*Biometrika* (2025), **112**, 1, asae056
https://doi.org/10.1093/biomet/asae056
Advance Access publication 30 October 2024

# Semi-supervised distribution learning

By MENGTAO WEN, YINXU JIA

*School of Statistics and Data Science, Nankai University,*
*94 Weijin Road, Tianjin 300071, China*
mtwen97@gmail.com yxjia@mail.nankai.edu.cn

HAOJIE REN

*School of Mathematical Sciences, Shanghai Jiao Tong University,*
*800 Dongchuan Road, Shanghai 200240, China*
haojieren@sjtu.edu.cn

ZHAOJUN WANG and CHANGLIANG ZOU

*School of Statistics and Data Science, Nankai University,*
*94 Weijin Road, Tianjin 300071, China*
zjwangnk@126.com nk.chlzou@gmail.com

## Summary

This study addresses the challenge of distribution estimation and inference in a semi-supervised setting. In contrast to prior research focusing on parameter inference, this work explores the complexities of semi-supervised distribution estimation, particularly the uniformity problem inherent in functional processes. To tackle this issue, we introduce a versatile framework designed to extract valuable information from unlabelled data by approximating a conditional distribution on covariates. The proposed estimator is derived using $K$-fold cross-fitting, and exhibits both consistency and asymptotic Gaussian process properties. Under mild conditions, the proposed estimator outperforms the empirical cumulative distribution function in terms of asymptotic efficiency. Several applications of the methodology are given, including parameter inference and goodness-of-fit tests.

*Some key words*: Asymptotic Gaussian process; Bias correction; Distributional regression; Functional delta theorem; Semi-supervised distribution test.

## 1. Introduction

Distribution estimation stands as one of the most fundamental problems in statistical theory and machine learning, encompassing applications such as goodness-of-fit testing, classification, regression and empirical Bayes methods. Traditionally, the empirical cumulative distribution function is a widely adopted estimator in this realm. Given a set of independent and identically distributed random variables $\{Y_i\}_{i=1}^n$ with a common cumulative distribution function $F$, the empirical cumulative distribution function (abbreviated as *ecdf* in subscript) is defined as $\hat{F}_{\mathrm{ecdf}}(y) = n^{-1}\sum_{i=1}^n \mathbb{1}_y(Y_i)$, where $\mathbb{1}_y(Y)$ is the indicator function of event $\{Y \leqslant y\}$. The empirical cumulative distribution function exhibits appealing properties, such as $\sqrt{n}$ consistency in the $\ell_\infty$ norm and asymptotic normality (Dvoretzky et al., 1956; van der Vaart, 1998).

Despite the effectiveness of the empirical cumulative distribution function as a nonparametric maximum likelihood estimator relying solely on $\{Y_i\}_{i=1}^n$, we study scenarios where obtaining $Y_i$ is challenging due to high cost, time constraints or practical limitations. Often in such situations, some associated side information, represented by covariates $X_i \in \mathbb{R}^p$, is readily available. This introduces a typical semi-supervised scenario, featuring two distinct datasets: a small or moderate labelled set $\mathcal{L} = \{(X_i, Y_i)\}_{i=1}^n$ and a large unlabelled set $\mathcal{U} = \{X_i\}_{i=n+1}^{n+m}$ with unobserved $\{Y_i\}_{i=n+1}^{n+m}$, where $\{(X_i, Y_i)\}_{i=1}^{n+m}$ are independent copies of $(X, Y)$ and $m \gg n$. This situation is pervasive, as seen in applications like electronic health records in healthcare, gene expression data in genomics and face recognition in computer science. In such cases, a pertinent question arises: can one more efficiently estimate the distribution of $Y$ by leveraging the substantial side information present in both labelled and unlabelled data, thereby surpassing the capabilities of $\hat{F}_{\rm ecdf}(y)$?

Semi-supervised inference has emerged as a valuable tool for harnessing both labelled and unlabelled data to enhance parameter inference and predictive capabilities, making significant advancements in both the statistics and machine learning communities over the past decade. Among others, Zhang et al. (2019) pioneered a method wherein the unlabelled $Y$ was replaced with its prediction derived from a linear projection of the conditional mean $E(Y \mid X)$. This approach involved averaging both labelled $Y$ and predictions of unlabelled $Y$ to yield a semi-supervised mean estimator. Building upon this, Zhang & Bradic (2022) extended the mean estimator to high-dimensional settings, addressing potential biases. In the realm of quantile inference, Chakrabortty et al. (2024) introduced a one-step semi-supervised estimator. The work of Yuval & Rosset (2022) proposed an empirical risk minimization framework, analysing the derived semi-supervised estimator for generalized linear models. Adopting a perspective rooted in $M$-estimation for finite-dimensional parameter estimation, Song et al. (2024) applied linear projection techniques to the loss function. Meanwhile, Angelopoulos et al. (2023) imputed the unlabelled $Y$ and aimed to enhance the score function from the standpoint of $Z$-estimation.

Despite the notable strides in semi-supervised inference methods, the focus of these advancements has mainly centred on a specific class of parameters, which often pertain to the expectation or quantiles of a random variable. However, when dealing with parameters involving multiple independent copies of a random variable, such as a one-sample $U$-statistic, the direct application of existing methodologies appears to face theoretical and practical difficulties. Fortunately, the aforementioned parameters can be expressed as continuous functions of the target cumulative distribution function $F$, denoted $\theta(F)$. If we have a distribution estimator $\hat{F}$ of $F$ and its asymptotic distribution, a consistent estimator $\theta(\hat{F})$ of $\theta(F)$ and its asymptotic distribution can be naturally derived. This observation underscores the importance of addressing the problem of semi-supervised distribution estimation and inference, which, to the best of our knowledge, has remained unexplored. The challenge lies in the transition from a finite-dimensional parameter to an infinite-dimensional function, introducing the complexity of uniformity.

This article introduces a framework for distribution estimation and inference within semi-supervised contexts. Motivated by the fact that $F(y) = E\{\mathbb{1}_y(Y)\} = E\{F(y \mid X)\}$, where $F(y \mid x) = \mathrm{pr}(Y \leqslant y \mid X = x)$, if we have a reliable estimator of $F(y \mid x)$, denoted $\tilde{F}(y \mid x)$, we can naturally derive an intuitive estimator of $F(y)$ through $\hat{F}_{\rm intu}(y) = (n+m)^{-1}\sum_{i=1}^{n+m} \tilde{F}(y \mid X_i)$ based on both labelled and unlabelled data. A rich body of literature exists on conditional cumulative distribution function estimations based on labelled data; see Kneib et al. (2023) for a comprehensive review. However, most estimation methods fall short in yielding a fast converged $\tilde{F}(y \mid X)$ without stringent and often impractical assumptions, significantly limiting their applicability in practice.

To address this issue, we observe that the potential bias term $E\{\mathbb{1}_y(Y) - \tilde{F}(y \mid X)\}$ can be effectively estimated using labelled data $(X_i, Y_i) \in \mathcal{L}$, when the estimated function $\tilde{F}(y \mid x)$ is independent of labelled data. Inspired by this fact, we propose a general estimator, termed the *bias-amended semi-supervised distribution* estimator (shortened to *basd* in subscript),

$$\tilde{F}_{\rm basd}(y) = \frac{1}{n+m}\sum_{i=1}^{n+m} \tilde{F}(y \mid X_i) + \frac{1}{n}\sum_{i=1}^{n}\{\mathbb{1}_y(Y_i) - \tilde{F}(y \mid X_i)\}. \tag{1}$$

In essence, the first term leverages both labelled and unlabelled data to enhance efficiency, while the second term is designed to rectify bias using the labelled data. To ensure the validity of the proposed estimator (1), we employ a cross-fitting technique to acquire $\tilde{F}(y \mid x)$, seamlessly integrating various conditional cumulative distribution function estimation methods into our framework. Our analysis establishes the $\sqrt{n}$ consistency of the cross-fitted estimator to $F(y)$ uniformly over $y \in \mathbb{R}$, necessitating only weak consistency of $\tilde{F}(y \mid x)$ instead of imposing specific requirements on the convergence rate of $\tilde{F}(y \mid x)$. The asymptotic Gaussian process of the proposed cross-fitted estimator is further established, where the asymptotic covariance matrix characterizes how the unlabelled data contribute to the estimation efficiency compared to $\hat{F}_{\text{ecdf}}(y)$. It is crucial to emphasize that the transition from parameter inference to functional inference yields benefits beyond functional inference itself. Leveraging the functional delta theorem, our approach, not only recovers specialized solutions for parameter inference found in the existing literature, but also offers a general and user-friendly method for inferring a broad class of parameters beyond the scope of prior works. Simulation studies demonstrate the practical efficacy and applicability of the proposed semi-supervised framework.

We now define some additional notation. For a set $S$, the space $\ell^\infty(S)$ is defined as the set of all uniform bounded and real functions on $S$. Let $\rightsquigarrow$ denote convergence in distribution, and $\overset{\text{D}}{=}$ denote equality in distribution. For a probability measure $P$, the $L_r(P)$ norm for some $r \geqslant 1$ is $\|f\|_{P,r} = (P|f|^r)^{1/r}$, where $Pf = \int f \,\mathrm{d}P$. For a function class $\mathcal{F}$, the bracketing number $N_{[\,]}\{\epsilon, \mathcal{F}, L_r(P)\}$ is the minimum number of $\epsilon$ brackets needed to cover $\mathcal{F}$, as described in van der Vaart (1998). Let $P_X$ be the probability measure of $X$, and $E_X$ be the expectation only with respect to $X$. Denote by $\mathbb{B} \circ f_y(X, Y)$ a centred Gaussian process indexed by $y$ with covariance $\text{cov}[\{f_s(X, Y), f_t(X, Y)\}^{\mathrm{T}}]$ for any $s, t \in \mathbb{R}$.

## 2. Bias-amended semi-supervised distribution

### 2.1. *Cross-fitted bias-amended semi-supervised distribution estimators*

Recall the proposed estimator in (1). We need to obtain the estimated conditional distribution function $\tilde{F}(y \mid x)$ based on the labelled data $\mathcal{L}$. While various techniques for estimating the conditional distribution function could be directly employed, technical challenges in the proposed estimator (1) arise from the intricate dependence between $\tilde{F}(y \mid x)$ and labelled data $(Y_i, X_i) \in \mathcal{L}$. To address this issue, we employ $K$-fold cross-fitting (Chernozhukov et al., 2018) to obtain the semi-supervised distribution estimator.

Concretely, we randomly partition data points in the labelled set $\mathcal{L} = \{(X_i, Y_i)\}_{i=1}^n$ and unlabelled set $\mathcal{U} = \{X_i\}_{i=n+1}^{n+m}$ into $K$ folds, denoted $\mathcal{L} = \mathcal{L}_1 \cup \cdots \cup \mathcal{L}_K$ and $\mathcal{U} = \mathcal{U}_1 \cup \cdots \cup \mathcal{U}_K$, respectively, where $K \geqslant 2$ is a fixed integer. The corresponding index sets are denoted $\mathcal{I}_1, \ldots, \mathcal{I}_K$ (labelled) and $\mathcal{J}_1, \ldots, \mathcal{J}_K$ (unlabelled), respectively. Without loss of generality, assume that $K$ divides both $n$ and $m$. Define $n_K = n/K$, $m_K = m/K$. For each $k \in \{1, \ldots, K\}$, letting $\mathcal{I}_{-k} = \{1, \ldots, n\} \setminus \mathcal{I}_k$ and $\mathcal{L}_{-k} = \mathcal{L} \setminus \mathcal{L}_k$, we utilize a conditional cumulative distribution function estimation algorithm $\mathcal{A}$ to obtain the estimator $\hat{F}_k(\cdot \mid x)$ with $\mathcal{L}_{-k}$. Then, we construct the $k$th fold estimator with all data in $\mathcal{I}_k \cup \mathcal{J}_k$ by (1),

$$\hat{F}_{k,\mathrm{B}}(y) = \frac{1}{n_K + m_K} \sum_{i \in \mathcal{I}_k \cup \mathcal{J}_k} \hat{F}_k(y \mid X_i) + \frac{1}{n_K} \sum_{i \in \mathcal{I}_k} \{\mathbb{1}_y(Y_i) - \hat{F}_k(y \mid X_i)\}, \qquad y \in \mathbb{R}.$$

These $K$ estimators are then aggregated into the cross-fitted estimator

$$\hat{F}_{\text{basd}}(y) = \frac{1}{K} \sum_{k=1}^{K} \hat{F}_{k,\mathrm{B}}(y), \qquad y \in \mathbb{R}.$$

Estimator $\hat{F}_{\text{basd}}(y)$ shares a conceptual similarity with Zhang & Bradic (2022) and Zrnic & Candès (2024), who focused on obtaining semi-supervised inference for a parameter. However, our aim is to estimate the entire distribution function rather than a single value. The uniformity introduced by the function estimation poses a new and challenging aspect, distinguishing our approach.

### 2.2. *Theoretical analysis*

It is important to highlight that the $\hat{F}_{\text{basd}}(y)$ remains unbiased, regardless of whether $\hat{F}_k(y \mid x)$ is misspecified or not well approximated. To develop the uniformity theory for $\hat{F}_{\text{basd}}(y)$, we introduce the following assumptions.

*Assumption* 1. There exists some nonrandom function $F_0(y \mid x)\colon \mathbb{R} \times \mathbb{R}^p \to [0,1]$, such that $U_k(x) = \sup_{y\in\mathbb{R}} |\hat{F}_k(y \mid x) - F_0(y \mid x)|$ satisfies $\varrho_n^2 = \max_{k=1,\ldots,K} E_X[\{U_k(X)\}^2] = o_p(1)$.

*Assumption* 2. The measurable function classes $\mathcal{G} = \{F_0(y \mid X)\colon y \in \mathbb{R}\}$ and $\mathcal{G}_k = \{\hat{F}_k(y \mid X)\colon y \in \mathbb{R}\}$ satisfy $\log N_{[\,]}\{\epsilon, \mathcal{G}, L_2(P_X)\} \lesssim \epsilon^{-\eta}$ and $\log N_{[\,]}\{\epsilon, \mathcal{G}_k, L_2(P_X)\} \lesssim \epsilon^{-\eta}$ for every $\epsilon > 0$, some $\eta \in (0,2)$ and each $k \in \{1,\ldots,K\}$.

Assumption 1 requires that the conditional cumulative distribution function estimator $\hat{F}_k(y \mid x)$ converges to $F_0(y \mid x)$ in each fold weakly in the sense that $E_X\{U_k(X)\}^2 = o_p(1)$. This can be guaranteed by the strong consistency $\sup_{y,x} |\hat{F}_k(y \mid x) - F_0(y \mid x)| = o_p(1)$, which is commonly provided by the literature (Henzi et al., 2021; Čevid et al., 2022). Assumption 2 holds if $\hat{F}_k(y \mid x)$ and $F_0(y \mid x)$ exhibit monotonic behaviour with respect to $y$ for each $x$, which can usually be guaranteed by the rearrangement strategy (Chernozhukov et al., 2010), as outlined in the Supplementary Material. Thus, Assumptions 1–2 are quite mild, providing flexibility in choosing the learning algorithm $\mathcal{A}$. We also suppose that both labelled and unlabelled samples are independently and identically distributed. Extending the proposed estimator to other scenarios, such as missing data (Zhang et al., 2023), remains an important area for future work.

Theorem 1. *Suppose that Assumptions 1–2 hold. Let $\gamma_n = m(n+m)^{-1}$. Then, there exist some constants $C, c > 0$ such that, for any $0 < \delta \leqslant 2K^{-1/2} n^{1/2} \varrho_n^{(2-\eta)/4}$,*

$$\text{pr}\Big[\sup_{y\in\mathbb{R}} |n^{1/2}\{\hat{F}_{\text{basd}}(y) - F(y)\}| \geqslant \{\overline{\sigma}(\mathcal{G}) + \omega_n \varrho_n^{(2-\eta)/4}\}\delta\Big] \leqslant C\exp(-c\delta^2), \tag{2}$$

*where*

$$\overline{\sigma}^2(\mathcal{G}) = \sup_{y\in\mathbb{R}}[(1-\gamma_n)\text{var}\{\mathbb{1}_y(Y)\} + \gamma_n \text{var}\{\mathbb{1}_y(Y) - F_0(y \mid X)\}]$$

$$\textit{and} \quad \omega_n = K^{1/2}\gamma_n + \{K\gamma_n(1-\gamma_n)\}^{1/2}.$$

*Moreover, assuming that $\gamma_n \to \gamma \in [0,1]$, we have, as $n, m \to \infty$,*

$$n^{1/2}\{\hat{F}_{\text{basd}}(y) - F(y)\} \rightsquigarrow \mathbb{F}(y;\mathcal{G}), \tag{3}$$

*uniformly for $y \in \mathbb{R}$, where*

$$\mathbb{F}(y;\mathcal{G}) = (1-\gamma)^{1/2}\mathbb{B}_1 \circ \mathbb{1}_y(Y) + \gamma^{1/2}\mathbb{B}_2 \circ \{\mathbb{1}_y(Y) - F_0(y \mid X)\},$$

*and $\mathbb{B}_1, \mathbb{B}_2$ are two independent Brownian bridges.*

Theorem 1 presents $\sqrt{n}$ consistency in (2) and the asymptotic Gaussian process limit in (3), similar to the results of the classical empirical cumulative distribution function $\hat{F}_{\text{ecdf}}(y)$ (Dvoretzky et al., 1956; Donsker, 1952). Parameter $\gamma$ represents the asymptotic proportion of unlabelled data. Our results are valid, not only in the classic semi-supervised setting ($\gamma = 1$), but also in scenarios where the labelled and unlabelled data have comparable sizes ($\gamma \in (0,1)$). For the special case $\gamma = 0$, which implies that the unlabelled data can be ignored asymptotically, Theorem 1 indicates that the behaviour of $\hat{F}_{\text{basd}}(y)$ is almost the same as that of $\hat{F}_{\text{ecdf}}(y)$. Theorem 1 holds provided that the conditional distribution of $Y \mid X$ and the corresponding estimator satisfy Assumptions 1–2. Extending this result to certain uniform versions over a reasonable class of distributions (e.g., Christgau et al., 2023) may provide more stable inference, meriting further study.

*Remark* 1 (*Monotonization via rearrangement*). The potential nonmonotonicity of the proposed estimator, caused by the subtraction term in (1), may be visually apparent and affect interpretation. Again, by applying the rearrangement operation (Chernozhukov et al., 2009), we can easily get the monotonized distribution estimator without altering any conclusions in Theorem 1.

*Remark* 2 (*Diverging $p$*). Estimator $\hat{F}_{\mathrm{basd}}(y)$ remains applicable in the context of diverging $p$ as long as the chosen conditional distribution estimation method satisfies the weak consistency of Assumption 1. Many machine learning–based conditional distribution estimators (e.g., Ćevid et al., 2022; Shen & Meinshausen, 2024) can be incorporated into this framework for high-dimensional data, with numerical studies indicating promising performance. Nevertheless, whether those more sophisticated machine learning–based estimators satisfy Assumption 1 remains an open question for future research.

Next, we study the efficiency of $\hat{F}_{\mathrm{basd}}$ compared with $\hat{F}_{\mathrm{ecdf}}$. According to Theorem 1, the asymptotic covariance of $\hat{F}_{\mathrm{basd}}$ corresponds to the covariance of a centred Gaussian process $\mathbb{F}(y; \mathcal{G})$, i.e.,

$$\operatorname{cov}\begin{pmatrix}\mathbb{F}(s;\mathcal{G})\\ \mathbb{F}(t;\mathcal{G})\end{pmatrix} = (1-\gamma)\operatorname{cov}\begin{pmatrix}\mathbb{1}_s(Y)\\ \mathbb{1}_t(Y)\end{pmatrix} + \gamma\operatorname{cov}\begin{pmatrix}\mathbb{1}_s(Y) - F_0(s \mid X)\\ \mathbb{1}_t(Y) - F_0(t \mid X)\end{pmatrix}$$

for any $s, t \in \mathbb{R}$. Notably, the permit of model misspecification (i.e., $F_0(y \mid x) \neq F(y \mid x)$) in Assumption 1 does not affect the consistency and weak convergence of $\hat{F}_{\mathrm{basd}}$, since $\hat{F}_{\mathrm{basd}}$ is always unbiased by its construction, but $F_0(y \mid x)$ does affect the asymptotic covariance. Conducting an efficiency comparison for a less restrictive function $F_0(y \mid x)$ is challenging, so we consider a specific case where the learning algorithm $\mathcal{A}$ yields $F_0(y \mid x)$ as a conditional distribution function of $Y$ on some projection of $X$.

Proposition 1. *Suppose that* $F_0(y \mid X) = \mathrm{pr}\{Y \leqslant y \mid h(X)\}$ *for some function* $h \colon \mathbb{R}^p \mapsto \mathbb{R}^q$ *for an integer* $q > 0$. *The asymptotic covariance of* $\hat{F}_{\mathrm{basd}}$ *satisfies*

$$\operatorname{cov}\begin{pmatrix}\mathbb{F}(s;\mathcal{G})\\ \mathbb{F}(t;\mathcal{G})\end{pmatrix} = \operatorname{cov}\begin{pmatrix}\mathbb{1}_s(Y)\\ \mathbb{1}_t(Y)\end{pmatrix} - \gamma\operatorname{cov}\begin{pmatrix}F_0(s \mid X)\\ F_0(t \mid X)\end{pmatrix}.$$

Proposition 1 demonstrates that the $\hat{F}_{\mathrm{basd}}(y)$ achieves higher efficiency than the empirical cumulative distribution function, since the asymptotic covariance of $\hat{F}_{\mathrm{ecdf}}(y)$ is known as $\operatorname{cov}[\{\mathbb{1}_s(Y), \mathbb{1}_t(Y)\}^{\mathrm{T}}]$. Numerous examples of algorithm $\mathcal{A}$ for such $F_0(y \mid x) = \mathrm{pr}\{Y \leqslant y \mid h(x)\}$ appear in many conditional distribution estimation methods, for example, the linear projection in Hall & Yao (2005) and generalized single-index models in Henzi et al. (2021). When $F_0(y \mid X) = \mathrm{pr}(Y \leqslant y \mid X = x)$ is the true conditional cumulative distribution function, the covariance of $\mathbb{F}(\cdot; \mathcal{G})$ attains the semiparametric efficiency lower bound.

*Remark* 3 (*Inference for $F$*). Theorem 1 allows us to construct a simultaneous confidence band for $F$. The $(1-\alpha)$th confidence band can be expressed as $[\hat{F}_{\mathrm{basd}}(y) - Ln^{-1/2}, \hat{F}_{\mathrm{basd}}(y) + Ln^{-1/2}]$, where threshold $L$ is determined such that $\mathrm{pr}[\sup_{y\in\mathbb{R}} |n^{1/2}\{\hat{F}_{\mathrm{basd}}(y) - F(y)\}| \leqslant L] \approx 1-\alpha$ with the significance level $\alpha$. In practice, the resampling strategy can be employed to approximate $L$. Specifically, we draw $B$ realizations $\mathbb{F}_1, \ldots, \mathbb{F}_B$ from the Gaussian process $\mathbb{F}(\cdot; \mathcal{G})$ in (3) (with estimated covariance matrices substituted) and compute $g_b = \sup_{y\in\mathbb{R}} |\mathbb{F}_b(y)|$ $(b = 1, \ldots, B)$. Threshold $L$ is then approximated by the $(1-\alpha)$th sample quantile of $\{g_1, \ldots, g_B\}$. Similarly, one can construct Kolmogorov–Smirnov goodness-of-fit tests. They are effectively applied in the problem of the local distributional treatment effect, with improved power over the standard test, as detailed in the Supplementary Material.

### 2.3. *Parameter inference from the proposed estimator*

In this section, we return to parameter inference from functional inference. We consider the parameter of interest as a continuous function of $F$, denoted $\theta(F)$, then its semi-supervised estimator can be directly the plug-in estimator $\theta(\hat{F}_{\text{basd}})$. Leveraging the uniform convergence of $\hat{F}_{\text{basd}}(y)$ in Theorem 1 and the functional delta theorem (van der Vaart, 1998), it is natural to derive the asymptotic normality of $\theta(\hat{F}_{\text{basd}})$ as follows.

Corollary 1. *Suppose that $\theta(F)$ is the function of interest with a mapping $\mathbb{D}_\theta \subset \ell^\infty(\mathbb{R})$ to $\mathbb{R}$, where $\mathbb{D}_\theta$ is some subspace determined by $\theta(\cdot)$, and $\theta(F)$ is Hadamard differentiable at $F$ with derivative $\nabla\theta_F(\cdot)$. If Theorem 1 holds then*

$$n^{1/2}\{\theta(\hat{F}_{\text{basd}}) - \theta(F)\} \rightsquigarrow \nabla\theta_F\{\mathbb{F}(\cdot;\mathcal{G})\}.$$

The proposed plug-in estimator aligns with existing results in the literature, such as those for the mean and quantiles; see Examples S1–S2 in the Supplementary Material. We provide the algorithm and computational details of the proposed method in the Supplementary Material, along with a numerical comparison of semi-supervised mean estimators. The performance of $\theta(\hat{F}_{\text{basd}})$ is comparable to that of specialized semi-supervised mean estimators, supporting the corresponding theoretical justification.

Additionally, it enables inference on more general parameters beyond the scope of the existing literature.

*Example* 1 (*One-sample U-statistics*). Let $\theta(F) = \int\int\int f(y_1,\ldots,y_r)\,\mathrm{d}F(y_1)\cdots\mathrm{d}F(y_r)$ be the parameter of interest, where $f(y_1,\ldots,y_r)$ is the kernel function of degree $r$. Then,

$$\nabla\theta_F\{\mathbb{F}(\cdot;\mathcal{G})\} \overset{\mathrm{D}}{=} \mathcal{N}\Bigg(0, (1-\gamma)r^2\mathrm{var}[E_{Y_2,\ldots,Y_r}\{f(Y, Y_2,\ldots,Y_r)\}] + \gamma r^2\mathrm{var}\Bigg[E_{Y_2,\ldots,Y_r}\{f(Y, Y_2,\ldots,Y_r)\} - \int E_{Y_2,\ldots,Y_r}\{f(y, Y_2,\ldots,Y_r)\}\,\mathrm{d}F_0(y \mid X)\Bigg]\Bigg).$$

The first term in the asymptotic variance coincides with that of the $U$-statistic. For the second part, we expect that $\int E_{Y_2,\ldots,Y_r}\{f(y, Y_2,\ldots,Y_r)\}\,\mathrm{d}F_0(y \mid X)$ is a good approximation of the conditional mean function $E\{E_{Y_2,\ldots,Y_r}f(Y, Y_2,\ldots,Y_r) \mid X\}$. Accordingly, the asymptotic variance of the proposed plug-in semi-supervised estimator is reduced.

*Remark* 4. While this article focuses primarily on the distribution estimation and inference for response $Y$, our framework can also accommodate the responses that are functions of $Y$ and $X$, such as the loss function of $M$-estimation (Song et al., 2024) or the score function of $Z$-estimation (Angelopoulos et al., 2023), as long as they can be written as a function of some distribution $F$. For example, for the squared loss $E\{(Y - X^{\mathrm{T}}\beta)^2\}$ in population, we can define $Z(\beta) = (Y - X^{\mathrm{T}}\beta)^2$ and use $\{(Z_i(\beta), X_i)\}_{i=1}^{n}$ and $\{X_i\}_{i=n+1}^{n+m}$ to get a semi-supervised loss.

## 3. Numerical studies

The data are generated from the model

$$Y_i = \mu(X_i) + \sigma(X_i)\varepsilon_i, \qquad i = 1,\ldots,n+m,$$

where $X_i = (X_{i1},\ldots,X_{ip})^{\mathrm{T}} \sim \mathcal{N}_p(0_p, I_p)$ and $\varepsilon_i \sim \mathcal{N}(0,1)$ are independent. We consider $\mu(X_i) = s^{-1/2}\sum_{j=1}^{s} X_{ij}$ and $\sigma(X_i) = 1/2$, where $s = \lceil 0.1p \rceil$ is the number of signals. We set the

Table 1. *Comparison of the proposed $\hat{F}_{\rm basd}$ and the empirical cumulative distribution function $\hat{F}_{\rm ecdf}$. The overall mean squared error of the empirical cumulative distribution function is denoted by* $\text{MSE}_0$. *The level of the confidence band is set as* $\alpha = 0.1$.

| | $p = 100$ | | | $p = 500$ | | |
|---|---|---|---|---|---|---|
| | $\text{MSE}/\text{MSE}_0$ | Coverage | Length $\times 10^2$ | $\text{MSE}/\text{MSE}_0$ | Coverage | Length $\times 10^2$ |
| $\tilde{F}^*_{\rm basd}$ | 0.50 | 0.92 | 5.51 | 0.46 | 0.90 | 5.53 |
| $\hat{F}_{\rm basd}$-GAMLSS | 0.51 | 0.90 | 5.57 | 0.88 | 0.88 | 7.04 |
| $\hat{F}_{\rm basd}$-Engression | 0.82 | 0.88 | 7.16 | 0.82 | 0.88 | 7.57 |
| $\hat{F}_{\rm basd}$-DRF | 0.90 | 0.90 | 7.05 | 0.99 | 0.86 | 7.35 |
| $\hat{F}_{\rm ecdf}$ | 1.00 | 0.90 | 7.36 | 1.00 | 0.87 | 7.37 |

$\tilde{F}^*_{\rm basd}$, $\tilde{F}_{\rm basd}$ with known $F(y \mid x)$; $\hat{F}_{\rm basd}$-GAMLSS, $\hat{F}_{\rm basd}$-Engression and $\hat{F}_{\rm basd}$-DRF represent the $\hat{F}_{\rm basd}$ estimators with $\mathcal{A}$ specified as the boosting method for fitting generalized additive models (Hofner et al., 2016), the engression estimator (Shen & Meinshausen, 2024) and distributional random forests (Ćevid et al., 2022), respectively.

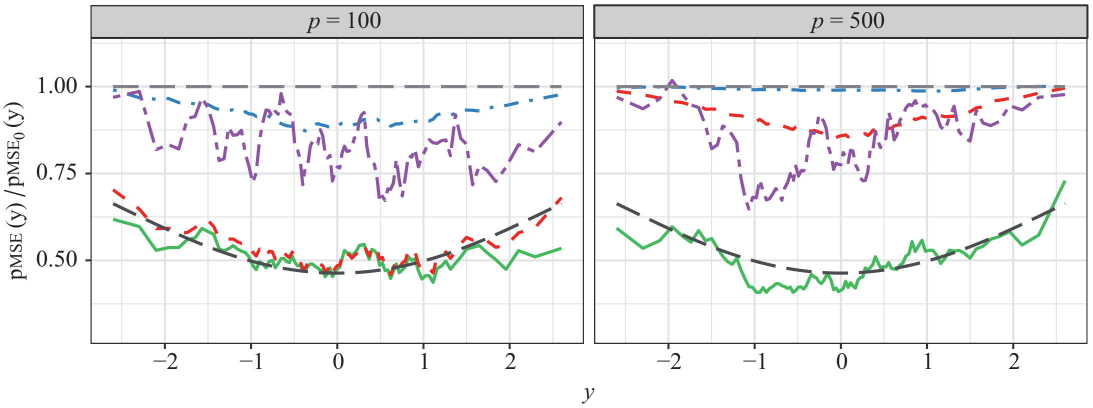

Fig. 1. The ratio of pointwise mean squared errors of $\hat{F}_{\rm basd}$. The pointwise mean squared error of $\hat{F}_{\rm ecdf}$ is denoted by $\text{pMSE}_0(y)$. The plot includes $\tilde{F}^*_{\rm basd}$ (green solid line), $\hat{F}_{\rm basd}$-GAMLSS (red dashed line), $\hat{F}_{\rm basd}$-Engression (purple long-dash–short-dash line) and $\hat{F}_{\rm basd}$-DRF (blue dash-dot line). The curve (dark grey long-dash line) is the theoretical relative semi-parametric efficiency lower bound. See the footnote to Table 1 for an explanation of the method notation.

number of cross-fitting folds as $K = 10$ in $\hat{F}_{\rm basd}$. Three conditional distribution estimation algorithms $\mathcal{A}$ are considered: the boosting method for fitting generalized additive models (Hofner et al., 2016), distributional random forests (Ćevid et al., 2022) and the neural network–based engression method (Shen & Meinshausen, 2024). We also provide the results of the proposed estimator with the true conditional distribution $F(y \mid x)$. Two metrics are considered for evaluating the distribution estimation: the pointwise mean squared error $\text{pMSE}(y) = E\{\hat{F}(y) - F(y)\}^2$ and the overall mean squared error $\text{MSE} = \int \text{pMSE}(y)\, \mathrm{d}y$. The average coverage and length of the confidence band are considered for distribution inference.

The results for $n = 1000$ and $m = 10\,000$ with different $p$ are reported in Table 1 and Fig. 1. A ratio of mean squared errors less than one means that the proposed $\hat{F}_{\rm basd}$ is more efficient than $\hat{F}_{\rm ecdf}$. The results show that the proposed $\hat{F}_{\rm basd}$ is usually at least as efficient as $\hat{F}_{\rm ecdf}$, and the improvement is significant especially when the conditional distribution can be well approximated.

Additional simulations can be found in the Supplementary Material, including the nonlinear model, conformal $p$-values and two-sample Kolmogorov–Smirnov tests.

## Acknowledgement

The authors are grateful to the referees, associate editor and editor for comments that significantly improved the article. Wen and Jia contributed equally to this work. Ren and Zou are the corresponding authors. This research was supported by the China National Key R&D Program and the National Natural Science Foundation of China. Wang and Zou are also affiliated with LPMC, KLMDASR and LEBPS at Nankai University.

## Supplementary material

The Supplementary Material contains all technical proofs; applications, including the semi-supervised conformal $p$-values and local distributional treatment effects; additional numerical studies and a real data analysis.

## References

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I. & Zrnic, T. (2023). Prediction-powered inference. *Science* **382**, 669–74.

Ćevid, D., Michel, L., Näf, J., Bühlmann, P. & Meinshausen, N. (2022). Distributional random forests: heterogeneity adjustment and multivariate distributional regression. *J. Mach. Learn. Res.* **23**, 1–79.

Chakrabortty, A., Dai, G. & Carroll, R. J. (2024). Semi-supervised quantile estimation: robust and efficient inference in high dimensional settings. *arXiv:* 2201.10208v2.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Economet. J.* **21**, C1–68.

Chernozhukov, V., Fernández-Val, I. & Galichon, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika* **96**, 559–75.

Chernozhukov, V., Fernández-Val, I. & Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica* **78**, 1093–125.

Christgau, A. M., Petersen, L. & Hansen, N. R. (2023). Nonparametric conditional local independence testing. *Ann. Statist.* **51**, 2116–44.

Donsker, M. D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* **23**, 277–81.

Dvoretzky, A., Kiefer, J. & Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **30**, 642–69.

Hall, P. & Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *Ann. Statist.* **33**, 1404–21.

Henzi, A., Kleger, G.-R. & Ziegel, J. F. (2021). Distributional (single) index models. *J. Am. Statist. Assoc.* **118**, 489–503.

Hofner, B., Mayr, A. & Schmid, M. (2016). gamboostLSS: an R package for model building and variable selection in the GAMLSS framework. *J. Statist. Softw.* **74**, 1–31.

Kneib, T., Silbersdorff, A. & Säfken, B. (2023). Rage against the mean – a review of distributional regression approaches. *Economet. Statist.* **26**, 99–123.

Shen, X. & Meinshausen, N. (2024). Engression: extrapolation through the lens of distributional regression. *arXiv:* 2307.00835v3.

Song, S., Lin, Y. & Zhou, Y. (2024). A general M-estimation theory in semi-supervised framework. *J. Am. Statist. Assoc.* **119**, 1065–75.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Yuval, O. & Rosset, S. (2022). Semi-supervised empirical risk minimization: using unlabeled data to improve prediction. *Electron. J. Statist.* **16**, 1434–60.

Zhang, A., Brown, L. D. & Cai, T. T. (2019). Semi-supervised inference: general theory and estimation of means. *Ann. Statist.* **47**, 2538–66.

Zhang, Y. & Bradic, J. (2022). High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika* **109**, 387–403.

Zhang, Y., Chakrabortty, A. & Bradic, J. (2023). Double robust semi-supervised inference for the mean: selection bias under MAR labeling with decaying overlap. *Info. Infer.* **12**, 2066–159.

Zrnic, T. & Candès, E. J. (2024). Cross-prediction-powered inference. *Proc. Nat. Acad. Sci. USA* **121**, e2322083121.

[*Received on* 18 *February* 2024. *Editorial decision on* 30 *September* 2024]