation alone. In other words, cross-prediction can enable a reduction in the number of necessary field visits to measure deforestation. Moreover, we saw that cross-prediction outperformed prediction-powered inference.

Here we argue another benefit of cross-prediction in this problem: It is a more stable solution than the baselines. Table 4 shows the SD of the endpoints of the confidence intervals constructed by cross-prediction and its competitors. Cross-prediction has a significantly lower variability of the endpoints than both classical inference and prediction-powered inference, while the latter two exhibit similar variability.

Finally, we provide the experimental details behind Fig. 2. We have $n_{\text{all}} = 3{,}192$ data points with gold-standard labels total. To simulate having unlabeled images, in each trial we randomly split the data into n points to serve as the labeled data, for varying $n \in \{0.1n_{\text{all}}, 0.2n_{\text{all}}, 0.3n_{\text{all}}\}$, and treat the remaining points as unlabeled. The target of inference is the fraction of deforested areas across the locations contained in the sample. We apply cross-prediction with $K = 10$ folds. For prediction-powered inference, we use $n_{\text{tr}} = 0.1n$ points for model tuning. We average the results over 100 trials.

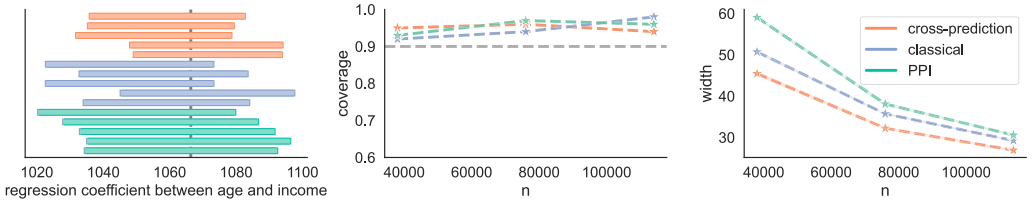

Fig. 7. Estimating the relationship between age, sex, and income in ACS data. Intervals from five randomly chosen trials (*Left*), coverage (*Middle*), and average interval width (*Right*) of cross-prediction, classical inference, and prediction-powered inference (PPI) in a linear regression problem on ACS PUMS data. The target θ^* is the linear regression coefficient when regressing income on age and sex.

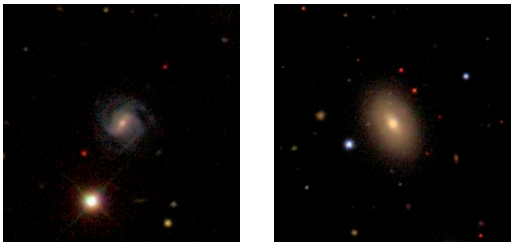

Fig. 8. Example images of a spiral galaxy (*Left*) and a nonspiral galaxy (*Right*).

Estimating Relationships between Socioeconomic Covariates in Survey Data. We evaluate cross-prediction on the American Community Survey (ACS) Public Use Microdata Sample (PUMS). We investigate the relationship between age, sex, and income in survey data collected in California in 2019 ($n_{\text{all}} = 377,575$ people total). High-quality survey data are generally difficult and time-consuming to collect. With this experiment, we hope to demonstrate how, by imputing missing information based on the available covariates, cross-prediction can achieve both powerful and valid inferences while reducing the requisite amount of survey data. See Fig. 6 for a subset of the available covariates in the ACS PUMS data.

We use the Folktables (58) interface to download the PUMS data, including income, age, sex, and 15 other demographic covariates. In each trial, we randomly sample n data points to serve as the labeled data, for varying n, and treat the remaining data points as the unlabeled data. We vary $n \in \{0.1n_{\text{all}}, 0.2n_{\text{all}}, 0.3n_{\text{all}}\}$. The target of inference is the linear regression coefficient when regressing income on age and sex: $\theta^* = \arg\min_\theta \mathbb{E}[(Y - X_{\text{ols}}^\top \theta)^2]$, where Y is income and X_{ols} encodes age and sex, $X_{\text{ols}} = (X_{\text{age}}, X_{\text{sex}})$. For the purpose of evaluating coverage, the corresponding coefficient computed on the whole dataset is taken as the ground-truth value of the estimand. To obtain the models $f^{(j)}$, we train gradient-boosted trees via XGBoost (57). Note that the predictors use all 17 covariates to predict the missing labels, even though the target of inference is only defined with respect to two covariates. We apply cross-prediction with $K = 5$ folds. For prediction-powered inference, we use $n_{\text{tr}} = 0.2n$ points for model training, and we also train gradient-boosted trees. The target error level is $\alpha = 0.1$ and we average the results over 100 trials.

In Fig. 7 we plot the coverage and interval width for the three baselines, together with five example intervals. All three methods cover the true target with the desired probability. Moreover, as before, cross-prediction outperforms prediction-powered inference. In this example, the predictive power of the trained models is not high enough for prediction-powered inference to outperform the classical approach; cross-prediction, however, outperforms both. In Table 4, we report on the stability of the three methods for $n = 0.1n_{\text{all}}$. We observe that cross-prediction is more stable than both alternatives. We also observe that prediction-powered inference has more stable intervals than the classical approach, despite the fact that they are wider on average.

Estimating the Prevalence of Spiral Galaxies from Galaxy Images. We next look at galaxy data from the Galaxy Zoo 2 dataset (59), consisting of human-annotated images of galaxies from the Sloan Digital Sky Survey (60). Of particular interest are galaxies with spiral arms, which are correlated with star formation in the discs of low-redshift galaxies, and thus contribute to the understanding of star formation. See Fig. 8 for example images of a spiral and a nonspiral galaxy. We show that, by leveraging the massive amounts of unlabeled galaxy imagery together with machine learning, cross-prediction can decrease the requisite number of human annotations for inference on galaxy demographics.

We have 167,434 annotated galaxy images. In each trial, we randomly split them up into n points to serve as the labeled data, for $n \in \{10{,}000, 20{,}000, 30{,}000\}$, and treat the remaining data points as unlabeled. The target of inference is the fraction of spiral galaxies in the dataset, equal to about 26.22%. To compute predictions, we fine-tune all layers of a pretrained ResNet50. We apply cross-prediction with $K = 3$ folds. For prediction-powered inference, we use $n_{\text{tr}} = 0.1n$ points for model training. The target error rate is $\alpha = 0.1$ and we average the results over 100 trials.

In Fig. 9 we plot the coverage and interval width of the three methods, as well as the intervals for five randomly chosen trials. Both cross-prediction and prediction-powered inference yield smaller intervals than the classical approach. Moreover, cross-prediction dominates prediction-powered inference. We observe satisfactory coverage for all three procedures. In Table 4 we evaluate the stability of the procedures for $n = 10,000$. Cross-prediction is significantly more stable than classical inference and prediction-powered inference. The latter two achieve a similar degree of stability.

Evaluating Heuristics

In Fig. 2, we saw that cross-prediction gave tighter CIs than the baseline approaches for the problem of deforestation analysis. In this section, we test two heuristic ways of reducing the variance of the classical approach and prediction-powered inference and compare the heuristics to cross-prediction.

The first heuristic removes the debiasing from the cross-prediction estimator and simply averages the predictions on the large unlabeled dataset:

$$\hat{\theta}_{\text{nodebias}} = \frac{1}{KN}\sum_{j=1}^{K}\sum_{i=1}^{N} f^{(j)}(\tilde{X}_i). \qquad [10]$$

This is akin to computing the classical estimator while pretending that the predicted labels are the ground truth. The second heuristic trains a model on all the labeled data, $f^{\text{all}} = \mathcal{A}_{\text{train}}(\{(X_i, Y_i)\}_{i=1}^n)$, and computes

$$\hat{\theta}_{\text{nofolds}} = \frac{1}{N}\sum_{i=1}^{N} f^{\text{all}}(\tilde{X}_i) - \frac{1}{n}\sum_{i=1}^{n}(f^{\text{all}}(X_i) - Y_i).$$

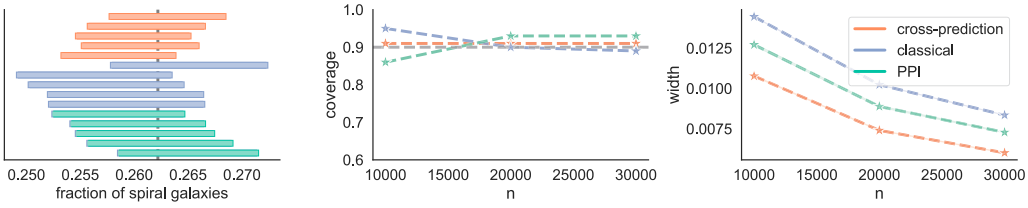

Fig. 9. Estimating the prevalence of spiral galaxies from galaxy images. Intervals from five randomly chosen trials (*Left*), coverage (*Middle*), and average interval width (*Right*) of cross-prediction, classical inference, and prediction-powered inference (PPI) in a mean estimation problem on galaxy image data. The target θ^* is the fraction of spiral galaxies.

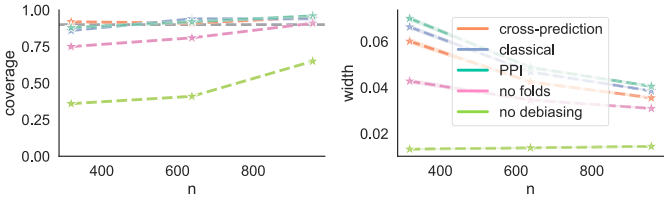

Fig. 10. Estimating the deforestation rate in the Amazon from satellite imagery (revisited). Coverage and average interval width of cross-prediction, classical inference, and prediction-powered inference (PPI), as well as two heuristics related to cross-fitting: one that removes the debiasing and one that trains on all labeled data instead of forming folds. The experimental setup is the same as in Fig. 2.

This estimator is akin to the prediction-powered estimator if we treated f^{all} as fixed and independent of the labeled dataset.

For both heuristics, we form confidence intervals based on the usual central limit theorem that assumes i.i.d. sampling. For the first heuristic this is done conditional on the trained models $f^{(j)}$, since the terms $(\frac{1}{K}\sum_{j=1}^{K} f^{(j)}(\tilde{X}_i))_{i\in[N]}$ are indeed conditionally independent given $f^{(1)}, \ldots, f^{(K)}$. Since the second heuristic proceeds under the assumption that f^{all} can be seen as being independent of the labeled data, we apply the central limit theorem to the two sums separately, as if f^{all} were fixed.

We see in Fig. 10 that removing the debiasing is detrimental to coverage; removing the folds has a more moderate effect that vanishes with n, but it is nevertheless significant. Cross-prediction yields wider intervals than both heuristics, and by doing so it maintains correct coverage.

Data, Materials, and Software Availability. Code and data for reproducing the experiments have been deposited in cross-ppi GitHub repository (56).

ACKNOWLEDGMENTS. We thank Anastasios Angelopoulos, Ying Jin, and Asher Spector for helpful suggestions and feedback on a draft of this manuscript, and Aditya Ghosh for catching and fixing a typo in a previous version of the manuscript. T.Z. was supported by Stanford Data Science through the Fellowship program. E.J.C. was supported by the Office of Naval Research grant N00014-20-1-2157, the NSF grant DMS-2032014, the Simons Foundation under award 814641, and the ARO grant 2003514594.

1. J. Jumper *et al.*, Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
2. K. Tunyasuvunakool *et al.*, Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
3. I. Bludau *et al.*, The structural context of posttranslational modifications at a proteome-wide scale. *PLoS Biol.* **20**, e3001636 (2022).
4. Z. Lin *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
5. Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes, O. M. Yaghi, ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *J. Am. Chem. Soc.* **145**, 18048–18062 (2023).
6. N. Jean *et al.*, Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
7. J. E. Steele *et al.*, Mapping poverty using mobile phone and satellite data. *J. R. Soc. Interface* **14**, 20160690 (2017).
8. J. E. Ball, D. T. Anderson, C. S. Chan, Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community. *J. Appl. Remote Sensing* **11**, 042609 (2017).
9. E. Rolf *et al.*, A generalizable and accessible approach to machine learning with global satellite imagery. *Nat. Commun.* **12**, 4392 (2021).
10. M. C. Hansen *et al.*, High-resolution global maps of 21st-century forest cover change. *Science* **342**, 850–853 (2013).
11. C. Robinson, F. Hohman, B. Dilkina, "A deep learning approach for population estimation from satellite imagery" in *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities* (2017), pp. 47–54.
12. A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, T. Zrnic, Prediction-powered inference. *Science* **382**, 669–674 (2023).
13. E. L. Bullock, C. E. Woodcock, C. Souza Jr, P. Olofsson, Satellite-based estimates reveal widespread forest degradation in the Amazon. *Global Chang. Biol.* **26**, 2956–2969 (2020).
14. J. O. Sexton *et al.*, Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of modis vegetation continuous fields with lidar-based estimates of error. *Int. J. Digital Earth* **6**, 427–448 (2013).
15. A. Zhang, L. D. Brown, T. T. Cai, Semi-supervised inference: General theory and estimation of means. *Ann. Stat.* **47**, 2538–2566 (2019).
16. E. L. Lehmann, J. P. Romano, *Testing Statistical Hypotheses* (Springer Nature, 2022), vol. 4.
17. A. W. van der Vaart, *Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics* (Cambridge University Press, 1998).
18. Y. Zhang, J. Bradic, High-dimensional semi-supervised learning: In search of optimal inference of the mean. *Biometrika* **109**, 387–403 (2022).
19. A. Chakrabortty, G. Dai, R. J. Carroll, Semi-supervised quantile estimation: Robust and efficient inference in high dimensional settings. arXiv [Preprint] (2022). http://arxiv.org/abs/2201.10208 (Accessed 14 December 2023).
20. D. Azriel *et al.*, Semi-supervised linear regression. *J. Am. Stat. Assoc.* **117**, 2238–2251 (2022).
21. T. Tony Cai, Z. Guo, Semisupervised inference for explained variance in high dimensional linear regression and its applications. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **82**, 391–419 (2020).
22. S. Song, Y. Lin, Y. Zhou, A general m-estimation theory in semi-supervised framework. *J. Am. Stat. Assoc.* 1–11 (2023), 10.1080%2F01621459.2023.2169699.
23. J. E. Van Engelen, H. H. Hoos, A survey on semi-supervised learning. *Mach. Learn.* **109**, 373–440 (2020).
24. X. Zhu, A. B. Goldberg, *Introduction to Semi-Supervised Learning* (Springer Nature, 2022).
25. A. N. Angelopoulos, J. C. Duchi, T. Zrnic, PPI++: Efficient prediction-powered inference. arXiv [Preprint] (2023). http://arxiv.org/abs/2311.01453 (Accessed 14 December 2023).
26. S. Wang, T. H. McCormick, J. T. Leek, Methods for correcting inference based on outcomes predicted by machine learning. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30266–30275 (2020).
27. K. Motwani, D. Witten, Valid inference after prediction. arXiv [Preprint] (2023). http://arxiv.org/abs/2306.13746 (Accessed 14 December 2023).
28. S. Dudoit, M. J. van der Laan, Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat. Methodol.* **2**, 131–154 (2005).
29. M. Austern, W. Zhou, Asymptotics of cross-validation. arXiv [Preprint] (2020). http://arxiv.org/abs/2001.11111 (Accessed 14 December 2023).
30. P. Bayle, A. Bayle, L. Janson, L. Mackey, Cross-validation confidence intervals for test error. *Adv. Neural Inf. Process. Syst.* **33**, 16339–16350 (2020).
31. N. Kissel, J. Lei, On high-dimensional Gaussian comparisons for cross-validation. arXiv [Preprint] (2022). http://arxiv.org/abs/2211.04958 (Accessed 14 December 2023).
32. S. Bates, T. Hastie, R. Tibshirani, Cross-validation: What does it estimate and how well does it do it? *J. Am. Stat. Assoc.* 1–12 (2023), 10.1080%2F01621459.2023.2197686.
33. B. Y. Levit, On the efficiency of a class of non-parametric estimates. *Theory Probab. Appl.* **20**, 723–740 (1976).
34. R. Z. Hasminskii, I. A. Ibragimov, "On the nonparametric estimation of functionals" in *Proceedings of the Second Prague Symposium on Asymptotic Statistics* (North-Holland Amsterdam, 1979), vol. 473, pp. 474–482.
35. C. A. Klaassen, Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Stat.* **15**, 1548–1562 (1987).
36. P. M. Robinson, Root-n-consistent semiparametric regression. *Econ. J. Econ. Soc.* **56**, 931–954 (1988).
37. P. J. Bickel *et al.*, *Efficient and Adaptive Estimation for Semiparametric Models* (Springer, 1993), vol. 4.
38. W. K. Newey, The asymptotic variance of semiparametric estimators. *Econom. J. Econom. Soc.* 1349–1382 (1994).
39. J. M. Robins, A. Rotnitzky, Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Stat. Assoc.* **90**, 122–129 (1995).
40. V. Chernozhukov *et al.*, Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21**, C1–C68 (2018).
41. H. Bang, J. M. Robins, Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973 (2005).
42. J. M. Robins, A. Rotnitzky, L. P. Zhao, Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.* **89**, 846–866 (1994).
43. A. Rotnitzky, J. M. Robins, D. O. Scharfstein, Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Am. Stat. Assoc.* **93**, 1321–1339 (1998).
44. W. K. Newey, D. McFadden, Large sample estimation and hypothesis testing. *Handb. Econ.* **4**, 2111–2245 (1994).
45. Y. Jin, D. Rothenhäusler, Tailored inference for finite populations: Conditional validity and transfer across distributions. *Biometrika* **111**, asad022 (2023).
46. V. Chernozhukov, J. C. Escanciano, H. Ichimura, W. K. Newey, J. M. Robins, Locally robust semiparametric estimation. *Econometrica* **90**, 1501–1535 (2022).
47. V. Chernozhukov, W. K. Newey, R. Singh, A simple and general debiased machine learning theorem with finite-sample guarantees. *Biometrika* **110**, 257–264 (2023).
48. D. B. Rubin, Inference and missing data. *Biometrika* **63**, 581–592 (1976).
49. D. Rubin, Multiple imputation for nonresponse in surveys. Wiley Series in Probability and Statistics (1987), p. 1.
50. D. B. Rubin, Multiple imputation after 18+ years. *J. Am. Stat. Assoc.* **91**, 473–489 (1996).
51. J. L. Schafer, Multiple imputation: A primer. *Stat. Methods Med. Res.* **8**, 3–15 (1999).
52. H. White, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econ. J. Econ. Soc.* **48**, 817–838 (1980).
53. H. White, Consequences and detection of misspecified nonlinear regression models. *J. Am. Stat. Assoc.* **76**, 419–433 (1981).
54. A. Buja *et al.*, Models as approximations I. *Stat. Sci.* **34**, 523–544 (2019).
55. A. Buja *et al.*, Models as approximations II. *Stat. Sci.* **34**, 545–565 (2019).
56. T. Zrnic, E. J. Candès, cross-ppi. GitHub. https://github.com/tijana-zrnic/cross-ppi. Accessed 1 March 2024.
57. T. Chen, C. Guestrin, "XGBoost: A scalable tree boosting system" in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* (2016), pp. 785–794.
58. F. Ding, M. Hardt, J. Miller, L. Schmidt, Retiring adult: New datasets for fair machine learning. *Adv. Neural Inf. Process. Syst.* **34**, 6478–6490 (2021).
59. K. W. Willett *et al.*, Galaxy zoo 2: Detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices R. Astron. Soc.* **435**, 2835–2860 (2013).
60. D. G. York *et al.*, The sloan digital sky survey: Technical summary. *Astron. J.* **120**, 1579 (2000).