Statistica Sinica Preprint No: SS-2025-0189	
Title	Generalized Linear Models in Distributed Environments: a Framework for Optimal Subset Selection
Manuscript ID	SS-2025-0189
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0189
Complete List of Authors	Hongmei Lin, Xinxin Xiao, Chen Guo, Liu Yanlin, Tiejun Tong and Riquan Zhang
Corresponding Authors	Chen Guo
E-mails	gc_1998@163.com
Notice: Accepted author version.	

Generalized Linear Models in Distributed Environments: A Framework for Optimal Subset Selection

Hongmei Lin[1], Xinxin Xiao[1], Chen Guo[2*], Yanlin Liu[1], Tiejun Tong[3] and Riquan Zhang[1]

[1] *School of Statistics and Data Science, Shanghai University of International Business and Economics, Shanghai, China*

[2] *KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, China*

[3] *Department of Mathematics, Hong Kong Baptist University, Hong Kong, China*

Abstract: Modern data analysis often involves challenges from distributed and high-dimensional data, especially in fields such as finance and genomics. While distributed statistical methods address fragmented datasets and subset selection tackles high-dimensionality, existing approaches lack an effective combination of these two aspects. This paper proposes a communication-efficient distributed algorithm for best-subset selection in generalized linear models. The method extends a state-of-the-art centralized best-subset selection approach to a distributed framework, enabling scalable variable selection while optimizing computational efficiency and communication cost. Theoretical guarantees are established, including statistical consistency and convergence rates. Simulation studies and real-world data analyses validate the method's superior performance in terms of predictive accuracy and computational efficiency, demonstrating its practical relevance for modern data challenges.

Key words and phrases: Distributed statistical learning, Best-subset selection, Generalized linear models, Splicing technique.

*Corresponding author. E-mail: gc_1998@163.com

1. Introduction

1.1 Literature Review

In modern data analysis, particularly in fields such as finance, healthcare, and genomics, practitioners often face datasets that are both high-dimensional and fragmented across multiple sources. For example, in large-scale genomic studies, data collected from different laboratories may need to be integrated for a comprehensive analysis, while maintaining data privacy. Similarly, in financial markets, trading data from various institutions is distributed across geographical locations, making centralized analysis infeasible due to privacy concerns and communication costs. These real-world challenges necessitate the development of statistical methods that can handle distributed and high-dimensional data efficiently.

1.1.1 Distributed statistical methods: challenges and advances

Distributed statistical methods have emerged as a powerful framework to address above challenges, enabling scalable data analysis across multiple computing nodes. Existing distributed statistical methods can be broadly classified into two categories: one-shot and multi-round iterative methods. One-shot methods aggregate independently computed local estimators in a single communication round, offering simplicity and low communication cost (Li et al. 2013; Chen & Xie 2014; Zhang et al. 2015; Rosenblatt & Nadler 2016; Battey et al. 2018; Lv & Lian 2022;). Despite their efficiency, one-shot approaches require at least $\sqrt{N}$ samples (where N is the total sample size) per machine to match the global convergence rate, motivating multi-round iterative methods that refine local estimators through multiple

communication (Shamir et al. 2014; Chen et al. 2019; Jordan et al. 2019; Wang et al. 2019). Notably, Jordan et al. (2019) proposed multi-round distributed M-estimators achieving optimal convergence under mild machine-size constraints (Jiang & Yu 2021; Luo et al. 2022), with a comprehensive review available in Gao et al. (2022).

In addition to distributed methods, federated learning (FL) has emerged as a privacy-preserving paradigm for decentralized data analysis. FL retains raw data on local clients and communicates only model updates, thereby enhancing privacy protection while adhering to regulatory constraints. Recent advances have extended FL to high-dimensional and heterogeneous settings, further broadening its applicability (Li et al. 2020; Liu et al. 2022; Allouah et al. 2023; Li et al. 2024).

1.1.2 Generalized linear models and subset selection methods

Generalized linear models (GLMs) extend linear regression by allowing the response to follow an exponential family distribution, such as Gaussian, binomial, or Poisson distributions (McCullagh 2019). Their flexibility makes them widely applicable across fields including economics, biology, and geography. Formally, let $\{\mathbf{x}_i, y_i\}_{i=1}^n$ represent a dataset with n observations, where y_i is the response variable and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is a p-dimensional predictor vector. GLMs assume that the conditional density function of y_i given $\mathbf{x}_i$ takes the form $f(y_i; \mathbf{x}_i^\top \boldsymbol{\beta}, \phi) = \exp\{y_i \mathbf{x}_i^\top \boldsymbol{\beta} - b(\mathbf{x}_i^\top \boldsymbol{\beta}) + c(y_i, \phi)\}$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the regression coefficient vector, $b(\cdot)$ and $c(\cdot)$ are some suitably chosen known functions. The choice of $b(\cdot)$ determines the distribution of y_i with $b'(\mathbf{x}_i^\top \boldsymbol{\beta}) = \mathbb{E}(y_i|\mathbf{x}_i)$.

In high-dimensional settings where $p \gg O(n)$, directly minimizing the

negative log-likelihood function $l_n(\boldsymbol{\beta}) = -\sum_{i=1}^{n}\{y_i\mathbf{x}_i^\top\boldsymbol{\beta} - b(\mathbf{x}_i^\top\boldsymbol{\beta}) + c(y_i,\phi)\}$, often yields unstable or non-unique estimates of $\boldsymbol{\beta}$. Enforcing sparsity is therefore crucial for model interpretability and stability. To this end, we consider the following best subset selection problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} l_n(\boldsymbol{\beta}) \ \text{ subject to } \ \|\boldsymbol{\beta}\|_0 \leq s, \tag{1.1}$$

where $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} I(\beta_j \neq 0)$, $I(\cdot)$ is the indicator function, and s is a data-driven sparsity parameter. The cardinality constraint on $\|\boldsymbol{\beta}\|_0 \leq s$ ensures that at most s predictors have non-zero coefficients, yielding a sparse solution. However, the non-convex cardinality constraint poses significant computational challenges, as finding the global optimum requires searching over all possible predictor subsets.

Subset selection for (1.1) can be broadly classified into model selection-based, penalty-based, and optimization algorithm-based methods. Model selection-based approaches, such as AIC and BIC (Hocking & Leslie 1967, Schwarz 1978, Akaike 1998, Burnham & Anderson 2002), identify the optimal variable subset. Penalty-based methods promote sparsity through continuous penalties including LASSO(Tibshirani 1996), SCAD (Fan & Li 2001), elastic net (Zou & Hastie 2005), MCP (Zhang 2010), and truncated ℓ_1-penalty (Shen et al. 2012). Optimization algorithm-based methods approximate best-subset solutions using computational strategies, from classical forward stepwise selection (Mallat & Zhang 1993, Lozano et al. 2011) to advanced techniques such as the primal-dual active set methods, and mixed-integer optimization (Hintermüller et al. 2002; Blumensath & Davies 2009; Bertsimas et al. 2016). Although ℓ_0-penalized GLMs differ from (1.1), coordinate descent algorithms (Beck & Eldar (2013); Dedieu et al. (2021))

have been develop but often lack guarantees for both statistical recovery and computational efficiency. Recently, Zhu et al. (2020) proposed a splicing algorithm that selects the true sparse set via a few local swaps with polynomial computational complexity. Several recent studies have further extended Iterate Hard Thresholding (IHT) algorithm (Jain et al. (2014)) to more complex modeling frameworks. For example, Wang et al. (2023) developed an improved IHT algorithm for high-dimensional best subset selection under non-smooth loss functions in the centralized framework. Zhao & Lian (2025) further extended the IHT framework to distributed quantile regression models.

1.2 Our Proposal and Contributions

Despite advances in distributed statistical methods and subset selection, effective best-subset selection for high-dimensional GLMs in distributed settings remains challenging. To address this, we propose a communication-efficient distributed algorithm for high-dimensional best-subset selection (DBESS-GLM). Unlike ℓ_1-based approaches, our ℓ_0-constraint approach yields truly sparse solutions, enhancing interpretability while reducing communication cost. Secondly, a stitching technique with sacrificial variable ranking makes the NP-hard ℓ_0 problem computationally feasible. Thirdly, we utilize the generalized information criterion (GIC) to determine the subset size in a data-driven manner, ensuring true sparsity and robust performance. Finally, we establish minimax-optimal error bounds, providing tighter guarantees than Jordan et al. (2019). By addressing both distributed data and high-dimensional modeling, DBESS-GLM offers a scalable, interpretable, and statistically robust framework.

1.3 Organization and Notation

The remainder of this paper is organized as follows. Section 2 introduces the distributed setup, followed by the proposed distributed optimal subset selection algorithm and an adaptive version, with detailed implementation pathways. Section 3 presents the theoretical analysis of the algorithms, including error bounds, consistency properties, and asymptotic results. Section 4 evaluates the numerical performance of the proposed methods, while Section 5 highlights their practical advantages through real-world data analysis. Finally, Section 6 concludes with future research directions. Additional simulations and complete proofs are provided in the supplementary materials.

In this paper, we adopt the following notational conventions. For any vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$, the ℓ_0-norm is defined as $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^p I(\beta_j \neq 0)$, and the ℓ_q-norm for $q \in [1, \infty)$ is defined as $\|\boldsymbol{\beta}\|_q = (\sum_{j=1}^p |\beta_j|^q)^{1/q}$. Let $\mathcal{S} = \{1, \ldots, p\}$ and $\mathcal{A} \subseteq \mathcal{S}$, where $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$, and $\mathcal{A}^\complement = \mathcal{S} \setminus \mathcal{A}$ is its complement. The support set of $\boldsymbol{\beta}$ is defined as $\text{supp}(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\}$, and $\boldsymbol{\beta}_\mathcal{A} = (\beta_j : j \in \mathcal{A}) \in \mathbb{R}^{|\mathcal{A}|}$ represents the subvector of $\boldsymbol{\beta}$ corresponding to the indices in $\mathcal{A}$. For a matrix $\boldsymbol{M} \in \mathbb{R}^{n \times p}$, the submatrix indexed by $\mathcal{A}$ and columns by $\mathcal{B}$ is denoted as $\boldsymbol{M}_{\mathcal{A} \times \mathcal{B}} = (M_{ij} : i \in \mathcal{A}, j \in \mathcal{B}) \in \mathbb{R}^{|\mathcal{A} \times \mathcal{B}|}$. For a vector $\boldsymbol{t} \in \mathbb{R}^p$ and an index set $\mathcal{A}$, the modified vector $\boldsymbol{t}_\mathcal{A}$ is a p-dimensional vector where the j-th entry $(\boldsymbol{t}_\mathcal{A})_j$ equals t_j if $j \in \mathcal{A}$ and is zero otherwise. Furthermore, we use $a_n \lesssim b_n$ ($a_n \gtrsim b_n$) to indicate that a_n is less than (greater than) b_n up to a constant multiple. The notation $a_n \asymp b_n$ (or equivalently $a_n = O(b_n)$) signifies that a_n and b_n are of the same order of magnitude.

2. Methodology and Algorithm

This section presents our DBESS-GLM algorithm for large-scale datasets split across multiple machines. We develop its two-stage strategy, splicing technique, and GIC-based adaptive optimal subset-size selection, with detailed procedures presented in following subsections.

2.1 Problem Setup and Algorithm Framework

We consider the high-dimensional GLM problem (1.1) in a distributed computing framework. The full dataset

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^N, \boldsymbol{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_N^\top)^\top, \boldsymbol{Y} = (y_1, \ldots, y_N)^\top$$

is partitioned across m interconnected computing nodes-one central and $m-1$ local machines. Each machine k stores a disjoint subset of n_k observations, denoted by $\{(\mathbf{x}_i, y_i)\}_{i \in M_k}$, where M_k is the index set of the samples stored locally, $\cup_{k=1}^m M_k = \{1, \ldots, N\}$ and $N = \sum_{k=1}^m |M_k| = \sum_{k=1}^m n_k$. For simplicity, we assume an even split across machines, i.e., $n_1 = \cdots = n_m = n$, so that $N = nm$. This balanced data distribution simplifies computation and facilitates efficient algorithm implementation in distributed systems. Under the distributed framework, the global problem (1.1) can be reformulated as the following best-subset selection over the full dataset

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} -\frac{1}{N} \sum_{i=1}^N \left\{ y_i \mathbf{x}_i^\top \boldsymbol{\beta} - b(\mathbf{x}_i^\top \boldsymbol{\beta}) + c(y_i, \phi) \right\} \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_0 \leq s.$$

Let $\boldsymbol{\beta}^*$ be the true sparse coefficients vector, with support set $\mathcal{A}^*$ and cardinality $s^* = |\mathcal{A}^*|$. On the k-th machine, a local loss function is defined as

$\mathcal{L}_k(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i\in M_k}\left\{y_i\mathbf{x}_i^\top\boldsymbol{\beta} - b(\mathbf{x}_i^\top\boldsymbol{\beta}) + c(y_i,\phi)\right\}$. The overall loss function across all nodes is then given by

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{m}\sum_{k=1}^{m}\mathcal{L}_k(\boldsymbol{\beta}) = -\frac{1}{N}\sum_{k=1}^{m}\sum_{i\in M_k}\left\{y_i\mathbf{x}_i^\top\boldsymbol{\beta} - b(\mathbf{x}_i^\top\boldsymbol{\beta}) + c(y_i,\phi)\right\}.$$

Inspired by pioneering work(Jordan et al. 2019), we adopt an initial value $\bar{\boldsymbol{\beta}}$ to construct a gradient-enhanced surrogate loss that locally approximates the global loss: $\tilde{\mathcal{L}}(\boldsymbol{\beta}) := \mathcal{L}_1(\boldsymbol{\beta}) + \left\langle \nabla\mathcal{L}(\bar{\boldsymbol{\beta}}) - \nabla\mathcal{L}_1(\bar{\boldsymbol{\beta}}), \boldsymbol{\beta}\right\rangle$, where $\nabla\mathcal{L}(\bar{\boldsymbol{\beta}})$ is the global gradient evaluated at $\bar{\boldsymbol{\beta}}$, $\mathcal{L}_1(\boldsymbol{\beta})$ and $\nabla\mathcal{L}_1(\bar{\boldsymbol{\beta}})$. Omitting constant terms, the resulting estimator is obtained by solving

$$\min_{\boldsymbol{\beta}}\tilde{\mathcal{L}}(\boldsymbol{\beta}) = \mathcal{L}_1(\boldsymbol{\beta}) + (\nabla\mathcal{L}(\bar{\boldsymbol{\beta}}) - \nabla\mathcal{L}_1(\bar{\boldsymbol{\beta}}))^\top\boldsymbol{\beta} \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_0 \le s. \tag{2.2}$$

Our new DBESS-GLM algorithm is formulated based on Eq. (2.2), assuming $\|\boldsymbol{\beta}\|_0 = s$, with the optimal s adaptively selected via the GIC. Let $\mathcal{A} \subseteq \{1,\ldots,p\}$ denote the active set with cardinality $|\mathcal{A}| = s$, and $\mathcal{I} = \mathcal{A}^{\complement}$ be its complement, termed the inactive set. The algorithm proceeds in two main phases:

Stage 1: Active set recovery. Initially, Eq. (2.2) is iteratively solved within the generalized linear model framework. The iterations continue until the active set stabilizes, providing an estimate of the true parameter support.

Stage 2: Parameter estimation. Once the active set is identified, parameter estimation is restricted to this subset. A straightforward approach, such as the one-shot average method, can be applied. A formal description of the algorithm is outlined in Algorithm 1.

Algorithm 1 Distributed Best-Subset Selection for GLM with a given support size s (DBESS-GLM.Fix).

Input: Datasets $\{(\boldsymbol{X}_k, \boldsymbol{Y}_k)\}_{k=1}^m$, initial value $\boldsymbol{\beta}_0$, number of iterations T, support size s.

1: The initial active set $\mathcal{A}_0 = \{1 \le j < p : |\hat{\beta}_j^M| \ge \gamma_s\}$ is determined according to Eq. (2.4), where γ_s represents the s-th largest value in the set $\{|\hat{\beta}_1^M|, \ldots, |\hat{\beta}_p^M|\}$.
2: **for** $t = 0, \ldots, T-1$ **do**
3: **for** $k = 1, 2, \ldots, m$ **do**
4: Compute $\nabla \mathcal{L}_k(\boldsymbol{\beta}_t)$ on Machine k and broadcast it to Machine 1
5: On Machine 1, use Algorithm 2 to compute:

$$\boldsymbol{\beta}_{t+1} \leftarrow \arg\min_{\|\boldsymbol{\beta}\|_0 = s} \mathcal{L}_1(\boldsymbol{\beta}) + [\frac{1}{m}\sum_{k=1}^m \nabla \mathcal{L}_k(\boldsymbol{\beta}_t) - \nabla \mathcal{L}_1(\boldsymbol{\beta}_t)]^\top \boldsymbol{\beta} \tag{2.3}$$

6: Update the active set: $\mathcal{A}_{t+1} = \{j : (\boldsymbol{\beta}_{t+1})_j \neq 0\}$.
7: **if** $\mathcal{A}_{t+1} = \mathcal{A}_t$ **then**
8: break
9: $t = t + 1$.
10: Machine 1 broadcasts active set $\hat{\mathcal{A}} = \mathcal{A}_{t+1}$ to the all machines.
11: **for** $k = 1, 2, \ldots, m$ **do**
12: On Machine k, perform maximum likelihood estimation restricted to $\hat{\mathcal{A}}$

$$\hat{\boldsymbol{\beta}}^k = \arg\min_{\boldsymbol{\beta}_{(\hat{\mathcal{A}})^\complement}=0} -\frac{1}{n}\sum_{i \in M_k} \left\{ y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - b(\boldsymbol{x}_i^\top \boldsymbol{\beta}) + c(y_i, \phi) \right\},$$

13: and send $\hat{\boldsymbol{\beta}}^k$ to Machine 1.
14: On Machine 1, compute: $\hat{\boldsymbol{\beta}} = \frac{1}{m}\sum_{k=1}^m \hat{\boldsymbol{\beta}}^k$.

Output: $\hat{\boldsymbol{\beta}}, \hat{\mathcal{A}}$.

Remark 1. *If the initial active set $\mathcal{A}_0$ is sufficiently close to the true solution, it can substantially enhance the convergence rate of the algorithm. A natural choice for $\mathcal{A}_0$ is provided by sure independent screening for GLMs (Raskutti et al. 2011). For each predictor X_j, a univariate GLM estimates its coefficient $\hat{\beta}_j$. Subsequently, all variables are ranked according to the*

absolute values of their coefficients $|\hat{\beta}_j|$*:*

$$\hat{\beta}_j^M \leftarrow \arg\min_{\beta} -\sum_{i=1}^{N} \left\{ y_i x_{ij} \beta - b(x_{ij}\beta) + c(y_i, \phi) \right\}, \quad j = 1, \ldots, p. \tag{2.4}$$

The initial active set is then $\mathcal{A}_0 = \{1 \le j < p : |\hat{\beta}_j^M| \ge \gamma_s\}$*, where* γ_s *is the* s*-th largest value in the set* $\{|\hat{\beta}_1^M|, \ldots, |\hat{\beta}_p^M|\}$*.*

Remark 2. *In Algorithm 1, lines 1-9 focus on identifying the active set using a CSL function, and lines 10-14 perform local MLE within this set. The final estimator is obtained by averaging the local estimates on Machine 1 across all nodes.*

Remark 3. *With an initial value* $\boldsymbol{\beta}_0$ *close to the true parameter* $\boldsymbol{\beta}^*$*, Algorithm 1 can recover the true active set in essentially one iteration for GLMs. For theory, we assume such a good initialization to show the statistical efficiency of the one-step DBESS-GLM estimator based on MLE within the identified active set.*

2.2 Distributed Splicing Method for GLM

In the preceding subsection, we outlined the distributed GLM subset selection framework. Its main challenge lies in recovering the active set under the ℓ_0 constraint. Here, we propose a splicing approach that iteratively updates the active set while keeping computation and communication low.

Specifically, we employ the splicing technique to solve the best subset selection problem (2.3) under the sparsity constraint $\|\boldsymbol{\beta}\|_0 = s$, i.e., $\min_{\|\boldsymbol{\beta}\|_0 = s} \tilde{\mathcal{L}}(\boldsymbol{\beta})$. Given an active set $\mathcal{A}$, the restricted estimator is $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}_{\mathcal{I}} = \mathbf{0}} \tilde{\mathcal{L}}(\boldsymbol{\beta})$, with gradient $\widehat{\mathbf{d}} = \frac{\partial \tilde{\mathcal{L}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}}$. The optimality conditions are given in the following two lemmas.

Lemma 1. *Backward sacrifice: For any $j \in \mathcal{A}$, the magnitude of removing the j-th variable from $\mathcal{A}$ is $\xi_j^* = \tilde{\mathcal{L}}(\hat{\boldsymbol{\beta}}|_{\mathcal{A}\setminus\{j\}}) - \tilde{\mathcal{L}}(\hat{\boldsymbol{\beta}})$. Forward sacrifice: For any $j \in \mathcal{I}$, adding the j-th variable from $\mathcal{I}$ to $\mathcal{A}$ is $\zeta_j^* = \tilde{\mathcal{L}}(\hat{\boldsymbol{\beta}}) - \tilde{\mathcal{L}}(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{t}}|_j)$, where $\hat{\boldsymbol{t}}|_j = \arg\min_{\boldsymbol{t}} \tilde{\mathcal{L}}(\hat{\boldsymbol{\beta}} + \boldsymbol{t}|_j)$.*

Lemma 2. *Let C denote the size of the exchanged subset of elements. For any positive integer C such that $C < |\mathcal{A}^q|$, the associated range for ρ during the q-th iteration is*

$$\rho \in \begin{cases} \left(\dfrac{\min_{j\in\mathcal{S}_{C,2}^q} |d_j^q|}{\max_{i\in\mathcal{S}_{C,1}^q} |\beta_i^q|}, +\infty \right) & \text{for } C = 0, \\ \left(\dfrac{\min_{j\in\mathcal{S}_{C+1,2}} |d_j^q|}{\max_{i\in\mathcal{S}_{C+1,2}} |\beta_j^q|}, \dfrac{\min_{j\in\mathcal{S}_{C,2}} |d_j^q|}{\max_{i\in\mathcal{S}_{C+1,2}} |\beta_j^q|} \right] & \text{for } 1 \le C < s, \\ \left(0, \dfrac{\min_{j\in\mathcal{S}_{C,2}^q} |d_j^q|}{\max_{i\in\mathcal{S}_{C,1}^q} |\beta_i^q|} \right] & \text{for } C = s. \end{cases}$$

The magnitude of ξ_j^* quantifies the influence of the j-th active variable on the response variable, while ζ_j^* measures the relative importance of the j-th inactive variable. Lemma 2 clarifies the interaction between the penalty parameter ρ in the augmented Lagrange and the exchange counts C between the active and inactive sets. While C controls the swap rate, convergence is guaranteed for any C due to the finite combinations C_n^s. Identifying an ideal C value is simpler than tuning ρ; a practical strategy is to select C that yields a significant loss reduction after set updates. By integrating splicing into the iterative process, we obtain the splicing algorithm, as detailed in Algorithm 2, to efficiently solve problem (2.3).

For any stitching scale $C \le s$, we define sets $S_{C,1}$ and $S_{C,2}$ as follows:

$$S_{C,1} = \left\{ j \in \mathcal{A} : \sum_{i\in\mathcal{A}} I(\xi_j \ge \xi_i) \le C \right\}, S_{C,2} = \left\{ j \in \mathcal{I} : \sum_{i\in\mathcal{I}} I(\zeta_j \le \zeta_i) \le C \right\}.$$

We perform a swap operation between $\mathcal{S}_{C,1}$ and $\mathcal{S}_{C,2}$ to generate a candidate active set, which is denoted as $\tilde{\mathcal{A}} = (\mathcal{A} \setminus \mathcal{S}_{C,1}) \cup \mathcal{S}_{C,2}$. Correspondingly, the inactive set is represented as $\tilde{\mathcal{I}} = (\tilde{\mathcal{A}})^{\complement}$, the parameters are updated to $\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}_{\tilde{\mathcal{I}}}=0} \tilde{\mathcal{L}}(\boldsymbol{\beta})$, and the resulting loss function is given by $\tilde{\mathcal{L}}(\tilde{\boldsymbol{\beta}})$. If $\tilde{\mathcal{L}}(\tilde{\boldsymbol{\beta}})$ is significantly smaller than $\tilde{\mathcal{L}}(\hat{\boldsymbol{\beta}})$, it implies that $\tilde{\mathcal{A}}$ is more advantageous than $\mathcal{A}$, thereby triggering an update of the active set from $\mathcal{A}$ to $\tilde{\mathcal{A}}$. In essence, by exchanging elements between active and inactive sets, the active set can undergo iterative updates until no significant reduction in the loss function is observed. For a detailed calculation process, see Algorithm 2.

Remark 4. *A threshold $\tau_s > 0$ can enhance the efficiency of the algorithm by accelerating convergence and eliminating redundant iterations. The parameter $C_{\max}$, a positive integer no greater than s, defines the maximum number of exchanges allowed in Algorithm 1.*

2.3 Adaptive Distributed Best-Subset Selection Algorithm

Sparsity is a fundamental tool for analyzing high-dimensional data and improving predictive accuracy, typically assessed via cross-validation and information criteria such as Extended BIC, High-Dimensional BIC, and SIC (Chen & Chen 2008; Wang et al. 2013; Zhu et al. 2020). In distributed systems, where communication costs are crucial, we propose a data-driven approach GIC to determine the optimal support size s:

$$\mathrm{GIC}(\hat{\boldsymbol{\beta}}) = \tilde{\mathcal{L}}(\hat{\boldsymbol{\beta}}) + \left|\mathrm{supp}(\hat{\boldsymbol{\beta}})\right| \log(p) \log\log(N). \tag{2.5}$$

Intuitively, the penalty term $|\mathrm{supp}(\hat{\boldsymbol{\beta}})| \log(p) \log\log(N)$ prevents over-fitting, while the slowly diverging $\log\log(N)$ guards against under-fitting. The GIC

combines the loss function with this complexity penalty, and the optimal support size is obtained by minimizing GIC($\hat{\boldsymbol{\beta}}$). For each $s = 1, \ldots, s_{\max}$, Algorithm 1 is executed, and the sparsity level that yields the smallest GIC is selected. The complete procedure is summarized in Algorithm 3.

Algorithm 2 Distributed Splicing Algorithm for GLM.

Input: $\{(\boldsymbol{X}_k, \boldsymbol{Y}_k)\}_{k=1}^m$, initial active set $\mathcal{A}^0$ with s elements , maximum number of swaps $C_{\max} \leq s$, and τ_s.

Initialize : $q \leftarrow -1$, $\mathcal{I}^0 = (\mathcal{A}^0)^{\complement}$, $\boldsymbol{\beta}^0 \leftarrow \arg\min_{\boldsymbol{\beta}_{\mathcal{I}^0}=0} \tilde{\mathcal{L}}(\boldsymbol{\beta})$, $\mathbf{d}^0 \leftarrow \frac{\partial \tilde{\mathcal{L}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0}$.

repeat

$$q \leftarrow q+1, L \leftarrow \tilde{\mathcal{L}}(\boldsymbol{\beta}^q)$$

  for $C = 1, 2, \ldots, C_{\max}$ **do**

$$\xi_j \leftarrow \frac{\partial^2 \tilde{\mathcal{L}}(\boldsymbol{\beta})}{\partial \beta_j^2}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^q} (\beta_j^q)^2, \quad j \in \mathcal{A}^q,$$

$$\zeta_j \leftarrow \left(\frac{\partial^2 \tilde{\mathcal{L}}(\boldsymbol{\beta})}{(\partial \beta_j)^2}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^q}\right)^{-1} \left(d_j^q\right)^2, \quad j \in \mathcal{I}^q,$$

    Let

$$\mathcal{S}_{C,1} = \{j \in \mathcal{A}^q : \sum_{i \in \mathcal{A}^q} I(\xi_j \geq \xi_i) \leq C\},$$

$$\mathcal{S}_{C,2} = \{j \in \mathcal{I}^q : \sum_{i \in \mathcal{I}^q} I(\zeta_j \leq \zeta_i) \leq C\},$$

    Update $\tilde{\mathcal{A}} \leftarrow (\mathcal{A}^q \setminus \mathcal{S}_{C,1}) \cup \mathcal{S}_{C,2}$, $\tilde{\mathcal{I}} \leftarrow \tilde{\mathcal{A}}^{\complement}$,

    solve $\tilde{\boldsymbol{\beta}} \leftarrow \arg\min_{\boldsymbol{\beta}_{\tilde{\mathcal{I}}=0}} \tilde{\mathcal{L}}(\boldsymbol{\beta})$, $\tilde{\mathbf{d}} \leftarrow \frac{\partial \tilde{\mathcal{L}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}$.

    if $L - \tilde{\mathcal{L}}(\tilde{\boldsymbol{\beta}}) > \tau_s$ **then**

      $L \leftarrow \tilde{\mathcal{L}}(\tilde{\boldsymbol{\beta}})$, $(\mathcal{A}^{q+1}, \mathcal{I}^{q+1}, \boldsymbol{\beta}^{q+1}, \mathbf{d}^{q+1}) = (\tilde{\mathcal{A}}, \tilde{\mathcal{I}}, \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{d}})$.

until $\mathcal{A}^{q+1} = \mathcal{A}^q$

Output: $\boldsymbol{\beta}$, $\mathcal{A}$.

Algorithm 3 Adaptive Distributed Best-Subset Selection Algorithm for GLM (DBESS-GLM)

Input: Dataset $\{(\boldsymbol{X}_k, \boldsymbol{Y}_k)\}_{k=1}^m$, initial estimate $\boldsymbol{\beta}_0$, maximum sparsity $s_{\max}$.

1: **for** $s = 1, 2, \ldots, s_{\max}$ **do**
2: $(\hat{\boldsymbol{\beta}}^s, \hat{\mathcal{A}}^s) \leftarrow$ **DBESS-GLM.Fix**$(\{(\boldsymbol{X}_k, \boldsymbol{Y}_k)\}_{k=1}^m, \boldsymbol{\beta}_0, s)$
3: Compute $\mathrm{GIC}(s) = \mathrm{GIC}(\hat{\boldsymbol{\beta}}^s)$
4: $s_{\min} \leftarrow \arg\min_s \mathrm{GIC}(s)$

Output: $\hat{\boldsymbol{\beta}}^{s_{\min}}$, $\hat{\mathcal{A}}^{s_{\min}}$.

3. Theoretical Properties

This section presents the statistical guarantees and convergence analysis of our new algorithms, with technical conditions and complete proofs provided in the supplementary materials.

Theorem 1. *When $s^* \leq s$, let $(\widehat{\boldsymbol{\beta}}, \widehat{\mathcal{A}})$ be the solution of Algorithm 1. Under (A1)-(A6) in the supplementary materials, we have $\mathbb{P}\left(\mathcal{A}^* \subseteq \widehat{\mathcal{A}}\right) \geq 1 - \delta_1 - \delta_2 - \delta_3$, where $\delta_i = O(\exp\{\log p - K_i \frac{N}{s} \min_{j \in \mathcal{A}^*} |\beta_j^*|^2\})$. Asymptotically, $\lim_{N\to\infty} \mathbb{P}(\mathcal{A}^* \subseteq \widehat{\mathcal{A}}) = 1$. Especially, if $s = s^*$, then we have $\lim_{N\to\infty} \mathbb{P}(\mathcal{A}^* = \widehat{\mathcal{A}}) = 1$.*

Theorem 1 demonstrates that the DBESS-GLM algorithm can precisely identify the true active set. If initial set is incomplete, iterative exchanges between active and inactive sets reduce the loss until convergence, yielding the correct active set and the minimized loss.

Theorem 2. *Suppose that (A1)-(A7) in the supplementary materials are satisfied with respect to $s_{\max}$. Let $(\widehat{\boldsymbol{\beta}}^{s_{\min}}, \widehat{\mathcal{A}}^{s_{\min}})$ denote the solution obtained from Algorithm 3. Under the GIC, there exists a positive constant $\alpha > 0$ such that, with probability $1 - O(p^{-\alpha})$, for a sufficiently large sample size N, the algorithm correctly identifies the true active set, i.e., $\widehat{\mathcal{A}}^{s_{\min}} = \mathcal{A}^*$.*

Theorem 2 shows that, even in the absence of knowledge about the true sparsity of the model, the GIC in Eq. (2.5) allows DBESS-GLM to estimate it accurately with high probability. Combined with Theorem 1, the DBESS-GLM precisely identifies relevant variables while excluding irrelevant ones. The minimax risk, a standard metric for estimator performance, has been widely studied (Raskutti et al. 2011, Abramovich & Grinshtein 2016, Lee & Courtade 2020). Here we consider the ℓ_2 minimax risk over the sparse parameter space $\mathbb{B}_0(s) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 \leq s\}$. For the true parameter $\boldsymbol{\beta}^* \in \mathbb{B}_0(s^*)$ and its estimator $\widehat{\boldsymbol{\beta}}$, the risk is defined as follows:

$$\mathcal{R}_2(\mathbb{B}_0(s^*)) = \min_{\widehat{\boldsymbol{\beta}}} \max_{\boldsymbol{\beta}^* \in \mathbb{B}_0} \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\|_2^2.$$

Theorem 3. *(1) Under (A1)-(A3) in the supplementary materials, there exists a lower bound for the ℓ_2 minimax risk in the space $\mathbb{B}_0(s^*)$ as*

$$\mathcal{R}_2(\mathbb{B}_0(s^*)) \geq \frac{c_0}{32 c_2 M_{2s^*}} \frac{s^* \log(p/s^*)}{N}.$$

(2) Let $\widehat{\boldsymbol{\beta}}_{\widehat{s}}$ denote the output from Algorithm 3. Under the assumptions of Theorem 1, the following inequality holds:

$$\max_{\boldsymbol{\beta}^* \in \mathbb{B}_0(s^*)} \left\| \widehat{\boldsymbol{\beta}}_{\widehat{s}} - \boldsymbol{\beta}^* \right\|_2^2 \leq \frac{10 M_{s^*}}{c_2 c_0^2 m_{s^*}^2} \frac{s^* \log(p/s^*)}{N}.$$

Theorem 3 shows that for GLMs, the minimax ℓ_2 risk on $\mathbb{B}_0(s^*)$ is of order $O(\frac{s^* \log(p \backslash s^*)}{N})$, matching the known rate for linear models (Raskutti et al. 2011). Moreover, our algorithm attains this rate with a tight ℓ_2 upper bound, confirming its optimality.

The previous results primarily focus on the support recovery and min-

imax convergence rates. Once the optimal active set $\mathcal{A}^*$ is consistently recovered, it becomes important to conduct classical statistical inference on the nonzero coefficients. Accordingly, we study the asymptotic distribution of the final one-shot estimator obtained in Stage 2 of Algorithm 3.

Recall that $\boldsymbol{\beta}^*$ denotes the true sparse coefficient vector. For each observation, we write $X_{i,\mathcal{A}^*}$ for the sub-vector of covariates restricted to $\mathcal{A}^*$ and define the conditional mean $\mu_i^* = b'(X_{i,\mathcal{A}^*}^\top \beta_{\mathcal{A}^*}^*)$ and the score contribution $\psi_i = X_{i,\mathcal{A}^*}(Y_i - \mu_i^*)$. Let $I_* := \mathbb{E}\big[\, b''(X_{\mathcal{A}^*}^\top \boldsymbol{\beta}_{\mathcal{A}^*}^*)\, X_{\mathcal{A}^*} X_{\mathcal{A}^*}^\top \big]$ be the Fisher information matrix for $\boldsymbol{\beta}_{\mathcal{A}^*}^*$, where the expectation is taken with respect to the joint distribution of (X, Y); for fixed-design, I_* is the limit of the sample information matrices.

Theorem 4. *Suppose (A1)–(A10) in the supplementary materials hold, and* $(\widehat{\boldsymbol{\beta}}, \widehat{\mathcal{A}})$ *is the output of Algorithm 3. If* $\mathbb{P}(\widehat{\mathcal{A}} = \mathcal{A}^*) \to 1$, *then on the event* $\{\widehat{\mathcal{A}} = \mathcal{A}^*\}$, *the one-shot averaged estimator admits the linear representation* $\sqrt{N}\left(\widehat{\beta}_{\mathcal{A}^*} - \boldsymbol{\beta}_{\mathcal{A}^*}^*\right) = I_*^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi_i + o_p(1)$. *Consequently,*

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{\mathcal{A}^*} - \boldsymbol{\beta}_{\mathcal{A}^*}^*\right) \xrightarrow{d} \mathcal{N}\big(0,\, I_*^{-1}\big).$$

Moreover, $\widehat{\boldsymbol{\beta}}_{(\mathcal{A}^*)^c} = 0$ *with probability tending to one.*

Theorem 4 shows that, conditional on the successful recovery of the true active set $\mathcal{A}^*$, the asymptotic distribution of the one-shot averaged local restricted MLE yielded by DBESS-GLM is identical to that of the centralized oracle MLE obtained by fitting the sparse GLM directly on the full dataset of size N. In other words, under mild regularity conditions such as $m = o(\sqrt{N})$, the distributed framework incurs no loss of first-order statistical efficiency, retaining $\sqrt{N}$-consistency and the asymptotic variance

characterized by the Fisher information matrix I_*. It crucially relies on a second-order expansion of the local restricted MLE on each machine, with the remainder term of order $O_p(n^{-1})$. The growth condition in (A10) guarantees that the aggregated remainder is asymptotically negligible at the $\sqrt{N}$-scale.

Corollary 1. *Under the conditions of Theorem 4, for any fixed vector $a \in \mathbb{R}^{s^*}$, the linear contrast $\boldsymbol{a}^\top \boldsymbol{\beta}^*_{\mathcal{A}^*}$ satisfies*

$$\sqrt{N}\,\frac{\boldsymbol{a}^\top(\widehat{\boldsymbol{\beta}}_{\mathcal{A}^*} - \boldsymbol{\beta}^*_{\mathcal{A}^*})}{\sqrt{\boldsymbol{a}^\top I_*^{-1}\boldsymbol{a}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Moreover, if we estimate I_* by the plug-in estimator, in a distributed implementation, $\widehat{I} = \frac{1}{m}\sum_{k=1}^{m}\frac{1}{n}\sum_{i\in\mathcal{M}_k} b''\big(X_{i,\mathcal{A}^*}^\top \widehat{\boldsymbol{\beta}}_{\mathcal{A}^*}\big)\, X_{i,\mathcal{A}^*}X_{i,\mathcal{A}^*}^\top$, then $\widehat{I} \xrightarrow{P} I_*$ and the corresponding Wald confidence intervals for $\boldsymbol{a}^\top\boldsymbol{\beta}^*_{\mathcal{A}^*}$ are asymptotically valid.

Based on the preceding deduction, we can construct the Wald confidence intervals for any fixed linear combination $\boldsymbol{a}^\top\boldsymbol{\beta}_{\mathcal{A}^*}$, enabling standard parameter inference in distributed environments.

4. Numerical Experiments

We evaluate the proposed algorithms through numerical experiments in three parts. Section 4.1 details the simulation setup, including synthetic data, comparison methods, and evaluation metrics; Section 4.2 examines convergence via estimation errors; and Section 4.3 evaluates active set recovery accuracy. This systematic approach comprehensively assesses both optimization efficiency and statistical performance, demonstrating the algorithms' effectiveness for distributed GLMs.

4.1 Experimental Setup

The distributed dataset $D = \{\boldsymbol{X}_i, \boldsymbol{Y}_i\}_{i=1}^N$ is stored across m machines, each with n independent samples from $\mathcal{N}_p(\mathbf{0}, \Sigma)$, forming a local $n \times p$ design matrix and n-dimensional response vector. The global dataset combines into design matrix $\boldsymbol{X} \in \mathbb{R}^{N\times p}$ and $\boldsymbol{y} \in \mathbb{R}^N$, with $N = mn$. Responses are generated via a GLM using the true s-sparse coefficient vector $\boldsymbol{\beta}^*$. Simulations are conducted using both logistic and Poisson models; only the results for the logistic model are presented here, while the Poisson results are provided in the supplementary materials.

In this paper, we primarily compare the following four algorithms to evaluate the performance of different algorithms in a distributed environment: (1) DBESS-GLM, our proposed method with Algorithm 3, which uses the GIC for model selection; (2) CSL-GLM, a pseudo-likelihood based approach with ℓ_1 penalty, where λ is tuned via cross-validation; (3) CEASE-GLM, implementing the framework of Fan et al. (2023), which combines pseudo-likelihood function with Rockafellar (1976)'s PPA algorithm. (4) DIHT, implementing the distributed iterative hard thresholding algorithm for ℓ_0-constrained quantile regression proposed by Zhao & Lian (2025).

4.2 Analysis of Convergence Rates and Estimation Errors

We consider logistic regression with sparsity level $s^* = 10$ in $p = 100$ dimensions, where the non-zero coefficients $\beta^*_{\mathcal{A}^*} = (1, -1, -1, 1, 1, -1, 1, 1, 1, 1)^\top$. A total of $N = 50000$ samples are distributed across $m \in \{20, 40, 50\}$ machines and each run allows up to 10 iterations. Estimation error is quantified by $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2$ under two initializations: zero-vector initialization and one-shot averaging. All results are averaged over 100 trials.

As shown in Figures 1 and 2, the four methods exhibit different convergence patterns under distributed logistic regression settings. DBESS-GLM reduces the estimation error rapidly during the first few iterations and maintains relatively low estimation error across different values of m. Such behavior is consistent with the surrogate likelihood based optimization framework together with the splicing refinement strategy, which allows more effective use of the global curvature information in distributed GLMs.

In contrast, DIHT continues to reduce the estimation error over multiple iterations and requires substantially more iterations to attain a comparable level of accuracy. Moreover, its estimation error becomes increasingly sensitive to the number of machines, particularly under zero initialization. Although DIHT remains competitive in sparse support recovery (see Table 1), these results suggest that its estimation stability may deteriorate as the local sample size decreases. CSL-GLM and CEASE-GLM exhibit stable convergence behavior, but their estimation errors remain consistently higher than those of DBESS-GLM. While one-shot initialization generally improves the convergence behavior of all methods, DBESS-GLM remains highly stable under both initialization schemes.

4.3 Recovery of Sparse Subsets

Using the same logistic regression setup as in Section 4.2, we investigate the performance of different algorithms in recognizing the true active set in a distributed environment under one-shot initialization. Performance is measured by four metrics: True Positive Rate (TPR) for active feature detection accuracy, True Negative Rate (TNR) for inactive feature exclusion, Matthews Correlation Coefficient (MCC) for overall classification quality,

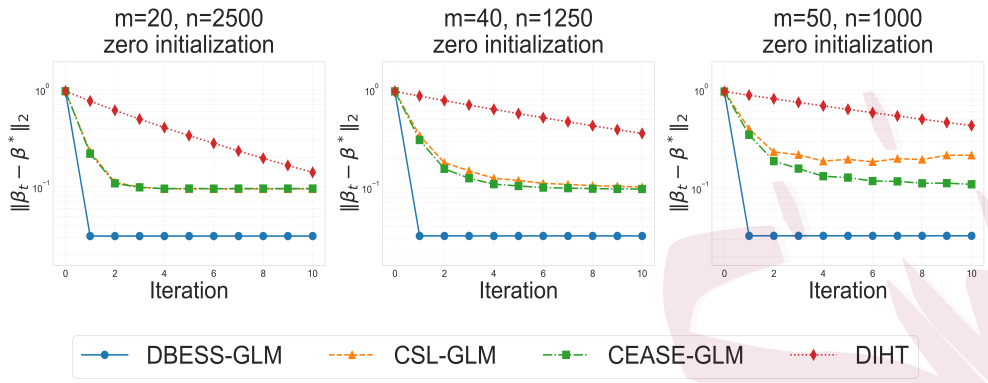

Figure 1: Comparison of convergence rates for one-shot initialization when m takes values of {20, 40,50}.

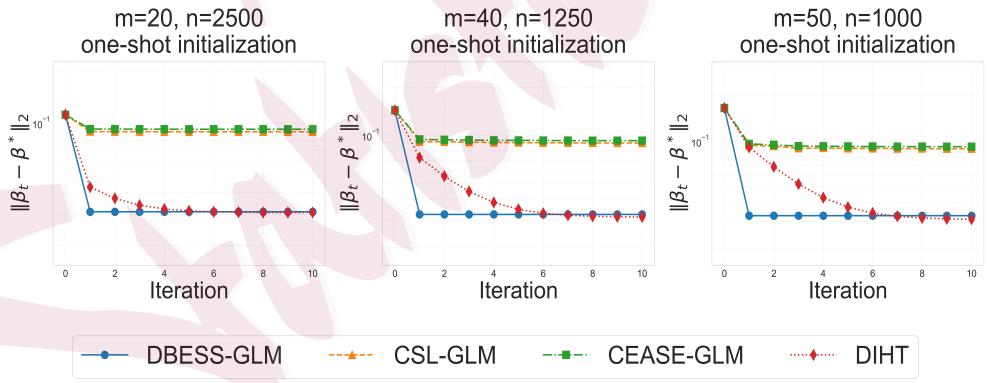

Figure 2: Comparison of convergence rates for zero initialization when m takes values of {20, 40,50}.

Table 1: Comparison of TPR, TNR, MCC and ReEE in Logistic Regression

Method	TPR	TNR	MCC	ReEE
		m=20		
CSL-GLM	1.0000(0.0000)	0.8776(0.0335)	0.6521(0.0611)	0.1118(0.0071)
CEASE-GLM	1.0000(0.0000)	0.7377(0.0440)	0.4715(0.0418)	0.0604(0.0075)
DBESS-GLM	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	0.0323(0.0077)
DIHT	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	0.0320(0.0075)
		m=40		
CSL-GLM	1.0000(0.0000)	0.9038(0.0384)	0.7059(0.0791)	0.1144(0.0068)
CEASE-GLM	1.0000(0.0000)	0.7736(0.0516)	0.5101(0.0567)	0.0492(0.0076)
DBESS-GLM	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	0.0332(0.0080)
DIHT	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	0.0320(0.0076)
		m=50		
CSL-GLM	1.0000(0.0000)	0.9082(0.0284)	0.7110(0.0603)	0.1160(0.0072)
CEASE-GLM	1.0000(0.0000)	0.7656(0.0499)	0.5010(0.0538)	0.0489(0.0085)
DBESS-GLM	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	0.0339(0.0082)
DIHT	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	0.0321(0.0076)

and Relative Error in Estimation (ReEE) for parameter recovery precision. These metrics are defined as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}, \quad \mathrm{ReEE} = \frac{\|\hat{\beta}-\beta^*\|_2}{\|\beta^*\|_2},$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}},$$

where $\mathrm{TP} = |\tilde{\mathcal{A}} \cap \mathcal{A}^*|$, $\mathrm{TN} = |\tilde{\mathcal{I}} \cap \mathcal{I}^*|$, $\mathrm{FN} = |\tilde{\mathcal{I}} \cap \mathcal{A}^*|$, and $\mathrm{FP} = |\tilde{\mathcal{A}} \cap \mathcal{I}^*|$. As CSL-GLM and CEASE-GLM algorithms may yield non-sparse estimates, we follow the truncation method by Battey et al. (2018) and apply a hard threshold ν, setting $|\hat{\beta}_j| < \nu$ to zero.

Table 1 shows the comparison results in terms of variable selection and parameter estimation. All four methods achieve nearly perfect TPR, indicating strong ability in identifying relevant variables. Among them, DBESS-GLM and DIHT achieve better performance in terms of TNR and

MCC, suggesting more accurate recovery of the underlying sparse structure. In contrast, CSL-GLM and CEASE-GLM tend to retain more irrelevant variables, leading to relatively lower TNR and MCC values.

Regarding parameter estimation accuracy, our new DBESS-GLM maintains relatively low ReEE across different values of m, while DIHT exhibits comparatively same estimation error. These results further demonstrate that DBESS-GLM achieves a favorable balance between variable selection accuracy and estimation stability in distributed GLM settings.

5. Real Data Analysis

Diabetes, a global chronic disease with no cure, poses a major public health challenge, highlighting the need for accurate prediction models to enable early intervention, improve diagnosis, and optimize healthcare resource allocation. For real data analysis, we consider the electronic medical data from the WiDS Datathon 2021 (`https://www.kaggle.com/competitions`), comprising 26,032 patient samples with 155 features each, where diabetes_mellitus serves as the binary response variable indicating diabetes status(0 or 1). The dataset was preprocessed to retain 103 relevant features after removing highly (more than 70%) missing values and imputing remaining gaps; descriptive statistical analysis is shown in the supplementary materials.

To investigate factors influencing diabetes prevalence, we employ logistic regression to the full dataset and compare DBESS-GLM, CSL-GLM and CEASE-GLM using the TPR, AUC (area under curve) and BS (Brier Score). The BS quantifies prediction accuracy through probability calibration, calculated as $\mathrm{BS} = \frac{1}{N}\sum_{k=1}^{m}\sum_{i\in M_k}(f_i - o_i)^2$, where f_i denotes the predicted probability and $o_i \in \{0, 1\}$ is the true label of the i-th sample. Lower

BS indicates better prediction accuracy. We use 10-fold cross-validation to select the regularization parameter with $m = 8$ machines, splitting the dataset 4:1 for training and testing, and repeating the experiment 100 times.

Table 2: Results of the experiment repeated 100 times.

Method	Model Size	TPR	AUC	Brier Score
CSL-GLM	17.128(1.813)	0.749(0.036)	0.808(0.027)	0.224(0.056)
CEASE-GLM	16.832(1.462)	0.772(0.018)	0.813(0.021)	0.216(0.043)
DBESS-GLM	14.852(1.384)	0.796(0.023)	0.824(0.017)	0.198(0.026)

As shown in Table 2, DBESS-GLM outperforms CSL-GLM and CEASE-GLM by selecting a sparser model while accurately identifying 15 clinically relevant variables, including Age, BMI index, Arf_apache (whether acute renal failure occurred), Glucose_apache (blood glucose concentration leading to the highest APACHE-III score), d1_glucose_max (maximum glucose concentration in serum or plasma within the first 24 hours) and so on. It reduces computational complexity, mitigates overfitting, and achieves higher AUC and lower BS, demonstrating superior predictive accuracy.

6. Discussion

Our work develops a communication-efficient distributed algorithm for best-subset selection in high-dimensional sparse generalized linear models, addressing challenges posed by large-scale data distributed across decentralized machines. By combining a theoretically justified subset screening strategy with a data-driven information criterion, the proposed method achieves consistent variable selection without requiring prior knowledge of sparsity levels. We show that the algorithm attains exact support recovery under mild conditions while maintaining polynomial computational complexity.

Empirical studies on logistic and Poisson regression models demonstrate improved convergence and estimation accuracy compared with existing approaches.

While this work provides important advances, several promising directions remain for future research. The framework can be extended to more challenging scenarios, including dependent or adversarial environments, privacy-constrained decentralized architectures, and models with structured sparsity. Further theoretical work may also relax distributional assumptions and clarify trade-offs between communication efficiency and statistical accuracy. These developments would strengthen both the theoretical foundations and the practical applicability of distributed sparse learning in complex data settings.

Supplementary Materials

The supplementary materials contain the technical conditions required for the theorems, the complete proofs of all lemmas and theorems, and the additional numerical results.

Acknowledgements

We sincerely thank the editor, the associate editor and two referees for their valuable comments and constructive suggestions. Hongmei Lin's research was supported in part by the Shanghai Oriental Talent Program (Youth Project) and the National Natural Science Foundation of China (12171310). Tiejun Tong's research was supported in part by the General Research Fund of Hong Kong (HKBU12300123) and the Initiation Grant for Faculty Niche Research Areas of Hong Kong Baptist University (RC-

FNRA-IG/23-24/SCI/03). Riquan Zhang's research was supported in part by the National Nature Science Foundation of China (12371272, 12531013).

References

Abramovich, F. & Grinshtein, V. (2016), 'Model selection and minimax estimation in generalized linear models', *IEEE Transactions on Information Theory* **62**(6), 3721–3730.

Akaike, H. (1998), Information theory and an extension of the maximum likelihood principle, *in* 'Selected papers of Hirotugu Akaike', Springer, pp. 199–213.

Allouah, Y., Guerraoui, R., Gupta, N., Pinot, R. & Stephan, J. (2023), On the privacy-robustness-utility trilemma in distributed learning, *in* 'International Conference on Machine Learning', PMLR, pp. 569–626.

Battey, H., Fan, J., Liu, H., Lu, J. & Zhu, Z. (2018), 'Distributed testing and estimation under sparse high dimensional models', *The Annals of Statistics* **46**(3), 1352–1382.

Beck, A. & Eldar, Y. C. (2013), 'Sparsity constrained nonlinear optimization: Optimality conditions and algorithms', *SIAM Journal on Optimization* **23**(3), 1480–1509.

Bertsimas, D., King, A. & Mazumder, R. (2016), 'Best subset selection via a modern optimization lens', *The Annals of Statistics* **44**(2), 813 – 852.

Blumensath, T. & Davies, M. E. (2009), 'Iterative hard thresholding for compressed sensing', *Applied and Computational Harmonic Analysis* **27**(3), 265–274.

Burnham, K. P. & Anderson, D. R. (2002), *Model selection and multimodel inference: a practical information-theoretic approach, 2nd Edition*, Springer.

Chen, J. & Chen, Z. (2008), 'Extended Bayesian information criteria for model selection with large model spaces', *Biometrika* **95**(3), 759–771.

Chen, X., Liu, W. & Zhang, Y. (2019), 'Quantile regression under memory constraint', *The Annals of Statistics* **47**(6), 3244–3273.

Chen, X. & Xie, M.-G. (2014), 'A split-and-conquer approach for analysis of extraordinarily

large data', *Statistica Sinica* **24**(4), 1655–1684.

Dedieu, A., Hazimeh, H. & Mazumder, R. (2021), 'Learning sparse classifiers: Continuous and mixed integer optimization perspectives', *Journal of Machine Learning Research* **22**(135), 1–47.

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

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.

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

Hintermüller, M., Ito, K. & Kunisch, K. (2002), 'The primal-dual active set strategy as a semismooth Newton method', *SIAM Journal on Optimization* **13**(3), 865–888.

Hocking, R. R. & Leslie, R. (1967), 'Selection of the best subset in regression analysis', *Technometrics* **9**(4), 531–540.

Jain, P., Tewari, A. & Kar, P. (2014), 'On iterative hard thresholding methods for high-dimensional m-estimation', *Advances in Neural Information Processing Systems* **1**, 685–693.

Jiang, R. & Yu, K. (2021), 'Smoothing quantile regression for a distributed system', *Neurocomputing* **466**, 311–326.

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

Lee, K.-Y. & Courtade, T. (2020), Minimax bounds for generalized linear models, *in* H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin, eds, 'Advances in Neural Information Processing Systems', Vol. 33, Curran Associates, Inc., pp. 9372–9382.

Li, M., Tian, Y., Feng, Y. & Yu, Y. (2024), 'Federated transfer learning with differential privacy',

arXiv preprint arXiv:2403.11343 .

Li, R., Lin, D. K. & Li, B. (2013), 'Statistical inference in massive data sets', *Applied Stochastic Models in Business and Industry* **29**(5), 399–409.

Li, T., Sahu, A. K., Talwalkar, A. & Smith, V. (2020), 'Federated learning: Challenges, methods, and future directions', *IEEE Signal Processing Magazine* **37**(3), 50–60.

Liu, K., Hu, S., Wu, S. Z. & Smith, V. (2022), 'On privacy and personalization in cross-silo federated learning', *Advances in Neural Information Processing Systems* **35**, 5925–5940.

Lozano, A., Swirszcz, G. & Abe, N. (2011), Group orthogonal matching pursuit for logistic regression, *in* G. Gordon, D. Dunson & M. Dudík, eds, 'Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics', Vol. 15 of *Proceedings of Machine Learning Research*, PMLR, Fort Lauderdale, FL, USA, pp. 452–460.

Luo, J., Sun, Q. & Zhou, W.-X. (2022), 'Distributed adaptive Huber regression', *Computational Statistics & Data Analysis* **169**, 107419.

Lv, S. & Lian, H. (2022), 'Debiased distributed learning for sparse partial linear models in high dimensions', *Journal of Machine Learning Research* **23**(2), 1–32.

Mallat, S. G. & Zhang, Z. (1993), 'Matching pursuits with time-frequency dictionaries', *IEEE Transactions on Signal Processing* **41**(12), 3397–3415.

McCullagh, P. (2019), *Generalized linear models*, Routledge.

Raskutti, G., Wainwright, M. J. & Yu, B. (2011), 'Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls', *IEEE Transactions on Information Theory* **57**(10), 6976–6994.

Rockafellar, R. T. (1976), 'Monotone operators and the proximal point algorithm', *SIAM Journal on Control and Optimization* **14**(5), 877–898.

Rosenblatt, J. D. & Nadler, B. (2016), 'On the optimality of averaging in distributed statistical learning', *Information and Inference: A Journal of the IMA* **5**(4), 379–404.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Shamir, O., Srebro, N. & Zhang, T. (2014), Communication-efficient distributed optimization using an approximate Newton-type method, *in* 'International Conference on Machine Learning', PMLR, pp. 1000–1008.

Shen, X., Pan, W. & Zhu, Y. (2012), 'Likelihood-based selection and sharp parameter estimation', *Journal of the American Statistical Association* **107**(497), 223–232.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society Series B: Statstical Methodology* **58**(1), 267–288.

Wang, L., Kim, Y. & Li, R. (2013), 'Calibrating non-convex penalized regression in ultra-high dimension', *The Annals of Statistics* **41**(5), 2505–2536.

Wang, X., Yang, Z., Chen, X. & Liu, W. (2019), 'Distributed inference for linear support vector machine', *Journal of Machine Learning Research* **20**(113), 1–41.

Wang, Y., Lu, W. & Lian, H. (2023), 'Best subset selection for high-dimensional non-smooth models using iterative hard thresholding', *Information Sciences* **625**, 36–48.

Zhang, C.-H. (2010), 'Nearly unbiased variable selection under minimax concave penalty', *The Annals of Statistics* **38**(2), 894–942.

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

Zhao, Z. & Lian, H. (2025), 'Distributed estimation for ℓ_0-constrained quantile regression using iterative hard thresholding', *Mathematics* **13**(4), 669.

Zhu, J., Wen, C., Zhu, J., Zhang, H. & Wang, X. (2020), 'A polynomial algorithm for best-subset selection problem', *Proceedings of the National Academy of Sciences* **117**(52), 33117–33123.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal*

of the Royal Statistical Society Series B: Statistical Methodology **67**(2), 301–320.

Hongmei Lin, Shanghai University of International Business and Economics, Shanghai, China

E-mail: hmlin@suibe.edu.cn

Xinxin Xiao, Shanghai University of International Business and Economics, Shanghai, China

E-mail: x12948500840022@163.com

Chen Guo, East China Normal University, Shanghai, China

E-mail: gc_1998@163.com

Yanlin Liu, Shanghai University of International Business and Economics, Shanghai, China

E-mail: ylliu_acad@163.com

Tiejun Tong, Hong Kong Baptist University, Hong Kong, China

E-mail: tongt@hkbu.edu.hk

Riquan Zhang, Shanghai University of International Business and Economics, Shanghai, China

E-mail: zhangriquan@163.com