Statistica Sinica Preprint No: SS-2025-0406	
Title	Distributed Learning of Single Index Model with Differential Privacy and Massive Data
Manuscript ID	SS-2025-0406
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0406
Complete List of Authors	Lingrui Wang Fode Zhang, and Hua Liang
Corresponding Authors	Fode Zhang
E-mails	fredzh@swufe.edu.cn
Notice: Accepted author version.	

Statistica Sinica: Newly accepted Paper

Statistica Sinica

Distributed Learning of Single-Index Model with Differential Privacy and Massive Data

Lingrui Wang[1], Fode Zhang[1], Hua Liang[2]

[1] *Southwestern University of Finance and Economics*

[2] *The George Washington University*

Abstract: Distributed learning enables scalable data analysis by reducing storage and computational burdens, although the communication of local outputs in divide-and-conquer (DC) implementations may introduce privacy risks. We study distributed estimation and inference for the single-index model (SIM), a flexible semiparametric framework with an interpretable low-dimensional index structure. We propose a DC estimator for the SIM and establish convergence rates and error bounds for the nonparametric component, together with asymptotic normality of the index parameter estimator. Under mild conditions, the aggregated estimator attains the optimal nonparametric convergence rate $O_{\mathbb{P}}\left(N^{-p/(2p+1)}\right)$. To provide rigorous privacy protection, we further develop differentially private (DP) algorithms for the SIM under the DC framework that satisfy (ε, δ)-DP. We establish a high-probability bound on the average squared gradient mapping, thereby characterizing how privacy noise and model complexity affect optimization accuracy. Simulations and a real-data application demonstrate accurate estimation and effective privacy protection.

Statistica Sinica: Newly accepted Paper

Key words and phrases: Distributed learning; Differential privacy; Single-index model; Optimal convergence rate; Asymptotic properties

1. Introduction

The SIM captures the effect of a multivariate predictor through a one-dimensional index obtained by projecting the covariates onto an unknown direction. As a result, SIM serves as an effective dimension-reduction tool in regression (Dong et al., 2016). Compared with fully nonparametric regression, SIM alleviates the curse of dimensionality and can be viewed as a flexible extension of the linear model with an unknown link function (Cui et al., 2011; Huang et al., 2020; Han et al., 2023). By retaining a low-dimensional index structure, SIM offers a useful compromise among modeling flexibility, interpretability, and simplicity (Lian et al., 2014).

SIM has been widely used in applications such as finance, economics, and biomedical studies, where complex relationships may exist but interpretability remains important (Horowitz, 2009; Han et al., 2023). From a methodological standpoint, a variety of estimators have been developed for the SIM, including B-spline-based estimation (Wang and Yang, 2009), least-squares estimation for monotone SIM (Balabdaoui et al., 2019), and penalized smoothing-spline approaches (Kuchibhotla and Patra, 2020). De-

Statistica Sinica: Newly accepted Paper

spite this extensive literature, less attention has been paid to SIM under modern distributed computing constraints, especially to algorithm design with rigorous differential privacy guarantees. This motivates the developments in this paper.

In the era of big data, datasets are often too large for storage and analysis on a single machine, which has motivated extensive work on distributed estimation and inference. A widely used baseline is DC, where local estimators are computed in parallel and then aggregated into a global estimator. Beyond one-shot aggregation, a broader literature studies communication-efficient distributed inference, with the goal of reducing both the number of communication rounds and the communication load by transmitting low-dimensional summaries or surrogate objectives (Jordan et al., 2019; Bao and Xiong, 2021; Fan et al., 2023). In this work, we adopt DC as the primary implementation because of its simplicity and scalability. For the SIM, constructing a suitable surrogate loss is challenging because its semiparametric structure involves the simultaneous estimation of the index parameter and the unknown link function.

The DC paradigm partitions a sample of size N across s machines, computes local estimators in parallel, and aggregates them on a central server to form a global estimator (Tu et al., 2024). DC-based estimation

and inference have been studied in many settings, including linear regression (Chen and Xie, 2014), kernel ridge regression (Zhang et al., 2015), and single-index models with massive data (Jiang and Peng, 2023). More recent work has further developed distributed inference under DC, addressing issues such as cross-machine heterogeneity and valid inference in more complex semiparametric settings (Gao et al., 2022; Gu and Chen, 2023; Chen et al., 2024).

However, distributed computation under the DC framework is not, by itself, a privacy-preserving mechanism. Although raw individual-level data are not directly shared with the central server, the information exchanged between local machines and the central server, such as local estimators and other summary statistics, may still reveal sensitive information. Therefore, avoiding data centralization alone does not provide a formal privacy guarantee. In our work, the single-index model is adopted for its modeling flexibility as a semiparametric alternative to fully parametric specifications, rather than for any inherent privacy advantage. To provide explicit and quantifiable protection for the communicated quantities in distributed estimation, we incorporate differential privacy (Dwork et al., 2006; Dong et al., 2022).

The combination of differential privacy and DC is particularly relevant

in multi-institution data analyses, where individual-level data often cannot be centralized because of regulatory, legal, or contractual constraints. Examples include multi-center clinical studies, hospital networks, and other cross-site collaborations in which data sharing is restricted but joint statistical analysis is still required. In such settings, distributed estimation is needed because raw data cannot be pooled across sites, while privacy protection is needed because the released summaries may still reveal sensitive information. More broadly, this motivation is aligned with the growing literature on differentially private statistical inference and data analysis, including work on parametric inference (Avella-Medina, 2021), summary-statistic release (Barrientos et al., 2024), regression analysis (Cai et al., 2021; Györfi and Kroll, 2025), network data analysis (Yan, 2021), and private hypothesis testing (Awan and Vadhan, 2023).

To provide differential privacy for the SIM in the distributed computing framework, we adopt a joint design that uses a single privacy budget for the joint release of the index and nonparametric estimates. This unified formulation simplifies privacy accounting and avoids splitting the privacy budget across multiple subproblems. In particular, each local machine communicates only a differentially private version of its final local estimates by adding appropriately calibrated noise before transmission, and the cumu-

Statistica Sinica: Newly accepted Paper

lative privacy loss is tracked through standard DP composition arguments. This design leads to a practical and scalable privacy-preserving procedure while maintaining competitive statistical accuracy.

1.1 Our contributions

This paper proposes a distributed estimation procedure and differentially private algorithms for the SIM under the DC framework, with the goal of enabling large-scale inference under formal privacy guarantees. Our main contributions are as follows.

- The proposed method is straightforward to implement and communication-efficient. The global estimator is obtained via one-shot aggregation of local estimators, thereby avoiding multiple rounds of communication between the central server and local nodes.

- We develop a differentially private estimation procedure for the SIM in the distributed setting. Leveraging the model structure, we employ a joint noisy gradient scheme that updates the index parameter and the nonparametric component within a unified privacy budget, using Gaussian perturbations and privacy accounting via the moments accountant. The resulting algorithm satisfies (ε, δ)-differential privacy under regularity conditions and admits a high-probability bound on

Statistica Sinica: Newly accepted Paper

the average squared norm of the gradient mapping, thereby providing an approximate stationarity guarantee for the empirical loss.

- We establish theoretical guarantees for the proposed estimator. In particular, we derive convergence rates for the nonparametric component and prove asymptotic normality for the parametric component under the scaling condition $n \gg N^{5/8}$, where N denotes the total sample size and n denotes the within-subpopulation sample size.

- Simulation studies under both distributed and differentially private settings, as well as a real-data application, show that the proposed method maintains good estimation accuracy while ensuring rigorous privacy protection for the SIM.

1.2 Organization

The rest of the paper is organized as follows. Section 2 introduces the estimator under the DC framework and establishes its theoretical properties. Section 3 develops differentially private algorithms for the SIM and studies their theoretical guarantees. Section 4 reports simulation results and a real-data application. Section 5 concludes with a discussion. Proofs and additional experimental results are deferred to the Appendix.

Statistica Sinica: Newly accepted Paper

1.3 Notation

We use the following notation. For a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|$ and $\|\boldsymbol{x}\|_\infty$ denote the Euclidean and ℓ_∞ norms, respectively; for a matrix $\boldsymbol{C}$, $\|\boldsymbol{C}\|_2$ denotes the spectral norm. For L_2-integrable functions φ_1 and φ_2, define $\langle \varphi_1, \varphi_2 \rangle = E\{\varphi_1(U)\varphi_2(U)\}$, where U is a generic real-valued random variable, and let $\|\varphi\|_2 = \sqrt{\langle \varphi, \varphi \rangle}$. For sequences $\{a_n\}$ and $\{b_n\}$, write $a_n \lesssim b_n$, $a_n \gtrsim b_n$, and $a_n \asymp b_n$ in the usual sense. We also write $a_n = O(b_n)$ if a_n/b_n is bounded and $a_n = o(b_n)$ if $a_n/b_n \to 0$. For $\boldsymbol{\nu} \in \mathbb{R}^d$ and $R > 0$, let $\Pi_R(\boldsymbol{\nu})$ denote the projection of $\boldsymbol{\nu}$ onto the Euclidean-ball $\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\| \le R\}$. Finally, $O_{\mathbb{P}}(\cdot)$ and $o_{\mathbb{P}}(\cdot)$ denote the usual stochastic order symbols in probability.

2. Distributed Estimation of Single-Index Model

2.1 Single-index model

Let $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^N$ denote the full sample, which is evenly partitioned across s machines. Let G_j denote the index set on the j-th machine, with $|G_j| = n$ and $N = sn$. For $i \in G_j$, the observations follow the single-index model

$$Y_i = f_0\big(\boldsymbol{\alpha}_0^\top \boldsymbol{X}_i\big) + \varepsilon_i, \qquad i \in G_j, \quad j = 1, \ldots, s. \tag{2.1}$$

Here f_0 is an unknown smooth function, and $\boldsymbol{\alpha}_0 \in \mathbb{R}^d$ is an unknown index parameter. The noise terms $\{\varepsilon_i\}$ satisfy $E(\varepsilon_i \mid \boldsymbol{X}_i) = 0$.

Statistica Sinica: Newly accepted Paper

For the estimation of unknown parameters, we impose the identifiability constraints $\|\boldsymbol{\alpha}_0\| = 1$ and $\alpha_{01} > 0$. To avoid constrained optimization on the unit sphere, we use the delete-one-component method (Yu and Ruppert, 2002). Specifically, let $\boldsymbol{\tau}_0 = (\alpha_{02}, \ldots, \alpha_{0d})^\top \in \mathbb{R}^{d-1}$. Then $\alpha_{01} = \sqrt{1 - \|\boldsymbol{\tau}_0\|^2}$, and we define $\boldsymbol{\alpha}(\boldsymbol{\tau}) = (\sqrt{1 - \|\boldsymbol{\tau}\|^2}, \boldsymbol{\tau}^\top)^\top$ for $\|\boldsymbol{\tau}\| < 1$. Therefore, estimating $\boldsymbol{\alpha}_0 \in \mathbb{R}^d$ under $\|\boldsymbol{\alpha}_0\| = 1$ and $\alpha_{01} > 0$ is equivalent to estimating $\boldsymbol{\tau}_0 \in \mathbb{R}^{d-1}$.

The Jacobian matrix of $\boldsymbol{\alpha}(\boldsymbol{\tau})$ is the $d \times (d-1)$ matrix

$$\mathrm{J}(\boldsymbol{\tau}) = \begin{bmatrix} -(1 - \|\boldsymbol{\tau}\|^2)^{-1/2}\boldsymbol{\tau}^\top \\ \boldsymbol{I}_{(d-1)\times(d-1)} \end{bmatrix}.$$

Then the single-index model (2.1) can be reparametrized as

$$Y_i = f_0\big(\boldsymbol{\alpha}^\top(\boldsymbol{\tau}_0)\boldsymbol{X}_i\big) + \varepsilon_i, \qquad i \in G_j, \quad j = 1, \ldots, s. \tag{2.2}$$

For the nonparametric component, we approximate the smooth function f_0 using a polynomial spline. Let $\mathcal{S}_n$ denote the space of polynomial splines on $[a, b]$ of degree $\rho \geq 1$. We consider a sequence of J_n interior knots:

$$t_{-\rho} = t_{-\rho+1} = \cdots = t_0 = a < t_1 < \cdots < t_{J_n} < b = t_{J_n+1} = \cdots = t_{J_n+\rho+1},$$

where J_n increases with the sample size n.

Statistica Sinica: Newly accepted Paper

2.2 Assumptions

We assume that $\{\boldsymbol{X}_i\}_{i=1}^N$ are i.i.d. samples from a random vector $\boldsymbol{X}$. Define $\Phi(\boldsymbol{x}) = E\left[\boldsymbol{X} \mid \boldsymbol{\alpha}_0^\top \boldsymbol{X} = \boldsymbol{\alpha}_0^\top \boldsymbol{x}\right]$ for $\boldsymbol{x} \in \mathbb{R}^d$. Let $\tilde{\boldsymbol{X}}_i = \boldsymbol{X}_i - \Phi(\boldsymbol{X}_i)$ and $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \Phi(\boldsymbol{X})$. For any matrix or vector $\boldsymbol{C}$, define $\boldsymbol{C}^{\otimes 2} = \boldsymbol{C}\boldsymbol{C}^\top$. Let r be a positive integer and $v \in (0, 1]$ such that $p = r + v > 2$. Let $\mathcal{H}$ be the collection of function f on $[a, b]$ whose r-th derivative exists and satisfies the Lipschitz condition of order v:

$$|f^{(r)}(x) - f^{(r)}(x')| < L|x - x'|^v, \quad \text{for any } x, x' \in [a, b], \tag{2.3}$$

where L is a positive constant. In order to derive asymptotic results, the following assumptions are required:

(A1) The function f_0 satisfies the Lipschitz condition (2.3), i.e., $f_0 \in \mathcal{H}$.

(A2) The r-th order partial derivatives of the joint density function of $\boldsymbol{X}$ satisfy the Lipschitz condition of order v. The distribution of $\boldsymbol{\alpha}^\top \boldsymbol{X}$ is absolutely continuous, and its density function is bounded away from zero and infinity on $[a, b]$.

(A3) For any $\boldsymbol{\tau}$, there exist constants $0 < c < C < \infty$ such that

$$c\boldsymbol{I}_{d+1} \leq E\left[\begin{pmatrix} \boldsymbol{X} \\ 1 \end{pmatrix}\begin{pmatrix} \boldsymbol{X} \\ 1 \end{pmatrix}^\top \middle| \boldsymbol{\alpha}^\top(\boldsymbol{\tau})\boldsymbol{X} = \boldsymbol{\alpha}^\top(\boldsymbol{\tau})\boldsymbol{x}\right] \leq C\boldsymbol{I}_{d+1}.$$

Statistica Sinica: Newly accepted Paper

(A4) The number of knots J_n satisfies $n^{1/2(p+1)} \ll J_n \ll n^{1/5}$ for $p = r+v > 2$.

(A5) There exists a positive constant C_0 such that $E(\varepsilon^2 \mid \boldsymbol{\alpha}_0^\top \boldsymbol{X}) \leq C_0$.

Condition (A1) imposes a smoothness requirement on $f_0(\cdot)$ which determines the approximation error of the spline space $\mathcal{S}_n$. Condition (A2) requires a smoothness condition on the joint and marginal density functions of the covariates, which is often assumed in the asymptotic analysis of nonparametric regression problems. Condition (A3) requires the conditional second-moment matrix in (A3) to have eigenvalues bounded away from 0 and ∞. This condition is useful for establishing the local nonsingularity of the Hessian matrix and the asymptotic normality of the parametric component. Condition (A4) gives the growth rate of the dimension of the spline spaces relative to the sample size. These assumptions are commonly used in the analysis of the asymptotic properties of semiparametric models.

2.3 Estimation Based on Each Subpopulation

For each $j = 1, \ldots, s$, we obtain the local estimators of $\boldsymbol{\tau}_0$ and f_0 by minimizing the empirical loss over the observations indexed by G_j:

$$(\hat{\boldsymbol{\tau}}^{(j)}, \hat{f}^{(j)}) = \underset{\boldsymbol{\tau} \in \mathbb{R}^{d-1}, f \in \mathcal{S}_n}{\arg\min} \left\{ L_n^{(j)} = \frac{1}{2n} \sum_{i \in G_j} \left(Y_i - f(\boldsymbol{\alpha}^\top(\boldsymbol{\tau}) \boldsymbol{X}_i) \right)^2 \right\}. \quad (2.4)$$

Statistica Sinica: Newly accepted Paper

Let $b_{m,\rho}(u), m = 1, \ldots, K_n$ be the B-spline basis functions of degree ρ with $K_n = J_n + \rho + 1$. For any $f \in \mathcal{H}$, the function f can be approximated by $f(u) = \boldsymbol{\gamma}^\top \boldsymbol{B}(u)$ (De Boor, 1978), where $\boldsymbol{B}(u) = (b_{1,\rho}(u), \ldots, b_{K_n,\rho}(u))^\top$ is the spline basis, and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_{K_n})^\top$ is the spline coefficient vector. Then, the Eq. (2.4) becomes

$$(\hat{\boldsymbol{\tau}}^{(j)}, \hat{\boldsymbol{\gamma}}^{(j)}) = \underset{\boldsymbol{\tau}\in\mathbb{R}^{d-1}, \boldsymbol{\gamma}\in\mathbb{R}^{K_n}}{\arg\min} \left\{ l_n^{(j)} = \frac{1}{2n} \sum_{i \in G_j} \left(Y_i - \boldsymbol{\gamma}^\top \boldsymbol{B}(\boldsymbol{\alpha}^\top(\boldsymbol{\tau}) \boldsymbol{X}_i) \right)^2 \right\}. \quad (2.5)$$

Remark 1. We solve (2.5) using an alternating algorithm. Specifically, given the current index $\boldsymbol{\tau}$, we update the spline coefficients for the non-parametric component, and then update $\boldsymbol{\tau}$ by a quasi-Newton-type minimization of $l_n^{(j)}(\boldsymbol{\tau}, \boldsymbol{\gamma})$, followed by a step to ensure that $\|\boldsymbol{\tau}\| < 1$. The iterations terminate when the successive changes in both $\boldsymbol{\tau}$ and the spline coefficients are below a prescribed tolerance; see Appendix S1.1.

This leads to the plug-in estimator $\hat{f}^{(j)}(u) = \left(\hat{\boldsymbol{\gamma}}^{(j)}\right)^\top \boldsymbol{B}(u)$, which admits the following theorem.

Theorem 1. *Under Assumptions (A1)–(A5), for each subpopulation, $j = 1, ..., s$, we have*

$$\left\| \hat{f}^{(j)} - f_0 \right\|_2 = O_{\mathbb{P}}(J_n^{1/2} n^{-1/2} + h^p),$$

where $h = 1/(J_n + 1)$ is the distance between the neighboring knots.

Statistica Sinica: Newly accepted Paper

The proof can be found in the supplementary material. Theorem 1 indicates that the convergence rate of the estimator for the nonparametric part of each subpopulation is related to the sample size of each subpopulation and the number of knots J_n. Let β be a positive number between 0 and 1. Assume that $s = O(N^{1-\beta})$ and $n = O(N^{\beta})$, to achieve the optimal convergence rate, we select the optimal number of knots $O(n^{\frac{1}{2p+1}})$. This choice yields the optimal convergence rate of $O_{\mathbb{P}}(n^{-\frac{p}{2p+1}})$, or equivalently $O_{\mathbb{P}}(N^{-\frac{p\beta}{2p+1}})$.

Under the same assumptions, we establish the asymptotic normality of the estimators of $\boldsymbol{\tau}_0$ and $\boldsymbol{\alpha}_0$.

Theorem 2. *Let $U_{\tau,0} = \boldsymbol{\alpha}^\top(\boldsymbol{\tau}_0)\boldsymbol{X}$. Under Assumptions (A1)–(A5), for each subpopulation, we have*

$$\sqrt{n}(\hat{\boldsymbol{\tau}}^{(j)} - \boldsymbol{\tau}_0) \to \mathcal{N}(\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}), \quad j = 1, ..., s,$$

and

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}^{(j)} - \boldsymbol{\alpha}_0) \to \mathcal{N}(\mathbf{0}, \mathrm{J}(\boldsymbol{\tau}_0)\boldsymbol{A}^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\mathrm{J}^\top(\boldsymbol{\tau}_0)),$$

where

$$\boldsymbol{A} = E\left[\{f_0'\}^2(U_{\tau,0})\mathrm{J}^\top(\boldsymbol{\tau}_0)\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top\mathrm{J}(\boldsymbol{\tau}_0)\right], \quad \boldsymbol{\Sigma} = E\left[\{\varepsilon f_0'(U_{\tau,0})\mathrm{J}^\top(\boldsymbol{\tau}_0)\tilde{\boldsymbol{X}}\}^{\otimes 2}\right].$$

The proof can be found in the supplementary material. Theorem 2

Statistica Sinica: Newly accepted Paper

indicates that the estimator for the parametric part of each subpopulation $\hat{\boldsymbol{\tau}}^{(j)}$ or $\hat{\boldsymbol{\alpha}}^{(j)}$ is $\sqrt{n}$-consistent and is asymptotically normal.

2.4 Aggregation of Local Estimators

We aggregate the local estimators from s subpopulations with total sample size $N = sn$, and establish theoretical guarantees for both the nonparametric and parametric components. For the nonparametric component, we use the aggregated estimator $\bar{f} = \frac{1}{s}\sum_{j=1}^{s} \hat{f}^{(j)}$. The following theorem establishes the convergence rate of the aggregated estimator $\bar{f}$.

Theorem 3. *Under Assumptions (A1)–(A5), then we have*

$$\left\|\bar{f} - f_0\right\|_2 = O_{\mathbb{P}}(J_n^{1/2} N^{-1/2} + h^p).$$

The proof can be found in the supplementary material. Theorem 3 shows that, if $n \gg N^{5/(2p+1)}$, choosing the number of knots such that $J_n = O(N^{1/(2p+1)})$ yields the optimal convergence rate $O_{\mathbb{P}}\left(N^{-p/(2p+1)}\right)$.

For the parametric component, we focus on the aggregated estimator $\bar{\boldsymbol{\tau}} = \frac{1}{s}\sum_{j=1}^{s} \hat{\boldsymbol{\tau}}^{(j)}$ as the final estimator of $\boldsymbol{\tau}$. The corresponding estimator of the original index parameter is defined as $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}(\bar{\boldsymbol{\tau}})$. Under stronger assumptions on the numbers n and J_n, we have the asymptotic normality of the aggregated estimators.

Statistica Sinica: Newly accepted Paper

Theorem 4. *Under Assumption(A1)–(A5), if* $n \gg N^{5/8}$, *then we have*

$$\sqrt{N}(\bar{\boldsymbol{\tau}} - \boldsymbol{\tau}_0) \to \mathcal{N}(\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}),$$

and

$$\sqrt{N}(\bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \to \mathcal{N}(\mathbf{0}, \mathrm{J}(\boldsymbol{\tau}_0)\boldsymbol{A}^{-1}\boldsymbol{\Sigma}\boldsymbol{A}^{-1}\mathrm{J}^\top(\boldsymbol{\tau}_0)),$$

where

$$\boldsymbol{A} = E\left[\{f_0'\}^2(U_{\tau,0})\mathrm{J}^\top(\boldsymbol{\tau}_0)\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top\mathrm{J}(\boldsymbol{\tau}_0)\right], \quad \boldsymbol{\Sigma} = E\left[\{\varepsilon f_0'(U_{\tau,0})\mathrm{J}^\top(\boldsymbol{\tau}_0)\tilde{\boldsymbol{X}}\}^{\otimes 2}\right].$$

The proof can be found in the supplementary material. Theorem 4 implies that $\bar{\boldsymbol{\tau}}$ and $\bar{\boldsymbol{\alpha}}$ are $\sqrt{N}$-consistent and asymptotically normal provided that $J_n \gg n^{1/2(p+1)}$ and $n \gg N^{5/8}$. Compared with Theorem 3, this result requires a stronger growth condition on the local sample size.

Remark 2. Recall that $N = sn$ and $N \to \infty$. The number of machines s may be fixed or may diverge with N, provided that the local sample size $n = N/s$ satisfies the growth conditions in Theorems 3 and 4. In particular, the condition $n \gg N^{5/(2p+1)}$ in Theorem 3 is equivalent to $s = o\left(N^{2(p-2)/(2p+1)}\right)$, while the condition $n \gg N^{5/8}$ in Theorem 4 is equivalent to $s = o(N^{3/8})$. Therefore, s need not remain fixed for the asymptotic results, although its allowable growth rate depends on the smoothness parameter p and the inferential objective. In practice, increasing s improves

computational scalability but reduces the local sample size n, which may deteriorate statistical accuracy and invalidate the asymptotic guarantees if s grows too quickly. Thus, the choice of s should balance computational efficiency and inferential accuracy, with particular attention to whether the primary goal is nonparametric estimation or inference for the parametric component.

3. Estimation with Differential Privacy

In the DC setting, the quantities communicated from local machines to the central server, such as local estimators and aggregated summaries, could still reveal sensitive information about individual records. To provide formal and quantifiable privacy protection for the communicated outputs, we develop a differentially private procedure for computing each local estimator before aggregation. For completeness, we first review several basic definitions and results from differential privacy.

3.1 Differential privacy

Differential privacy provides a formal framework for quantifying the privacy risk of a randomized algorithm. Roughly speaking, it requires that the distribution of the algorithm's output remain nearly unchanged when

Statistica Sinica: Newly accepted Paper

one individual record in the input dataset is modified. We begin with the definition of (ε, δ)-differential privacy.

Definition 1 (Differential privacy (Dwork et al., 2006))**.** A randomized algorithm $\mathcal{M} : \mathbb{X}^n \to \mathbb{R}^d$ is said to be (ε, δ)-differentially private if for every pair of adjacent datasets $X, X' \in \mathbb{X}^n$ differing in one individual record, and every measurable set $S \subseteq \mathbb{R}^d$,

$$\mathbb{P}\{\mathcal{M}(X) \in S\} \leq e^{\varepsilon}\mathbb{P}\{\mathcal{M}(X') \in S\} + \delta,$$

where the probability is taken over the randomness of $\mathcal{M}$.

This definition implies that, for small values of ε and δ, the output distributions under adjacent datasets are nearly indistinguishable. Smaller values of ε and δ correspond to stronger privacy protection. We next introduce the Gaussian mechanism, which will be used throughout this section.

Definition 2 (Gaussian Mechanism (Dwork et al., 2006))**.** Let $M : \mathbb{X}^n \to \mathbb{R}^d$ be any (non-private) algorithm. Suppose its ℓ_2-sensitivity is bounded as $\sup_{X \sim X'} \|M(X) - M(X')\|_2 \leq B$ for all neighboring datasets X, X'. The Gaussian mechanism is defined by $\mathcal{M}(X) = M(X) + \boldsymbol{\omega}$, where $\boldsymbol{\omega} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{d\times d})$, and is (ε, δ)-differentially private when $\sigma^2 \geq 2B^2 \log(2/\delta)/\varepsilon^2$.

To analyze the cumulative privacy loss, we use standard privacy accounting tools based on composition and Rényi differential privacy. We

Statistica Sinica: Newly accepted Paper

then develop a differentially private estimation algorithm for the SIM on each local machine under the DC framework.

3.2 Differentially private algorithms for the single-index model

The SIM admits two natural approaches for differentially private estimation: a joint design using a single privacy budget for the entire procedure and a componentwise design that privatizes the parametric and nonparametric components separately. In our setting, the componentwise design is less suitable because alternating optimization complicates privacy-budget allocation and noise calibration across subproblems, and the nonconvex update for the index parameter can become the main computational bottleneck. Therefore, we adopt a joint design in which both components are updated under a unified differentially private mechanism. This choice simplifies privacy accounting and allows consistent calibration of the injected noise. Specifically, we employ a differentially private gradient descent (DP-GD) procedure, where Gaussian perturbations are added to the gradient updates. This approach is well suited for nonconvex objectives and provides utility guarantees (Wang et al., 2017).

For the j-th subpopulation, we consider the empirical loss

$$L_n^{(j)}(\boldsymbol{\gamma}, \boldsymbol{\tau}) = \frac{1}{2n} \sum_{i \in G_j} \left\{ Y_i - \boldsymbol{\gamma}^\top \boldsymbol{B} \left(\boldsymbol{\alpha}^\top(\boldsymbol{\tau}) \boldsymbol{X}_i \right) \right\}^2,$$

Statistica Sinica: Newly accepted Paper

and aim to estimate both the projection parameter $\boldsymbol{\tau}$ and the spline coefficient vector $\boldsymbol{\gamma}$ in a privacy-preserving manner. We use a joint noisy gradient descent procedure, where Gaussian noise is calibrated according to the sensitivity of the optimization steps; see Algorithm 1 for details.

Algorithm 1 Differentially Private Algorithm for the SIM

Input: Local dataset $\{(Y_i, \boldsymbol{X}_i) : i \in G_j\}$, initial values $\boldsymbol{\tau}_0 \in \mathbb{R}^{d-1}$, $\boldsymbol{\gamma}_0 \in \mathbb{R}^{J_n+\rho+1}$, truncation level R, c_γ, c_τ, step size η, privacy budgets (ε, δ), and noise scales σ_θ, total rounds T.

1: **for** $t = 0$ to $T-1$ **do**

2: Compute $\hat{f}_t(z) := \boldsymbol{\gamma}_t^\top \boldsymbol{B}(z)$, $\hat{f}_t'(z) := \boldsymbol{\gamma}_t^\top \boldsymbol{B}'(z)$ and $u_{ti} = \boldsymbol{\alpha}(\boldsymbol{\tau}_t)^\top \boldsymbol{X}_i$.

3: Sample noises $\boldsymbol{\omega}_t^{(\tau)} \sim \mathcal{N}(0, \sigma_\theta^2 I_{d-1})$ and $\boldsymbol{\omega}_t^{(\gamma)} \sim \mathcal{N}(0, \sigma_\theta^2 I_{J_n+\rho+1})$.

4: // **Update** $\boldsymbol{\theta} = \binom{\gamma}{\boldsymbol{\tau}}$

5: Compute $\boldsymbol{\gamma}_{t+1} = \Pi_{c_\gamma}\Big(\boldsymbol{\gamma}_t - \eta\Big[\frac{1}{n}\sum_{i\in G_j}.(\hat{f}_t(u_{ti}) - \Pi_R(Y_i))\,\boldsymbol{B}(u_{ti}) + \boldsymbol{\omega}_t^{(\gamma)}\Big]\Big)$

$\boldsymbol{\tau}_{t+1} = \Pi_{c_\tau}\Big(\boldsymbol{\tau}_t - \eta\Big[\frac{1}{n}\sum_{i\in G_j}.(\hat{f}_t(u_{ti}) - \Pi_R(Y_i))\,\hat{f}_t'(u_{ti})\,\mathrm{J}^\top(\boldsymbol{\tau}_t)\boldsymbol{X}_i + \boldsymbol{\omega}_t^{(\tau)}\Big]\Big)$

6: **end for**

Output: Sample $t \sim \mathrm{Unif}\{0, \ldots, T-1\}$ and return $\boldsymbol{\theta}_{\mathrm{out}} = \binom{\gamma_t}{\boldsymbol{\tau}_t}$.

Before analyzing convergence properties of the resulting estimators, we introduce the following technical assumptions:

(B1) The covariate vector is almost surely bounded: $\|\boldsymbol{X}_i\| \le c_X$.

Statistica Sinica: Newly accepted Paper

(B2) The true index parameter satisfies $\|\boldsymbol{\tau}_0\| \leq \kappa < 1$ and its Jacobian matrix satisfies $\sup_{\|\boldsymbol{\tau}\|\leq\kappa} \|\mathrm{J}(\boldsymbol{\tau})\|_2 \leq C_\kappa$, $\sup_{\|\boldsymbol{\tau}\|\leq\kappa} \left\| \frac{\partial \mathrm{J}(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}} \right\|_2 \leq C'_\kappa$.

(B3) The true regression function $f_0 \in \mathcal{H}$ is uniformly bounded on $[a, b]$, i.e., $|f_0(u)| \leq Q$, and the noisy $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is independent of $\boldsymbol{X}$.

(B4) The spline basis vector $\boldsymbol{B}(u)$ and its first and second derivatives are uniformly bounded on $[a, b]$: $\|\boldsymbol{B}(u)\| \leq c_b$, $\|\boldsymbol{B}'(u)\| \leq c'_b$, and $\|\boldsymbol{B}''(u)\| \leq c''_b$.

(B5) The optimal approximation coefficient satisfies $\|\tilde{\boldsymbol{\gamma}}\| \leq c_0$.

Assumptions (B1)–(B5) establish standard regularity conditions for single-index models in the context of differentially private estimation. Specifically, (B1)–(B2) ensure the boundedness of the covariates and index parameters, which is crucial for identifiability and privacy calibration. (B3)–(B4) ensure the link function and the spline basis are uniformly bounded, providing the tight sensitivity control required for the calibration of DP noise. Finally, (B5) bounds the coefficient vector, further constraining the model's sensitivity. Collectively, these conditions help control sensitivity and support both privacy guarantees and utility analysis.

Under these conditions, we now state a lemma establishing the Lipschitz continuity and smoothness of the empirical loss, which plays a key role in

Statistica Sinica: Newly accepted Paper

both convergence analysis and privacy calibration.

Lemma 1. *Suppose Assumptions (B1)–(B5) hold,* $|Y_i| \le R$ *almost surely, and* $\|\boldsymbol{\tau}^{(j)}\| \le \kappa < 1$, $\|\boldsymbol{\gamma}^{(j)}\| \le c_0$. *Then the empirical loss on the* j*-th machine is jointly* L_θ*-Lipschitz continuous and* M_θ*-smooth with respect to* $\boldsymbol{\theta} := \binom{\gamma}{\boldsymbol{\tau}}$, *where*

$$L_\theta = (c_0 c_b + R) c_0 c_b' C_\kappa c_X,$$

$$M_\theta = \max\{c_b^2, (c_0 c_b')^2 C_\kappa^2 c_X^2 + (c_b c_0 + R) c_0 c_b'' C_\kappa^2 c_X^2 + (c_b c_0 + R) c_0 c_b' C_\kappa' c_X^2\}$$
$$+ c_0 c_b' C_\kappa c_X c_b + (c_b c_0 + R) C_\kappa c_X c_b'.$$

The proof is deferred to the supplementary material. Under standard regularity conditions, Lemma 1 shows that the empirical loss on each machine is jointly Lipschitz continuous and smooth in $\boldsymbol{\theta}$. We next state the corresponding privacy guarantee.

Theorem 5. *Suppose that the noise variance satisfies* $\sigma_\theta^2 \ge C \frac{L_\theta^2 T \log(1/\delta)}{n^2 \varepsilon^2}$. *Then Algorithm 1 guarantees* (ε, δ)*-differential privacy.*

The proof is provided in the supplementary material and is based on composing the privacy guarantees of the noisy gradient updates via the moments accountant. The privacy guarantee for the aggregated estimator follows from the parallel composition and post-processing properties of

Statistica Sinica: Newly accepted Paper

differential privacy. We next present the convergence result for the differentially private algorithm.

Theorem 6. *Let* $\{(\boldsymbol{X}_i, Y_i) : i \in G_j\}$ *be the local sample on the* j*-th machine, where* $|G_j| = n$*. Let the step size* $\eta = 1/M_\theta$*, projection radius* $R \lesssim \sigma\sqrt{\log n}$*,* $c_\gamma = c_0$*,* $c_\tau = \kappa$*, initial value* $\boldsymbol{\gamma}_0 = \mathbf{0}$*, iteration number* $T = \frac{\sqrt{M_\theta} n\varepsilon}{L_\theta\sqrt{\log(1/\delta)}}$ *and noise scale* $\sigma_\theta^2 \geq C\frac{L_\theta^2 T\log(1/\delta)}{n^2\varepsilon^2}$*. If* $n \gtrsim \frac{L_\theta\sqrt{M_\theta\log n}}{\varepsilon\sqrt{\log(1/\delta)}}$*, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left\|G_\eta(\boldsymbol{\theta}_t^{(j)})\right\|^2 \lesssim \frac{L_\theta\sqrt{M_\theta\log(1/\delta)}}{n\varepsilon} + \frac{L_\theta(d+J_n+\rho)\sqrt{M_\theta\log(1/\delta)}}{n\varepsilon}$$

with probability at least $1 - c_2\exp(-c_3 n) - \exp(-c_1\sqrt{T}(d+J_n+\rho))$*, where* $G_\eta(\boldsymbol{\theta}_t^{(j)})$ *is defined as*

$$G_\eta(\boldsymbol{\theta}_t^{(j)}) := \frac{1}{\eta}\left(\boldsymbol{\theta}_t^{(j)} - \Pi_{\mathcal{C}}\left(\boldsymbol{\theta}_t^{(j)} - \eta\nabla_\theta L_n^{(j)}(\boldsymbol{\theta}_t^{(j)})\right)\right).$$

The proof is deferred to the supplementary material. Theorem 6 provides a high-probability bound on the average squared norm of the gradient mapping generated by the differentially private algorithm. The bound is of order $1/(n\varepsilon)$ up to factors involving the model dimension, the smoothness constants, and the privacy parameter δ. It therefore decreases as either the local sample size n or the privacy budget ε increases, explicitly characterizing the trade-off between privacy and optimization accuracy. In particular, conditional on the stated high-probability event, an iterate selected uni-

formly from $\{0, \ldots, T-1\}$ satisfies the same bound in expectation with respect to the random choice of the output iterate.

4. Simulation Studies and Real Data Analysis

In this section, we investigate the finite-sample performance of the proposed methods through simulation. The data are generated from the following model:

$\mathcal{M}_1$: $Y = \exp(\boldsymbol{\alpha}^\top \boldsymbol{x}) + \varepsilon$, where $\boldsymbol{\alpha} = (1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5})^\top$, $\varepsilon \sim \mathcal{N}(0,1)$, and the components of $\boldsymbol{x}$ are i.i.d. from $U[-1,1]^5$.

4.1 Distributed estimation of SIM

We use cubic splines ($\rho = 3$) with equally spaced knots. The total sample size N is chosen from $\{2^{11}, 2^{12}, 2^{13}, 2^{14}, 2^{15}\}$, and the number of machines s is selected from $\{1, 4, 16, 64, 128\}$, such that each machine has $n = N/s$ observations. The spline basis dimension is set to $J_n \in \{n^{1/5}, n^{1/6}, n^{1/7}\}$. Each configuration is replicated 100 times.

We first assess the performance of the aggregated estimator $\bar{f}$ for f_0 using the mean squared error (MSE) under different choices of s and N. Figure 1 reports the average MSE of $\bar{f}$. The estimation accuracy deteriorates when the local sample size is too small. Once the local sample size is

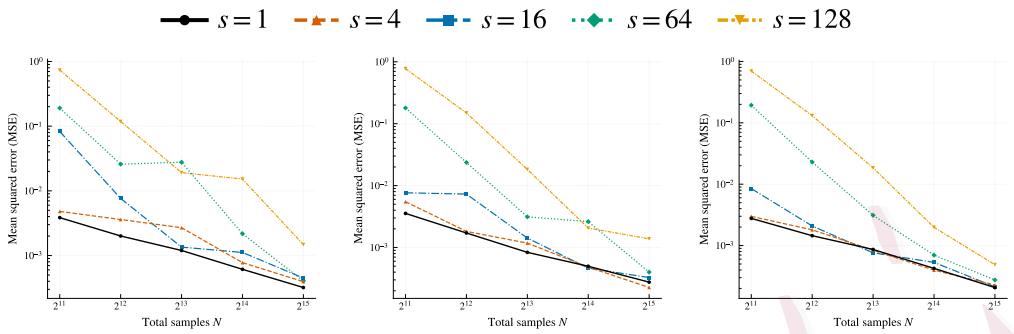
Statistica Sinica: Newly accepted Paper

sufficiently large, however, the performance of $\bar{f}$ remains relatively stable over a range of s. In addition, the MSE generally decreases as the number of knots approaches the theoretically motivated order.

Table 1 compares the proposed DC estimator with the two-stage refitting strategy and three benchmark methods for $N = 16384$ and $J_n = n^{1/7}$. For SIR-Spline and MAVE-Spline, the index direction is estimated locally using SIR and a MAVE-type procedure, respectively (Li, 1991; Xia and Härdle, 2006), and then aggregated using sample-size weights. IDC-Spline is adapted from the iterative divide-and-conquer method of Jiang and Peng (2023). For all benchmarks, the nonparametric component is estimated by the same B-spline procedure with the aggregated index direction held fixed. The proposed one-shot DC estimator remains competitive for small or moderate s, achieving comparable accuracy with only one round of aggregation. When s becomes large, the two-stage strategy and benchmark methods yield smaller nonparametric MSEs but require additional communication or a separate index-estimation step. Thus, the proposed estimator provides greater communication efficiency with competitive accuracy unless the data are split too aggressively. The corresponding index-parameter results are reported in the Appendix.

We next study finite-sample inference for the aggregated parametric

Statistica Sinica: Newly accepted Paper



(a) $J_n = n^{1/5}$ (b) $J_n = n^{1/6}$ (c) $J_n = n^{1/7}$

Figure 1: Average MSE of the aggregated estimator $\bar{f}$ under different choices of s and N, with curves to different numbers of machines s.

estimator using two types of confidence regions: marginal coordinatewise intervals and a chi-square calibrated joint region for the index parameter. Figure 2 reports empirical coverage under nominal 95% inference for different choices of s, N, and J_n. The top row shows marginal coverage averaged over coordinates, and the bottom row shows joint coverage. The results support the theoretical scaling requirement: when $n \gg N^{5/8}$, both marginal and joint coverages remain close to the nominal level 0.95 across the considered choices of J_n. When this condition is violated, coverage deteriorates substantially, with the joint region exhibiting more severe undercoverage. Overall, these results confirm that $n \gg N^{5/8}$ is essential for well-calibrated inference for the aggregated parametric estimator. Additional finite-sample

Statistica Sinica: Newly accepted Paper

Table 1: Nonparametric MSE comparison for N=16384.

s	Proposed DC	SIR-Spline	IDC-Spline	MAVE-Spline	Two-stage refit
1	0.000425	0.000425	0.000425	0.000425	0.000425
4	0.000397	0.000395	0.000395	0.000396	0.000395
16	0.000531	0.000495	0.000494	0.000495	0.000494
64	0.000698	0.000404	0.000403	0.000404	0.000403
128	0.001700	0.000573	0.000549	0.000557	0.000553

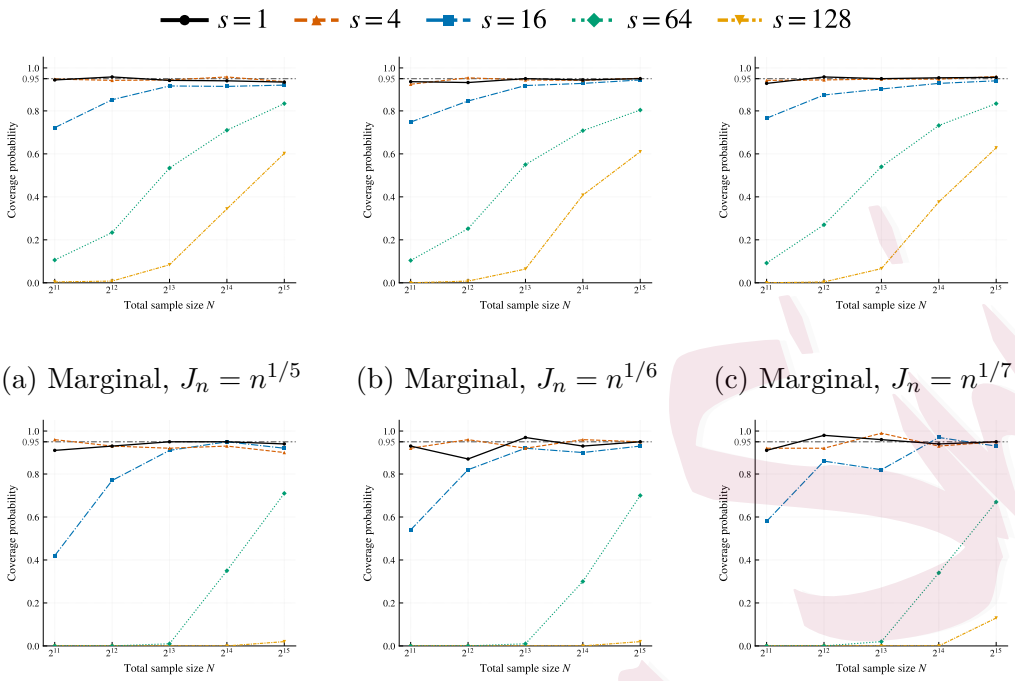
results are deferred to the Appendix.

4.2 Estimation with differential privacy

We first study the proposed differentially private algorithm in the single-machine setting. We consider $n \in [5000, 50000]$, use cubic splines with $\rho = 3$ and equally spaced knots, and set $J_n = n^{1/7}$. The privacy levels are $(\varepsilon, \delta) \in \{(0.2, 1/n), (0.5, 1/n), (0.8, 1/n)\}$, with 100 replications for each configuration.

Figure 3a and Table 2 summarize the finite-sample performance in the single-machine setting. Figure 3a shows that the MSE of the private nonparametric estimator decreases with n, while stronger privacy protection leads to larger estimation error. Table 2 reports the mean squared error (MSE) $\frac{1}{t}\sum_{j=1}^{m}\sum_{i=1}^{t}(\hat{\alpha}_{ji} - \alpha_{ji})^2$ and the absolute bias sum (ABS)

Statistica Sinica: Newly accepted Paper



(a) Marginal, $J_n = n^{1/5}$ (b) Marginal, $J_n = n^{1/6}$ (c) Marginal, $J_n = n^{1/7}$

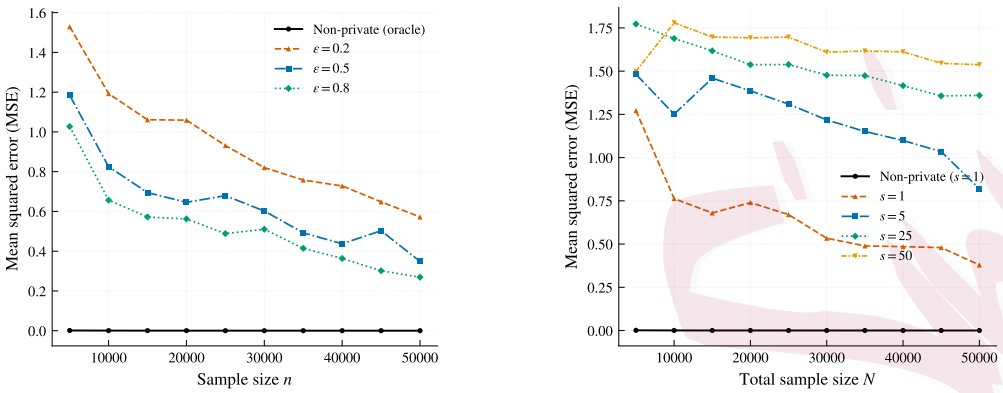
(d) Joint, $J_n = n^{1/5}$ (e) Joint, $J_n = n^{1/6}$ (f) Joint, $J_n = n^{1/7}$

Figure 2: Empirical coverage probabilities of nominal 95% confidence regions for the aggregated parametric estimator. The top and bottom rows show marginal and joint coverage, respectively; columns correspond to spline basis dimensions J_n, and curves to numbers of machines s.

$\frac{1}{t}\sum_{j=1}^{m}\sum_{i=1}^{t}|\hat{\alpha}_{ji} - \alpha_{ji}|$, where m is the dimension of $\boldsymbol{\alpha}$ and t is the number of repetitions. The table shows the same overall pattern for the parametric estimator $\boldsymbol{\alpha}$: both ABS and MSE decrease with n, while privacy constraints introduce additional error relative to the non-private case. Overall, the re-

Statistica Sinica: Newly accepted Paper

sults agree with the theory that larger sample sizes improve accuracy and reduce the relative impact of privacy noise.



(a) Single-machine DP setting.

(b) Distributed DP setting.

Figure 3: Average MSE of the differentially private estimator for f_0 under two simulation settings. Panel (a) considers the single-machine DP case with varying privacy levels, while Panel (b) considers the distributed DC case with a common privacy budget across machines.

Next, we study the private aggregated estimators $(\bar{f}_T, \bar{\boldsymbol{\alpha}}_T)$ in the DC setting. We consider total sample sizes $N \in [5000, 50000]$ and numbers of machines $s \in \{1, 5, 25, 50\}$, so that each local machine handles $n = N/s$ observations. Under parallel composition, each machine uses the same privacy budget $(\varepsilon, \delta) = (0.5, 1/N)$. All results are based on 100 replications.

Figure 3b and Table 3 summarize the performance of the distributed

Statistica Sinica: Newly accepted Paper

Table 2: The ABS and MSE of the estimator $\boldsymbol{\alpha}$ for the model $\mathcal{M}_1$ under privacy levels $(0.2, 1/n)$, $(0.5, 1/n)$, and $(0.8, 1/n)$

n		5000	10000	15000	20000	25000	30000	35000	40000	45000	50000
non	ABS	0.06460	0.04280	0.03560	0.03175	0.02835	0.02460	0.02170	0.02125	0.01955	0.01860
private	MSE	0.00145	0.00065	0.00045	0.00035	0.00025	0.00020	0.00015	0.00015	0.00015	0.00010
$\varepsilon = 0.2$	ABS	0.79525	0.89790	0.88680	0.86000	0.83405	0.87685	0.89200	0.92825	0.91705	0.86920
	MSE	0.17885	0.24025	0.23570	0.21185	0.19820	0.23765	0.23440	0.25940	0.25785	0.24210
$\varepsilon = 0.5$	ABS	0.83215	0.86525	0.81040	0.82250	0.75760	0.67520	0.61460	0.55680	0.53650	0.50420
	MSE	0.19960	0.21770	0.20040	0.19955	0.17345	0.14190	0.12365	0.09885	0.09480	0.09195
$\varepsilon = 0.8$	ABS	0.77595	0.80295	0.69510	0.65680	0.56885	0.48090	0.46985	0.42815	0.34140	0.36060
	MSE	0.17345	0.18660	0.14355	0.13520	0.10340	0.07855	0.07355	0.06630	0.04330	0.04980

private estimators. Figure 3b shows that, for fixed s, the MSE of the aggregated nonparametric estimator $\bar{f}_T$ decreases as N increases, while for fixed N, the estimation error becomes larger as the number of machines increases. Compared with the non-private benchmark, differential privacy introduces additional estimation error, and this degradation becomes more pronounced in highly distributed settings. Table 3 shows a similar pattern for the aggregated parametric estimator $\bar{\boldsymbol{\alpha}}_T$. In the non-private case, both ABS and MSE decrease with N for each fixed s, while larger s leads to lower estimation accuracy. Under privacy constraints, the accuracy loss becomes more evident, especially when the computation is distributed across more machines. Overall, these results demonstrate the impact of differential pri-

Statistica Sinica: Newly accepted Paper

vacy on the accuracy of distributed estimation.

Table 3: The ABS and MSE of the aggregated estimator $\bar{\boldsymbol{\alpha}}_T$ under privacy constraints

N			5000	10000	15000	20000	25000	30000	35000	40000	45000	50000
non private	$s=1$	ABS	0.06460	0.04281	0.03561	0.03173	0.02833	0.02460	0.02172	0.02125	0.01953	0.01860
		MSE	0.00146	0.00063	0.00045	0.00035	0.00027	0.00021	0.00016	0.00015	0.00013	0.00012
	$s=5$	ABS	0.08000	0.05326	0.04334	0.03788	0.03319	0.02972	0.02610	0.02499	0.02318	0.02229
		MSE	0.00195	0.00085	0.00058	0.00044	0.00033	0.00026	0.00021	0.00019	0.00017	0.00015
	$s=25$	ABS	0.15558	0.09265	0.07013	0.05883	0.05124	0.04545	0.04040	0.03775	0.03542	0.03381
		MSE	0.00767	0.00266	0.00154	0.00107	0.00082	0.00064	0.00052	0.00045	0.00040	0.00036
	$s=50$	ABS	0.25637	0.14664	0.10396	0.08424	0.07229	0.06290	0.05691	0.05224	0.04919	0.04581
		MSE	0.01998	0.00681	0.00346	0.00226	0.00168	0.00130	0.00106	0.00089	0.00079	0.00068
$\varepsilon=0.5$	$s=1$	ABS	0.79118	0.86972	0.79220	0.77450	0.77951	0.70778	0.71630	0.62826	0.61101	0.51713
		MSE	0.17306	0.21966	0.18680	0.18188	0.18241	0.15367	0.15943	0.12600	0.12186	0.08599
	$s=5$	ABS	0.81710	0.83758	0.75782	0.77671	0.76281	0.75525	0.75987	0.75997	0.76210	0.78399
		MSE	0.17969	0.18529	0.15053	0.15501	0.15035	0.14873	0.14965	0.14994	0.15059	0.16422
	$s=25$	ABS	0.85570	0.85606	0.86245	0.85553	0.81234	0.81857	0.80456	0.79887	0.80795	0.79209
		MSE	0.18261	0.18262	0.18498	0.18265	0.16512	0.16761	0.16145	0.15971	0.16349	0.15729
	$s=50$	ABS	1.77083	0.87293	0.98302	0.86957	0.81587	0.87163	0.83751	0.80576	0.83576	0.82333
		MSE	0.67127	0.18706	0.23268	0.18567	0.16492	0.18610	0.17296	0.16077	0.17232	0.16752

4.3 Real data analysis

We illustrate the proposed methods using the 2015 U.S. flight delay dataset (`https://www.transtats.bts.gov/OT_Delay/`). After preprocessing, the resulting sample contains 5,714,008 observations. Let ArrDelay be the ar-

Statistica Sinica: Newly accepted Paper

rival delay in minutes, and define $Y = \log(\text{ArrDelay} - \min(\text{ArrDelay}) + 1)$. The covariates are scheduled departure hour HD, flight distance DIS, a night-flight indicator NF, and a weekend-flight indicator WF. The continuous covariates HD and DIS are rescaled to $[-1, 1]$, and the binary indicators are coded as ± 1: NF $= 1$ for nighttime departures and -1 otherwise, and WF $= 1$ for weekend departures and -1 otherwise. We fit the single-index model (SIM):

$$Y = f(\alpha_1 \text{HD} + \alpha_2 \text{DIS} + \alpha_3 \text{NF} + \alpha_4 \text{WF}) + \varepsilon,$$

and compare it with the linear regression (LR):

$$Y = \alpha_1 \text{HD} + \alpha_2 \text{DIS} + \alpha_3 \text{NF} + \alpha_4 \text{WF} + \varepsilon.$$

For the DC analysis, we partition the data by airline and treat each airline as a local machine, yielding $s = 14$. Within each machine, we use an 80/20 training/validation split, and sample-size weights are used in aggregation. This random splitting procedure is repeated 100 times to assess the stability of the results. The nonparametric component is approximated by cubic splines ($\rho = 3$) with equally spaced knots and basis dimension $J_n = n_{\min}^{1/7}$, where $n_{\min} = 61{,}248$ is the smallest local training sample size.

Table 4 reports the estimated SIM index coefficients, normalized linear

Statistica Sinica: Newly accepted Paper

regression coefficients, and validation RMSEs for the DC and full-sample estimators. Figure 4 presents the fitted link functions, whereas Table 5 reports the corresponding index-group summaries.

Table 5 provides a more direct interpretation of the estimated index. Under both estimators, the median transformed response increases from the lower to the upper index group, indicating that the estimated index provides a meaningful ordering of flights with respect to transformed arrival delay. Larger index values are mainly associated with flights scheduled later in the day, particularly short-distance weekday flights. The estimated index may therefore be viewed as summarizing a late-day flight profile. Although the underlying mechanisms are not examined directly, this pattern is consistent with the propagation of earlier disruptions, accumulated airport congestion, and reduced opportunities for schedule recovery. Figure 4 shows that the fitted link is approximately linear over the central part of the index distribution but exhibits a mild nonlinear departure near the upper tail. This pattern is consistent with the similar predictive performance of the SIM and linear regression for most flights, while indicating that the SIM offers additional flexibility for flights with relatively large index values.

The DC and full-sample estimators yield similar response ordering, fitted link functions, and validation RMSEs, indicating that the main em-

Statistica Sinica: Newly accepted Paper

pirical patterns are broadly consistent across the two estimators. A more noticeable difference is observed in the proportion of night flights assigned to the upper index group. Since the night-flight indicator is defined from scheduled departure hour, these two covariates contain overlapping temporal information, which may partly account for the differences in their estimated coefficients. Despite these differences, the fitted responses and flight rankings remain similar. Overall, the empirical conclusions are relatively stable, while the decomposition of the estimated index across closely related temporal covariates is more sensitive.

Table 4: Estimated SIM index coefficients, normalized LR coefficients, and validation RMSEs for the DC and full-sample estimators

	SIM					Linear				
Method	α_1	α_2	α_3	α_4	RMSE	α_1	α_2	α_3	α_4	RMSE
DC estimator	0.92351	-0.32215	-0.19113	-0.08209	0.28287	0.89969	-0.39942	-0.15067	-0.09121	0.28257
Full-sample	0.84200	-0.52522	-0.09980	-0.07218	0.28237	0.39118	0.01143	0.85679	-0.33579	0.28472

For the differentially private analysis, both estimators are evaluated at $\varepsilon \in \{0.2, 0.5, 0.8\}$ and $\delta = 10^{-5}$, with all other settings unchanged. Each privacy setting is repeated 100 times, and the reported results are averaged over these repetitions. Table 6 presents the estimated index coefficients and validation RMSEs for the private and non-private SIM estimators.

Statistica Sinica: Newly accepted Paper

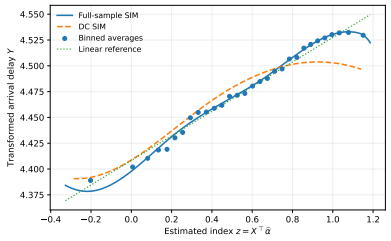


Figure 4: Fitted links for the DC and full-sample estimators.

Table 5: Index-group summaries for the DC and full-sample estimators.

Measure (DC/Full)	Low 10%	Mid. 80%	High 10%
Median Y	4.37/4.38	4.42/4.42	4.45/4.45
Dep. hour	6.68/7.14	13.62/13.46	19.03/19.89
Distance	1.27/1.50	0.81/0.79	0.53/0.40
Night (%)	3.90/3.90	12.20/8.50	6.40/36.00
Weekend (%)	51.60/47.60	25.70/26.20	3.80/4.00

Differential privacy increases the validation RMSE for both estimators, with the loss in predictive accuracy being more pronounced for the DC estimator, particularly at $\varepsilon = 0.2$. As ε increases, the validation RMSE decreases, exhibiting the expected privacy–utility trade-off. Across the privacy levels considered, scheduled departure time remains the dominant component of the estimated index, indicating that its association with transformed arrival delay is relatively robust to privacy perturbation. By contrast, the coefficients associated with flight distance and the night-flight and weekend-flight indicators change more substantially relative to their non-private estimates. The decomposition of the private index across these secondary covariates should therefore be interpreted with greater caution.

Table 6: Estimated index coefficients and validation RMSEs for the DC and full-sample SIM estimators, with and without differential privacy

DC						Full					
Setting	α_1	α_2	α_3	α_4	RMSE	Setting	α_1	α_2	α_3	α_4	RMSE
Non-private	0.9279	-0.3167	-0.1786	-0.0827	0.2950	Non-private	0.8442	-0.5216	-0.0999	-0.0728	0.2815
$\varepsilon = 0.2$	0.8886	0.0274	-0.4571	-0.0038	0.7492	$\varepsilon = 0.2$	0.8601	-0.1247	-0.4947	-0.0012	0.3997
$\varepsilon = 0.5$	0.8873	0.0446	-0.4584	-0.0030	0.5825	$\varepsilon = 0.5$	0.8613	-0.1189	-0.4940	-0.0018	0.3721
$\varepsilon = 0.8$	0.9007	0.0304	-0.4310	-0.0040	0.5230	$\varepsilon = 0.8$	0.8652	-0.1082	-0.4897	-0.0027	0.3364

5. Discussion

In this paper, we propose distributed estimators for the SIM for large-scale data and develop differentially private algorithms within the DC framework. We establish convergence rates and error bounds for the nonparametric component and study the asymptotic normality of the parametric component under DC. Under mild conditions, the aggregated estimator achieves the optimal nonparametric convergence rate. To address privacy risks in distributed computation, we further establish (ε, δ)-differential privacy guarantees and derive a high-probability bound on the average squared norm of the gradient mapping. Simulations and a real-data application show that the proposed methods achieve accurate estimation while providing effective privacy protection.

While our implementation focuses on one-shot aggregation, a natural

Statistica Sinica: Newly accepted Paper

direction for future work is to incorporate communication-efficient refinements, such as using surrogate objectives and one-step updates. Future studies could explore the trade-offs between statistical efficiency, privacy loss, and communication budget.

Supplementary Materials

The supplemental materials contain an appendix with proofs, technical lemmas, complete experimental results, and additional experiments.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Nos. 12071372, 12201395, 11901134), in part by the Sichuan Science and Technology Program (No. 2024ZYD0135).

References

Avella-Medina, M. (2021). Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association 116*(534), 969–983.

Awan, J. and S. Vadhan (2023). Canonical noise distributions and private hypothesis tests. *The Annals of Statistics 51*(2), 547–572.

Statistica Sinica: Newly accepted Paper

REFERENCES

Balabdaoui, F., C. Durot, and H. Jankowski (2019). Least squares estimation in the monotone single index model. *Bernoulli 25*(4B), 3276–3310.

Bao, Y. and W. Xiong (2021). One-round communication efficient distributed m-estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 46–54. PMLR.

Barrientos, A. F., A. R. Williams, J. Snoke, and C. M. Bowen (2024). A feasibility study of differentially private summary statistics and regression analyses with evaluations on administrative and survey data. *Journal of the American Statistical Association 119*(545), 52–65.

Cai, T. T., Y. Wang, and L. Zhang (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics 49*(5), 2825–2850.

Chen, X., W. Jing, W. Liu, and Y. Zhang (2024). Distributed estimation and inference for semiparametric binary response models. *The Annals of Statistics 52*(3), 922–947.

Chen, X. and M. Xie (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica 24*(4), 1655–1684.

Cui, X., W. K. Härdle, and L. Zhu (2011). The efm approach for single-index models. *The Annals of Statistics 39*(3), 1658–1688.

De Boor, C. (1978). *A practical guide to splines*, Volume 27. Springer New York.

Dong, C., J. Gao, and D. Tjøstheim (2016). Estimation for single-index and partially linear single-index integrated models. *The Annals of Statistics 44*(1), 425–453.

Statistica Sinica: Newly accepted Paper

REFERENCES

Dong, J., A. Roth, and W. J. Su (2022). Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(1), 3–37.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating noise to sensitivity in private data analysis. In *In TCC 2006*, pp. 265–284. Springer.

Fan, J., Y. Guo, and K. Wang (2023). Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association 118*(542), 1000–1010.

Gao, Y., W. Liu, H. Wang, X. Wang, Y. Yan, and R. Zhang (2022). A review of distributed statistical inference. *Statistical Theory and Related Fields 6*(2), 89–99.

Gu, J. and S. X. Chen (2023). Distributed statistical inference under heterogeneity. *Journal of Machine Learning Research 24*(387), 1–57.

Györfi, L. and M. Kroll (2025). On rate optimal private regression under local differential privacy. *Statistica Sinica 35*, 613–627.

Han, D., M. Han, J. Huang, and Y. Lin (2023). Robust inference for high-dimensional single index models. *Scandinavian Journal of Statistics 50*(4), 1590–1615.

Horowitz, J. L. (2009). Single index models. In *Semiparametric and Nonparametric Methods in Econometrics*, New York, NY, pp. 7–51. Springer Berlin Heidelberg.

Huang, H., J. Shangguan, X. Li, and H. Liang (2020). High-dimensional single-index models with censored responses. *Statistics in Medicine 39*(21), 2695–2854.

Jiang, R. and Y. Peng (2023). A short note on fitting a single-index model with massive data.

Statistica Sinica: Newly accepted Paper

Statistical Theory and Related Fields 7(1), 49–60.

Jordan, M. I., J. D. Lee, and Y. Yang (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association 114*(526), 668–681.

Kuchibhotla, A. K. and R. K. Patra (2020). Efficient estimation in single index models through smoothing splines. *Bernoulli 26*(2), 1587–1618.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*(414), 316–327.

Lian, H., H. Liang, and R. J. Carroll (2014). Variance function partially linear single-index models. *Journal of the Royal Statistical Society: Series B 77*(1), 171–194.

Tu, J., W. Liu, and X. Mao (2024). Distributed estimation on semi-supervised generalized linear model. *Journal of Machine Learning Research 25*, 1–41.

Wang, D., M. Ye, and J. Xu (2017). Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems 30*.

Wang, L. and L. Yang (2009). Spline estimation of single-index models. *Statistica Sinica 19*(2), 765–783.

Xia, Y. and W. Härdle (2006). Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis 97*(5), 1162–1184.

Yan, T. (2021). Directed networks with a differentially private bi-degree sequence. *Statistica Sinica 31*(4), 2031–2050.

Statistica Sinica: Newly accepted Paper

REFERENCES

Yu, Y. and D. Ruppert (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association 97*(460), 1042–1054.

Zhang, Y., J. Duchi, and M. Wainwright (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research 16*(1), 3299–3340.

Lingrui Wang, Center of Statistical Research, School of Statistics and Data Science, Southwestern University of Finance and Economics, Chengdu 611130, China.

E-mail: (wanglr9812@gmail.com)

Fode Zhang, Center of Statistical Research, School of Statistics and Data Science, Southwestern University of Finance and Economics, Chengdu 611130, China.

E-mail: (fredzh@swufe.edu.cn). ORCID ID: https://orcid.org/0000-0002-4733-9752

Hua Liang, Department of Statistics, The George Washington University, Washington, DC 20052, USA.

E-mail: (hliang@gwu.edu)