Statistica Sinica Preprint No: SS-2024-0329	
Title	Robust and Efficient Decoding from Heterogeneous 1-bit Compressive Measurements Over Networks
Manuscript ID	SS-2024-0329
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0329
Complete List of Authors	Canyi Chen, Zhengtian Zhu and Liping Zhu
Corresponding Authors	Liping Zhu
E-mails	zhu.liping@ruc.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

Efficient Decoding from Heterogeneous 1-Bit Compressive Measurements over Networks

Canyi Chen, Liping Zhu, and Zhengtian Zhu

Southeast University, Renmin University of China and Tongji University

Abstract: The 1-bit compressive sensing (CS) framework is an effective method for approximating high-dimensional signals using binary measurements. However, deploying 1-bit CS in decentralized networks presents significant challenges due to the discontinuity of the sign function, susceptibility to sign flips, and the heterogeneity of network environments. Existing approaches often fail to adapt to decentralized settings, where communication with a central coordinator is prohibited, and only local, neighbor-to-neighbor interactions are feasible, driven by privacy and communication cost concerns. This paper introduces an efficient decentralized estimation approach for 1-bit CS. We reformulate the 1-bit CS problem as a penalized least squares task, which enables us to develop a generalized alternating direction method of multipliers algorithm with a simple implementation for optimization. We provide rigorous algorithmic and statistical guarantees. The proposed algorithm achieves *linear* convergence and, after a finite number of communications, attains a near-oracle statistical rate of convergence. Additionally, it reliably recovers the signal support under mild conditions. Extensive numerical studies demonstrate the efficacy of the proposed method.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

Key words and phrases: Decentralized estimation, sufficient dimension reduction, support recovery, oracle rate, linear convergence.

1. Introduction

Compressive Sensing (CS) is an effective method for approximating a high-dimensional signal from underdetermined measurements (Candes and Tao, 2005; Candes et al., 2006; Donoho, 2006). This approach maintains high accuracy even with relatively low sampling rates (Shannon, 1949). Typically, continuous infinite-precision measurements are quantized to reduce memory and transmission costs via analog-to-digital converters. An extreme form of quantization is 1-bit CS (Boufounos and Baraniuk, 2008), where each measurement is quantized to a single-bit sign. This low-cost quantization significantly simplifies hardware requirements, making 1-bit CS suitable for large-scale applications such as smart grids, multi-agent robotic systems, and wireless sensor networks.

Many practical applications, however, bring three new challenges.

1. *Decentralized Systems:* The cost efficiency of 1-bit CS enables large-scale deployment. For example, in wireless sensor networks, numerous inexpensive sensors improve signal measurement accuracy. These sensors are typically interconnected through a mesh network. Aggregating

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

all measurements at a single central node, as suggested by Maly and Palzer (2019) and Genzel and Jung (2019), may fail to fully exploit the sensor network's capabilities and incur high communication costs in decentralized networks with general topologies. The primary task in decentralized networks is to decode a global signal by distributing computing tasks across all nodes and exchanging information solely among neighbors.

2. *Data Heterogeneity:* Variations in sensor quality and potential corruptions result in different noise levels and sign-flip probabilities. Without global coordination by a central node, local nodes in decentralized systems rely heavily on local data and neighboring information to decode the signal. This heterogeneity can, hence, significantly impair estimation efficiency.

3. *Large Dimensional Sensing Vector:* Challenges arise when the signal dimension p is quite large. In CS, the signal is often sparse in Fourier or wavelet bases (Mallat, 1999; Ito and Jin, 2015). Leveraging sparsity is a natural approach to addressing the issue of underdetermined measurements. However, applying this principle in heterogeneous, decentralized systems remains an open question in the literature on

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

1-bit CS.

A motivating application of this work concerns a distributed network of satellites or drones engaged in remote imaging or acoustic sensing, such as sparse vegetation mapping, wildfire detection, or surveillance (Rani et al., 2018). In these settings, each node acquires compressive measurements of the scene while operating under stringent communication constraints and potentially unreliable channels, which may induce sign flips and heterogeneous noise variances. To reduce communication burden and hardware complexity, each node quantizes its measurements to a single bit, retaining only sign information. Furthermore, the system lacks a centralized controller; instead, nodes exchange limited information with neighboring units to collaboratively reconstruct the global sparse signal. This context motivates the development of decentralized 1-bit CS, where efficient communication and resilience to node failures are essential. Additional applications are provided in Appendix D of the Supplementary Material.

In 1-bit CS, measurements are noisy, nonlinearly dependent, and potentially sign-flipped. Let m denote the number of nodes in the network. The 1-bit CS framework assumes the following model at the jth node:

$$Y_j = \xi_j \text{sign}(\mathbf{x}_j^{\mathrm{T}} \boldsymbol{\beta}^\dagger + \varepsilon_j), \quad j = 1, \ldots, m, \tag{1.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

where Y_j is the 1-bit measurement, ξ_j is a random variable modeling the sign flip of Y_j with $\mathbb{P}(\xi_j = 1) = 1 - \mathbb{P}(\xi_j = -1) = q_j$, $\mathbf{x}_j$ is a p-dimensional sensing vector, $\boldsymbol{\beta}^\dagger$ is an unknown parameter of interest, and ε_j is an independent error with mean zero and variance σ_j^2. The sign function satisfies $\text{sign}(z) = 1$ if $z \geq 0$ and -1 otherwise. Both $\boldsymbol{\beta}^\dagger$ and σ_j are unknown and therefore not identifiable (Knudson et al., 2016). The jth node collects measurements of size n_j from model (1.1), denoted by $\{(\mathbf{x}_{i,j}, Y_{i,j})\}_{i=1}^{n_j}$. All measurements are assumed to be independent. However, they are not necessarily identically distributed. We merely require that the measurements within each local node follow an identical distribution. The objective of this paper is to efficiently estimate the direction of $\boldsymbol{\beta}^\dagger$ over a decentralized network.

For the specific case of $m = 1$, significant efforts have addressed both the theoretical and computational challenges inherent in the aforementioned problem. These efforts encompass first-order methods (Boufounos and Baraniuk, 2008; Laska et al., 2011; Yan et al., 2012; Dai et al., 2016) and greedy algorithms (Jacques et al., 2013; Liu et al., 2016). Despite their efficacy, these algorithms are not easily adaptable to decentralized 1-bit CS.

Most advances in the literature of distributed learning focus on CS without considering nonlinear sign quantization and potential sign flips. These methods can be roughly divided into two categories: 1) *Centralized*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

Distributed Learning: Here, a central node plays a key role (Hector and Song, 2020, 2021; Tang et al., 2020; Zhou et al., 2023; Chen and Zhu, 2023). Limited bandwidth of the central node could impair training efficiency, and its failure could disrupt the entire system. 2) *Decentralized Distributed Learning:* When a central node is not available or economical, decentralized distributed learning is proposed as an alternative. Chang et al. (2015) developed an inexact consensus alternating direction method of multipliers (ADMM) with a linear convergence rate. Additionally, Shi et al. (2015) and Li et al. (2019) proposed decentralized (proximal) (sub-)gradient descent algorithms to avoid the Lagrangian dual. Sub-gradient variants are also considered in Wang and Li (2018), Wang et al. (2019), and Zhang et al. (2019). Liu et al. (2022); Qiao and Chen (2024) considered the robust estimation under the linear model. However, most works focus on optimization convergence rates, and the statistical guarantees for parameter estimation and support recovery remain unclear.

In this work, we aim to accelerate 1-bit CS over a decentralized network while allowing heterogeneous measurements. Our contributions are summarized as follows:

- *A Least Squares Loss Function:* We demonstrate that estimating the direction of $\boldsymbol{\beta}^\dagger$ can be achieved by minimizing the least squares (LS)

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

loss function under a so-called linearity condition on the sensing vector, which is satisfied by a wide range of probability distributions, including the Gaussian distribution. Compared to classical results (Huang et al., 2018), we allow a more flexible choice of the distribution of the sensing vector. The LS loss is further armed with the ℓ_1 penalty to encourage sparsity of the estimate.

- *Decentralized Generalized ADMM:* This new formulation of the loss function enables us to design a generalized ADMM algorithm over networks to cooperatively solve the reformulated ℓ_1-regularized LS problem, converging linearly. Each local node needs to find a closed-form solution for each locally approximated subproblem and communicate local estimates with its neighbors. This implementation, hence, greatly reduces communication and computation costs at each local node without sacrificing estimation efficiency.

- *Statistical Guarantees:* We establish two key statistical guarantees: estimation accuracy and support recovery. After sufficient ADMM iterations, the estimate achieves a near-optimal statistical convergence rate $\{s\log(N)/N\}^{1/2}$, where s is the number of nonzero entries of $\boldsymbol{\beta}^\dagger$. This rate coincides with the optimal rate $(s/N)^{1/2}$, assuming the true

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

support is known up to a logarithmic factor. The estimated support matches the true support of $\boldsymbol{\beta}^\dagger$ under mild conditions.

Our LS approach for 1-bit CS bears a close connection to sufficient dimension reduction (Zhu et al., 2010). The LS formulation was initially applied to the context of 1-bit CS by Huang et al. (2018). More recently, Chen and Zhu (2023) proposed a centralized distributed learning framework. In the special case where all observations are aggregated, which corresponds to the scenario where $m = 1$, other methods have been developed to recover $\boldsymbol{\beta}^\dagger$. Notable techniques for estimating $\boldsymbol{\beta}^\dagger$ include inverse regression methods (Li and Duan, 1989; Li, 1991; Zhu et al., 2010), forward regression approaches (Xia, 2007), and semiparametric methods (Ma and Zhu, 2012). Our decentralized approach naturally extends these methods to the decentralized setting.

We organize the rest of the paper as follows. We develop a decentralized 1-bit CS in Section 2 and establish the accuracy of parameter estimation and support recovery in Section 3. In Section 4, we conduct extensive numerical studies to demonstrate the superior finite sample performance of our method. Section 5 concludes this article with brief discussions. All technical proofs, detailed derivations, additional simulation results, and an application to decoding EEG signals are relegated to the online Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

The codes used in our experiments are publicly available at `https://github.com/t9fhb47qyr-coder/decentralized1bit`.

The following notation will be used repetitively in subsequent exposition. We use $C, C_0, C_1, \ldots, c, c_0, c_1, \ldots$ to denote generic constants that may vary at each appearance. For a vector $\mathbf{v} = (v_1, \ldots, v_p)^{\mathrm{T}}$, we denote its support by $\mathcal{S}(\mathbf{v}) \stackrel{\text{def}}{=} \mathrm{supp}(\mathbf{v}) = \{j \in \mathbb{N}_+ : v_j \neq 0, 1 \leq j \leq p\}$. We further define vector norms

$$|\mathbf{v}|_0 \stackrel{\text{def}}{=} |\mathrm{supp}(\mathbf{v})|, \quad |\mathbf{v}|_1 \stackrel{\text{def}}{=} \sum_{i=1}^{p} |v_i|, \quad |\mathbf{v}|_2 \stackrel{\text{def}}{=} \left(\sum_{i=1}^{p} v_i^2 \right)^{1/2}, \text{ and } \mathbf{v}^{\min} \stackrel{\text{def}}{=} \min_{i \in \mathcal{S}(\mathbf{v})} |v_i|.$$

For $\mathcal{S} \subseteq \{1, \ldots, p\}$ with length $|\mathcal{S}|$, $\mathbf{v}_{\mathcal{S}} \stackrel{\text{def}}{=} (v_i, i \in \mathcal{S}) \in \mathbb{R}^{|\mathcal{S}|}$. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p \times q}$, we define matrix norms

$$|\mathbf{A}|_\infty \stackrel{\text{def}}{=} \max_{1 \leq i \leq p, 1 \leq j \leq q} |a_{ij}|, \|\mathbf{A}\|_\infty \stackrel{\text{def}}{=} \max_{1 \leq i \leq p} \sum_{1 \leq j \leq q} |a_{ij}|, \text{ and } \|\mathbf{A}\|_{\mathrm{op}} \stackrel{\text{def}}{=} \max_{|\mathbf{v}|_2 = 1} |\mathbf{A}\mathbf{v}|_2.$$

For two subsets $\mathcal{S}_1 \subseteq \{1, \ldots, p\}$, $\mathcal{S}_2 \subseteq \{1, \ldots, q\}$, we let $\mathbf{A}_{\mathcal{S}_1 \times \mathcal{S}_2} = (a_{ij}, i \in \mathcal{S}_1, j \in \mathcal{S}_2)$. Denote the largest and the smallest singular values of $\mathbf{A}$ by $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$, respectively, and the condition number by $\kappa(\mathbf{A})$, which, by definition, is $\lambda_{\max}(\mathbf{A})/\lambda_{\min}(\mathbf{A})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

2. Decentralized 1-Bit Compressive Sensing over Networks

2.1 A least squares approach

In this subsection, we will formulate 1-bit CS as an LS problem to bypass the issues of nonlinearity, non-differentiability, and sign flips in model (1.1). We establish such a formulation by showing that the direction of the minimizer of the LS loss at the population level aligns with that of the true parameter $\boldsymbol{\beta}^\dagger$.

We begin with necessary notations and assumptions. Since the sensing vectors $\mathbf{x}_j$'s are fully controlled by system designers, we assume these sensing vectors have zero means, and the same covariance matrices denoted as $\boldsymbol{\Sigma}$. To ensure identifiability of $\boldsymbol{\beta}^\dagger$, we assume $\boldsymbol{\beta}^{\dagger\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^\dagger = 1$, and the first component of $\boldsymbol{\beta}^\dagger$ is positive. This assumption can be guaranteed by the well-known dithered technique (Knudson et al., 2016). In model (1.1), if the first element of $\boldsymbol{\beta}^\dagger$ is negative, we can simply replace ξ_j with $-\xi_j$.

The formulation of 1-bit CS as an LS problem critically relies on the following linearity assumption.

Assumption 1. The p-dimensional mean-zero sensing vector $\mathbf{x}_j$ satisfies linearity of expectation in the direction of $\boldsymbol{\beta}^\dagger$: $E(\mathbf{x}_j \mid \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^\dagger) = \boldsymbol{\Sigma}\boldsymbol{\beta}^\dagger(\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^\dagger)$.

The mean-zero assumption is not essential for the validity of the method,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

which can be easily removed by incorporating an intercept term in the model. The linearity Assumption 1 is generally regarded as mild and is widely employed in the literature on sufficient dimension reduction (Cook, 1998). Li (1991) demonstrated that Assumption 1 holds when $\mathbf{x}_j$ follows an elliptical distribution, with the normal distribution as a specific case. This normality assumption is commonly utilized in 1-bit CS. Hall and Li (1993) further established that Assumption 1 approximately holds when the dimension of $\mathbf{x}_j$, p, is sufficiently large. For further discussion, see Diaconis and Freedman (1984) and Zhu and Zhu (2009). Moreover, in CS problems, the sensing vector $\mathbf{x}_j$ is often designed by the system (Arjoune et al., 2018), thereby ensuring the validity of Assumption 1.

Under Assumption 1, we have a simple but useful observation that lays the basis of our work. From Theorem 2.1 of Li and Duan (1989), we note under model (1.1) and the normality of $\mathbf{x}_j$,

$$\arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{m} E\left\{(Y_j - EY_j) - \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}\right\}^2 = \kappa\boldsymbol{\beta}^{\dagger},$$

for some $\kappa \in \mathbb{R}$. More generally, Huang et al. (2018); Chen and Zhu (2023) have shown that

Lemma 1. Suppose $(\mathbf{x}_j, Y_j)$ follows model (1.1) and $\mathbf{x}_j$ fulfills the linearity

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

Assumption 1 for $j = 1, \ldots, m$. The minimizer of

$$\sum_{j=1}^{m} E\left\{(Y_j - EY_j) - \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}\right\}^2, \tag{2.1}$$

denoted by $\boldsymbol{\beta}^*$, is proportional to $\boldsymbol{\beta}^\dagger$. The proportional constant is

$$\kappa = m^{-1}\sum_{j=1}^{m}(2 - 4q_j)\mathrm{cov}\left\{(\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^\dagger), F_j(-\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^\dagger)\right\},$$

where F_j is the cumulative distribution function of ε_j in model (1.1). In particular, if ε_j follows $\mathcal{N}(0, \sigma_j^2)$ and $\mathbf{x}_j$ follows $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, a simple application of Stein's lemma (Stein, 1981) helps quantify the constant κ precisely:

$$\kappa = m^{-1}\sum_{j=1}^{m}(2q_j - 1)\left\{\pi(1 + \sigma_j^2)/2\right\}^{-1/2}.$$

A key consequence of Lemma 1 is that $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^\dagger$ share the same direction unless $\kappa = 0$. Moreover, while $\boldsymbol{\beta}^\dagger$ is not identifiable, $\boldsymbol{\beta}^*$ is. Thus, estimating $\boldsymbol{\beta}^*$ is sufficient to determine the direction of $\boldsymbol{\beta}^\dagger$ in model (1.1). This result is significant as it enables the recovery of the direction of $\boldsymbol{\beta}^\dagger$ by directly minimizing the ordinary LS loss function. This approach circumvents the issue of non-differentiability associated with the sign function in model (1.1) under the presence of noise and sign flips. In addition, Lemma 1 does

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

not necessitate that measurements from different nodes follow the same distribution, thereby allowing for heterogeneity.

Typically, $\kappa \neq 0$, enabling us to identify the direction of $\boldsymbol{\beta}^\dagger$ from $\boldsymbol{\beta}^*$. If $\kappa > 0$, both the direction and sign of $\boldsymbol{\beta}^*$ align with those of $\boldsymbol{\beta}^\dagger$. Conversely, if $\kappa < 0$, their directions remain the same, but their signs are inverted. According to (2.1), $\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}^*$ is the optimal linear predictor of Y_j. In this scenario, the direction of $\boldsymbol{\beta}^\dagger$ can be recovered using $-\boldsymbol{\beta}^*$. Furthermore, the sign of $\boldsymbol{\beta}^\dagger$ can be determined using observations at the jth node as long as $q_j > 1/2$. In numerous applications, such as image compression and sound detection, determining the sign is often straightforward. In the worst-case scenario, where all q_j values are exactly $1/2$, recovering $\boldsymbol{\beta}^\dagger$ becomes impossible since all $\mathbf{x}_j$'s and Y_j's are independent.

At the sample level, with N measurements $\{\mathbf{x}_{i,j}, Y_{i,j}\}_{i=1,j=1}^{n,m}$ in places, in the principle of sparsity, to accommodate the sparse structure of the signal $\boldsymbol{\beta}^\dagger$, we advocate using the sparsity-encouraging ℓ_1-norm penalty and consider the following penalized solution,

$$\widehat{\boldsymbol{\beta}} \overset{\text{def}}{=} \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{2N}\sum_{j=1}^{m}\sum_{i=1}^{n}(Y_{i,j} - \mathbf{x}_{i,j}^{\mathrm{T}}\boldsymbol{\beta})^2 + \lambda_N|\boldsymbol{\beta}|_1. \tag{2.2}$$

where $\lambda_N > 0$ is the regularization parameter. Under the sparse structure,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

the ℓ_1-penalized least squares estimate could outperform the ordinary least squares estimate implied by Lemma 1. Moreover, adaptive ℓ_1 and non-convex penalties can be employed to mitigate the bias introduced by the ℓ_1 penalty, thereby enhancing the statistical convergence rate.

In order to optimize the minimization problem mentioned above, it is possible to consolidate all measurements into a single central node if allowed. However, this approach may require significant memory, storage, and communication resources when dealing with large-scale measurements and incur great privacy concerns. In the context of decentralized systems, there is a strong preference for efficient decentralized learning algorithms with communications confined to neighbors.

2.2 Decentralized 1-bit compressive sensing over networks

In this subsection, we introduce the consensus form of (2.2) over decentralized networks. We consider a connected network without self-loops. Denote the network where the data resides as an undirected, connected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ with $|\mathcal{N}| = m$, where $\mathcal{N}$ and $\mathcal{E}$ denote the sets of nodes and edges, respectively. The connectivity of $\mathcal{G}$ is modeled using an adjacency matrix $\mathbf{W} \in \{0,1\}^{m\times m}$. Let $\mathcal{N}_j$ denote the jth node. Nodes $\mathcal{N}_j$ and $\mathcal{N}_k$ are connected, and communication between them is possible if and only if

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

$\mathbf{W}_{jk} = 1$. By no self-loops, we mean the diagonal elements of $\mathbf{W}$ are all zeros.

We assume the nodes can perform local computation, yet communications are restricted to their neighbor through edges in $\mathcal{E}$ due to communications costs (Ling and Tian, 2010). Mathematically, the neighbors of node $\mathcal{N}_j$ is defined as $\mathcal{N}(j) \stackrel{\text{def}}{=} \{k\colon (j,k) \in \mathcal{E}\}$. For ease of illustration, we assume the local sample sizes are all the same, that is, $n_j = n$ for $j = 1, \ldots, n$, and use $\mathcal{D}_j \stackrel{\text{def}}{=} \{(\mathbf{x}_{i,j}, Y_{i,j})\}_{i=1}^{n}$ to denote the subsample at node $\mathcal{N}_j$. Let $\mathcal{D} \stackrel{\text{def}}{=} \cup_{j=1}^{m} \mathcal{D}_j$ be the entire sample. We can easily extend our method to the setting with varying local sample sizes.

In decentralized networks, a central coordinate node is no longer available to maintain global parameters $\boldsymbol{\beta}$. To optimize problem (2.2) over the network, we first cast (2.2) into the consensus form:

$$\begin{aligned} \arg\min_{\boldsymbol{\beta}^{(j)} \in \mathbb{R}^p} \quad & \frac{1}{m} \sum_{j=1}^{m} \Big\{ \frac{1}{2n} \sum_{i=1}^{n} (Y_{i,j} - \mathbf{x}_{i,j}^{\mathrm{T}} \boldsymbol{\beta}^{(j)})^2 + \lambda_N |\boldsymbol{\beta}^{(j)}|_1 \Big\} \\ \text{s.t.} \quad & \boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(k)}, \ \forall \ (j,k) \in \mathcal{E}, \end{aligned} \tag{2.3}$$

where $\boldsymbol{\beta}^{(j)}$ is an introduced local parameter copy at node $\mathcal{N}_j$. The penalty parameter λ_N is universal across the network and is utilized to regulate the sparsity of the estimated signal. In problem (2.3), the constraints ensure

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

that the local parameter copy at each node equals that of its neighbors. This gives rise to the term "consensus". Clearly, the solution of the problem (2.2) solves problem (2.3) and vice versa (Nedic and Ozdaglar, 2009; Shi et al., 2014; Yuan et al., 2016).

Addressing the sparse learning problem in (2.3) is challenging for two primary reasons. First, the nonsmooth nature of the objective function renders decentralized gradient or Newton-based methods impractical. Second, due to the incorporation of the ℓ_1 penalty, closed-form solutions for the consensus loss function are not readily obtainable.

2.3 Generalized ADMM

Next, we develop an efficient decentralized algorithm for solving (2.2), which guarantees *linear* convergence while limiting communications within neighbors.

By introducing pseudo-variables $\{\mathbf{t}^{(jk)}\colon k \in \mathcal{N}(j)\}_{j=1}^{m}$, problem (2.3) is equivalent to:

$$\begin{aligned} \arg\min_{\boldsymbol{\beta}^{(j)}\in\mathbb{R}^p} \quad & \frac{1}{2n}\sum_{j=1}^{m}\sum_{i=1}^{n}(Y_{i,j}-\mathbf{x}_{i,j}^{\mathrm{T}}\boldsymbol{\beta}^{(j)})^2+\lambda_N|\boldsymbol{\beta}^{(j)}|_1 \\ \text{s.t.} \quad & \boldsymbol{\beta}^{(j)}=\boldsymbol{\beta}^{(k)}=\mathbf{t}^{(jk)},\ \forall\ (j,k)\in\mathcal{E}. \end{aligned} \tag{2.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

To solve the above problem, the structure of the linear constraints naturally motivates us to appeal to the popular ADMM algorithm. Let $\mathbf{y}^{(j)} = (Y_{1,j}, \ldots, Y_{n,j})^{\mathrm{T}} \in \mathbb{R}^n$ and $\mathbf{X}^{(j)} = (\mathbf{x}_{1,j}, \ldots, \mathbf{x}_{n,j})^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ for $j = 1, \ldots, m$. By the classic ADMM theory (Boyd, 2010), we can construct augmented Lagrangians with a penalty parameter $\tau > 0$ as:

$$\mathcal{L}_\tau(\{\boldsymbol{\beta}^{(j)}\}, \{\mathbf{t}^{(jk)}\}, \{\mathbf{u}^{(jk)}\}, \{\mathbf{v}^{(jk)}\}) = (2n)^{-1}\sum_{j=1}^{m} |\mathbf{y}^{(j)} - \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}|_2^2 + \lambda_N|\boldsymbol{\beta}^{(j)}|_1$$
$$+ \sum_{j=1}^{m}\sum_{k\in\mathcal{N}(j)}\Big(\langle \mathbf{u}^{(jk)}, \boldsymbol{\beta}^{(j)} - \mathbf{t}^{(jk)}\rangle + \langle \mathbf{v}^{(jk)}, \boldsymbol{\beta}^{(k)} - \mathbf{t}^{(jk)}\rangle + \frac{\tau}{2}|\boldsymbol{\beta}^{(j)} - \mathbf{t}^{(jk)}|_2^2$$
$$+ \frac{\tau}{2}|\boldsymbol{\beta}^{(k)} - \mathbf{t}^{(jk)}|_2^2\Big),$$

where $\{\mathbf{u}^{(jk)}\colon k \in \mathcal{N}(j)\}_{j=1}^m$ and $\{\mathbf{v}^{(jk)}\colon k \in \mathcal{N}(j)\}_{j=1}^m$ are two sets of dual variables. Define an axillary variable $\mathbf{p}_t^{(j)} \overset{\text{def}}{=} \sum_{k\in\mathcal{N}(j)}(\mathbf{u}_t^{(jk)} + \mathbf{v}_t^{(jk)})$ with $\mathbf{p}_0^{(j)} = \mathbf{0}$, respectively. To find the solution for the above augmented Lagrangian, we can solve (2.4) by recursively performing (2.5) – (2.6):

$$\mathbf{p}_{t+1}^{(j)} = \mathbf{p}_t^{(j)} + \tau\sum_{k\in\mathcal{N}(j)}(\boldsymbol{\beta}_t^{(j)} - \boldsymbol{\beta}_t^{(k)}), \tag{2.5}$$

$$\boldsymbol{\beta}_{t+1}^{(j)} = \arg\min_{\boldsymbol{\beta}^{(j)}} \frac{1}{2n}\left|\mathbf{y}^{(j)} - \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}\right|_2^2 + \lambda_N|\boldsymbol{\beta}^{(j)}|_1 + \left\langle \mathbf{p}_{t+1}^{(j)}, \boldsymbol{\beta}^{(j)}\right\rangle$$
$$+ \tau\sum_{k\in\mathcal{N}(j)}\left|\boldsymbol{\beta}^{(j)} - (\boldsymbol{\beta}_t^{(j)} + \boldsymbol{\beta}_t^{(k)})/2\right|_2^2. \tag{2.6}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

Unfortunately, when the design matrix $\mathbf{X}^{(j)}$ is not column-orthogonal, problem (2.6) generally lacks a closed-form solution. As a result, multiple optimization iterations are required to obtain an approximate solution (Tao et al., 2016). This can impose significant computational demands on local nodes. To tackle this issue, we consider the generalized ADMM framework (Chang et al., 2015; Deng and Yin, 2016; Zhu, 2017; Gu et al., 2018) by upper bounding the Hessian matrix $\mathbf{X}^{(j)\mathrm{T}}\mathbf{X}^{(j)}/n$ in $(2n)^{-1}\left|\mathbf{y}^{(j)}-\mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}\right|_2^2$ with $\rho_j\mathbf{I}$ in (2.6). This approach can be interpreted as a variant of the Majorization-Minimization algorithm or as a proximal step (Sun et al., 2017). This replacement yields the closed-form approximation to $\boldsymbol{\beta}_{t+1}^{(j)}$,

$$\boldsymbol{\beta}_{t+1}^{(j)} = \mathcal{S}_{\lambda_N\omega_j}\Bigg[\omega_j\Big\{\rho_j\boldsymbol{\beta}_t^{(j)} - \frac{1}{n}\mathbf{X}^{(j)\mathrm{T}}(\mathbf{X}^{(j)}\boldsymbol{\beta}_t^{(j)}-\mathbf{y}^{(j)}) - \mathbf{p}_{t+1}^{(j)} + \tau\sum_{k\in\mathcal{N}(j)}(\boldsymbol{\beta}_t^{(j)}+\boldsymbol{\beta}_t^{(k)})\Big\}\Bigg], \tag{2.7}$$

where $\omega_j = 1/(2\tau|\mathcal{N}(j)|+\rho_j)$, $\mathcal{S}_t(\mathbf{v}) \overset{\text{def}}{=} (\mathbf{v}-t\mathbf{1}_p)_+ - (-\mathbf{v}-t\mathbf{1}_p)_+$ is the coordinate-wise soft-thresholding operator, and for any vector $\mathbf{v} = (V_1,\ldots,V_p)^{\mathrm{T}} \in \mathbb{R}^p$, $(\mathbf{v})_+ \overset{\text{def}}{=} \{\max(V_1,0),\ldots,\max(V_p,0)\}^{\mathrm{T}} \in \mathbb{R}^p$. In conclusion, we summarize our proposed decentralized algorithm for solving (2.3) in Algorithm 1. The Supplementary Material details deriving Algorithm 1.

The flexibility inherent in our LS transformation and ADMM updates

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

facilitates the extension of our algorithm to accommodate various sparsity-inducing penalties. These include the ℓ_0 norm penalty, the smoothly clipped absolute deviation (SCAD) penalty introduced by Fan and Li (2001), and the minimax concave penalty (MCP) developed by Zhang (2010). For a comprehensive overview of recent developments in this domain, readers are directed to Hastie et al. (2015) and the references contained therein.

Algorithm 1 Decentralized 1-bit Compressive Sensing.

Input: Data $\{(\mathbf{x}_{i,j}, Y_{i,j})\}_{i=1,j=1}^{n,m}$, the number of iterations T, the regularization parameter λ_N, the initial estimates $\widehat{\boldsymbol{\beta}}_{\text{init}}^{(j)}$ at each node $\mathcal{N}_j$.

1: Set $\boldsymbol{\beta}_0^{(j)} = \widehat{\boldsymbol{\beta}}_{\text{init}}^{(j)}$ and $\mathbf{p}_0^{(j)} = \mathbf{0}$.
2: **for** $t = 0, \ldots, T$ **do**
3: Communicate local parameter $\boldsymbol{\beta}_t^{(j)}$ with neighboring nodes;
4: Update $\mathbf{p}_{t+1}^{(j)}$ and $\boldsymbol{\beta}_{t+1}^{(j)}$ with (2.5) and (2.7), respectively.
5: **end for**

Output: The final estimate $\widehat{\boldsymbol{\beta}}^{(j)} = \boldsymbol{\beta}_{T+1}^{(j)}$.

3. Theoretical Results

This section provides theoretical guarantees for our estimate of the estimation bound and support recovery. Denote $\mathcal{S} \stackrel{\text{def}}{=} \text{supp}(\boldsymbol{\beta}^*)$ as the support of $\boldsymbol{\beta}^*$ and let $s \stackrel{\text{def}}{=} |\mathcal{S}|$. Lemma 1 ensures that $\text{supp}(\boldsymbol{\beta}^*) = \text{supp}(\boldsymbol{\beta}^\dagger)$. We impose the following assumptions to ensure our generalized ADMM algorithm holds a linear convergence rate and to derive the statistical properties.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

Assumption 2. The peer-to-peer network $\mathcal{G}$ is connected and has no self-loops.

Assumption 3. The sensing vector $\mathbf{x}_{i,j}$'s are sub-Gaussian, i.e., for some constants $t > 0$ and $C > 0$, $\sup_{|\boldsymbol{\theta}|_2=1} E \exp\{t(\mathbf{x}_{i,j}^{\mathrm{T}}\boldsymbol{\theta})^2\} \leq C$.

Assumption 4. The dimension p over the local sample size n is $p/n \leq \tau$ for some $\tau \in (0, 1)$. Assume that there exists a positive constant c_0, such that $c_0^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq c_0$.

Assumption 5. For each $j = 1, \ldots, m$, the initial estimate $\widehat{\boldsymbol{\beta}}_{\text{init}}^{(j)}$ satisfies $|\widehat{\boldsymbol{\beta}}_{\text{init}}^{(j)} - \boldsymbol{\beta}^*|_2 = O_p(1)$.

Assumption 2 is commonly employed in decentralized distributed learning frameworks, as it guarantees that no individual node or subset of nodes is isolated within the network. This facilitates the convergence of all nodes to a consensus state. Assumption 3 puts a typical sub-Gaussian condition on the distribution of $\mathbf{x}_{i,j}$ over the network $\mathcal{G}$. Assumption 4 is standard on the population covariances, which ensures the largest and smallest eigenvalues of the local sample covariances are bounded from infinity and zero with high probability, respectively (Yaskov, 2014). Accordingly, the optimization problem is strongly convex with high probability. In Algorithm 1, the initial estimate $\widehat{\boldsymbol{\beta}}_{\text{init}}^{(j)}$ is obtained using the measurements at each local node.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

Such an initial estimate satisfies Assumption 5 under Assumptions 3 and 4. Assumption 5 in fact allows an inaccurate initial estimate. In particular, we can simply take $\widehat{\boldsymbol{\beta}}_{\text{init}}^{(j)} = \mathbf{0}$. Under the identifiability assumption that ${\boldsymbol{\beta}^{\dagger}}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^{\dagger} = 1$, we have $|\boldsymbol{\beta}^{\dagger}|_2 \leq \lambda_{\min}^{-1}(\boldsymbol{\Sigma}) \leq c_0$.This together with Theorem 3.1 of Huang et al. (2018) yields $|\widehat{\boldsymbol{\beta}}_{\text{init}}^{(j)} - \boldsymbol{\beta}^*|_2 = O_p(1)$.

We first demonstrate the optimization convergence properties of the proposed generalized ADMM Algorithm 1. With certain assumptions and proper choices of the step lengths ρ_j, we show that the algorithm holds a linear convergence rate in the connected network.

Proposition 1 (Linear Convergence)**.** Under Assumptions 2–4, by setting the step lengths $\rho_j > \lambda_{\max}(n^{-1}{\mathbf{X}^{(j)}}^{\mathrm{T}}\mathbf{X}^{(j)})$ for each $j = 1, \ldots, m$, we have $(\sum_{j=1}^{m} |\boldsymbol{\beta}_{T+1}^{(j)} - \widehat{\boldsymbol{\beta}}|_2^2)^{1/2} = O_p(\gamma^T)$, where $\gamma \in (0, 1)$.

Proposition 1 demonstrates that the generalized ADMM Algorithm 1 solves the ℓ_1-penalized surrogate least squares problem with linear convergence. The convergence factor γ is determined solely by the network topology $\mathbf{W}$ and the singularity properties of the covariance matrices. The choice of the step length, ρ_j, is straightforward, as it depends only on the maximum eigenvalue of the local covariance matrix. See Appendix C of the Supplementary Material for a detailed discussion on Proposition 1 and practical guidance for setting up our generalized ADMM algorithm.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

Let $\theta_{T+1}^{(j)}$ be the angle between $\boldsymbol{\beta}_{T+1}^{(j)}$ and $\boldsymbol{\beta}^{\dagger}$. The following theorem presents the convergence rate of $\boldsymbol{\beta}_{T+1}^{(j)}$ and $\theta_{T+1}^{(j)}$.

Theorem 1. Set $\lambda_N = C_0(\log N/N)^{1/2}$, where $C_0 > 0$ is a sufficiently large constant. Under Assumptions 1–5, we have for $\gamma \in (0,1)$,

$$\begin{aligned} |\boldsymbol{\beta}_{T+1}^{(j)} - \boldsymbol{\beta}^*|_2 &= O_p\left\{\sqrt{\frac{s\log p}{N}} + \gamma^T\right\}, \text{ and} \\ |\cos\left(\theta_{T+1}^{(j)}\right)| &= |\boldsymbol{\beta}^{\dagger\mathrm{T}}\boldsymbol{\beta}_{T+1}^{(j)}|/\left(|\boldsymbol{\beta}^{\dagger}|_2|\boldsymbol{\beta}_{T+1}^{(j)}|_2\right) \geq 1 - O_p(|\boldsymbol{\beta}_{T+1}^{(j)} - \boldsymbol{\beta}^*|_2^2). \end{aligned}$$

By Theorem 1, the accuracy of angle estimation is governed by the ℓ_2 convergence rate of the estimate. Compared to the initial estimate $\widehat{\boldsymbol{\beta}}_{\text{init}}^{(j)}$, the convergence rate of our estimate $\boldsymbol{\beta}_{T+1}^{(j)}$ improves to $\max\{(s\log p/N)^{1/2}, \gamma^T\}$. The first term is from the convergence rate of $|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_2$, where $\widehat{\boldsymbol{\beta}}$ is defined in (2.2). The second term is the optimization error resulting from the generalized ADMM Algorithm 1. By letting the number of loops sufficiently large that $T \geq \log(s\log p/N)/\log\gamma$, we have the convergence rate of the our decentralized estimate $\boldsymbol{\beta}_{T+1}^{(j)}$ attains the minimax optimal rate $O\{(s\log p/N)^{1/2}\}$. This is also the optimal rate even when all the data are pooled at a single central node (Wainwright, 2009). Furthermore, the following corollary implies that T only needs to be of constant order.

Corollary 1. Under the assumptions of Theorem 1, when $p = O\{\exp(N/s)\}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

we further have the number of loops $T \geq \log(s \log p/N)/\log \gamma$ only needs to be constant order.

We provide guarantees on the support recovery of the limiting point of our decentralized estimate $\widehat{\boldsymbol{\beta}}$.

Theorem 2. Assume that $\|\boldsymbol{\Sigma}_{\mathcal{S}^c \times \mathcal{S}} \boldsymbol{\Sigma}_{\mathcal{S} \times \mathcal{S}}^{-1}\|_\infty \leq 1 - \alpha$ for some constant $0 < \alpha < 1$, and $s^2 \log p/N = o(1)$. Under assumptions of Theorem 1, we have $\text{supp}(\widehat{\boldsymbol{\beta}}) \subseteq \mathcal{S}$ with probability approaching one. Suppose, in addition, for a sufficiently large positive constant C, $\boldsymbol{\beta}^{*\min} \geq C\|\boldsymbol{\Sigma}_{\mathcal{S} \times \mathcal{S}}^{-1}\|_\infty (\log p/N)^{1/2}$. Then we have $\text{supp}(\widehat{\boldsymbol{\beta}}) = \mathcal{S}$ with probability approaching one.

The "beta-min" condition is widely used in the high-dimensional statistical literature and matches the oracle rate for the "beta-min" condition under the linear model (Wainwright, 2009).

4. Numerical Studies

4.1 Experimental setup

To evaluate the finite-sample performance of our proposed method, we conduct a series of experiments. We generate a connected decentralized network comprising m nodes using an Erdős-Rényi random graph with a connection probability p_c. At each node $\mathcal{N}_j$, data is simulated following

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

model (1.1). Specifically, the sensing vectors $\mathbf{x}_{i,j}$ at node j are sampled from $\mathcal{N}(0, \mathbf{\Sigma}_j)$ where $\mathbf{\Sigma}_{j,\ell k} = \sigma^2 \rho_j^{|\ell - k|}$, and the noise terms $\varepsilon_{i,j}$ are drawn from $\mathcal{N}(0, \sigma_j^2)$ with $\sigma_j = 1$. The probability of sign-flips is characterized by $\mathbb{P}(\xi_{i,j} = 1) = 1 - \mathbb{P}(\xi_{i,j} = -1) = q_j$. The sparsity of the true coefficients $\boldsymbol{\beta}^\dagger$ is set to s, where the indices of the nonzero entries are randomly selected, and these values are assigned to either -1 or 1 with equal probability. Unless otherwise specified, we use $m = 20$, $n = 100$, $p_c = 0.3$, $\sigma_j^2 = 1$, $p = 100$, $s = 10$, $\rho_j = 0.1$ and $q_j = 0.9$. Using the standard definition of the signal-to-noise ratio (SNR), i.e., $\text{SNR} = \text{Var}\{E(Y \mid \mathbf{x})\}/\text{Var}\{Y - E(Y \mid \mathbf{x})\}$, this corresponds to a $\text{SNR} = 0.869$. Except for the experiments in Section 4.4, all data are generated independently and identically distributed.

At each node, the initial estimates are obtained using the ℓ_1-penalized LS method on the local data. The tuning parameter λ_N is selected based on the Bayesian information criterion (Wang et al., 2009, BIC). Practically, broadcasting a single tuning parameter and aggregating the resultant local BIC values across the network incurs minimal communication overhead via the gossip protocol. Following the convention in Ma and Huang (2017), we set the Lagrangian parameter $\tau = 1$.

We compare our proposed method with the following four methods: 1) Pooled ℓ_1-penalized LS (Pooled), which performs the ℓ_1-penalized LS on

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

the pooled data in a single node and serves as the benchmark; 2) Local ℓ_1-penalized LS (Local), which performs the ℓ_1-penalized LS on local data at each node, and serves as the worse performance benchmark; 3) Average ℓ_1-penalized LS (Avg), which obtains the estimate by averaging the local ℓ_1-penalized LS, and such an averaging operation is usually completed via gossip algorithms over decentralized networks; 4) Decentralized subgradient descent (subGD) (Wang and Li, 2018), where the nodes collaboratively solve (2.3) with local subgradient descent and network communication. For all the decentralized estimates, we fixed the iteration budget to $T = 500$.

We employ two metrics to assess the performance of the model. The first metric is the estimation error, defined as $(\sum_{j=1}^{m} |\boldsymbol{\beta}_{T+1}^{(j)} - \boldsymbol{\beta}^*|_2^2/m)^{1/2}$. The second metric is the F_1-score, calculated as the harmonic mean of recall and precision at each node $\mathcal{N}_j$. Here, precision is the proportion of true nonzero components among the estimated nonzero components, and recall is the proportion of correctly identified nonzero components. The F_1-score ranges from 0 to 1, with higher values indicating superior performance. This score is widely utilized in the literature to evaluate support recovery accuracy. We report all results based on 100 independent replications to ensure stable evaluations.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

4.2 Effect of sample size

We examine how the sample size N impacts the performance of different methods. We vary the sample size $N \in \{1000, \ldots, 6000\}$ by fixing the number of nodes m fixed at 20 and varying the local sample size $n \in \{50, 100, \ldots, 300\}$. Results are summarized in Figure 1. The ℓ_2-error of our proposal decreases linearly in sample size N on the log-log scale, which is in line with the theoretical results in Theorem 1. Among all the decentralized estimates, our proposal performs the best and matches the pooled estimate in terms of both metrics. Of note, our estimate yields the desirable sparse solutions and depicts superior support recovery performance. A similar plateauing behavior in support recovery—as the F_1-score approaches one with increasing sample size—has also been observed in prior work (Wainwright, 2009). The subGD method does not typically produce sparse estimates, as updates are made in proportion to the gradient of the objective function, which is often non-sparse. This accounts for its low F_1-score even when the local sample size is large.

4.3 Effect of network topology

We investigate the effect of network topology by examining the following two factors: the number of nodes m and the network sparsity p_c.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

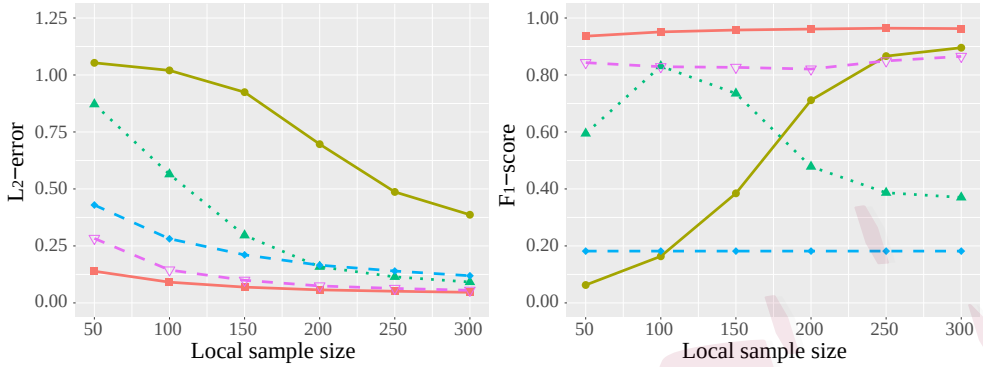

Figure 1: The ℓ_2-errors (left panel) and F_1-scores (right panel) of the Pooled (—■—), Local (—●—), Avg (···▲···), subGD (- -◆- -), and our estimates (- -▽- -) under different local sample sizes n with the number of nodes m fixed at 20.

To study the effect of the number of nodes, we fix the local sample size n at 100, and vary the number of nodes m in $\{5, 10, \ldots, 50\}$. Our finding, summarized in Figure 2, reveals that our estimate outperforms the decentralized competitors in terms of ℓ_2-errors in all settings and exhibits competitive support recovery performance. In addition, the performance of our decentralized estimate is robust to the number of nodes m, with nearly the same performance as the pooled estimate. These findings highlight the importance of leveraging information across nodes in decentralized networks to achieve superior estimation performance. The F_1-score exhibits a non-monotonic trend: it initially increases and subsequently declines. The low F_1-score at $m = 5$ reflects a violation of the condition $s^2 \log p/N = o(1)$, as stipulated in Theorem 2. The observed decline in performance arises because the generalized ADMM algorithm is not fully iterated so that the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

optimization error dominates compared to the statistical convergence rate $(s \log p/N)^{1/2}$ established in Theorem 1 as N increases.

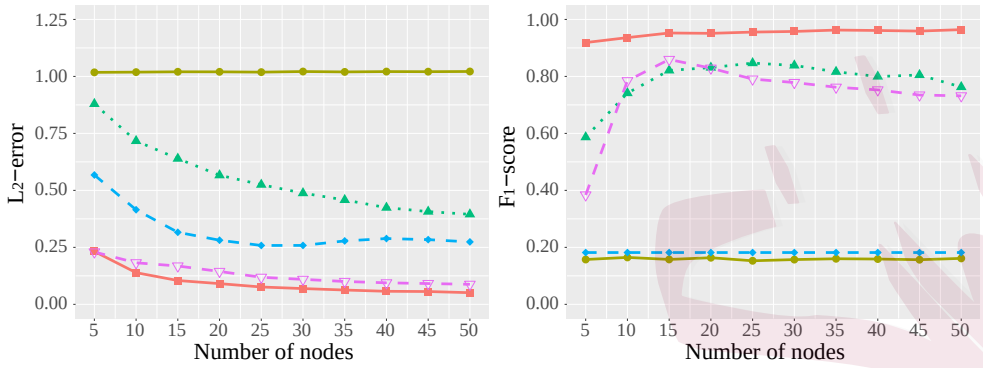

Figure 2: The ℓ_2-errors (left panel) and F_1-scores (right panel) of the Pooled (—■—), Local (—●—), Avg (···▲···), subGD (- -◆- -), and our estimates (- -▽- -) under different numbers of nodes with local sample size n fixed at 100.

Next, we examine the impact of network sparsity by varying the probability of network connection p_c in $\{0.3, 0.5, 0.8\}$. A smaller probability of network connection will lead to a sparser network. Our results, summarized in Figure 3, show that our proposal is resilient to network sparsity, with performance remaining similar across all settings. This is because the network sparsity only affects the convergence factor γ. However, the error term with γ diminishes at a linear rate (Proposition 1), so that it becomes negligible compared to the statistical error term.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

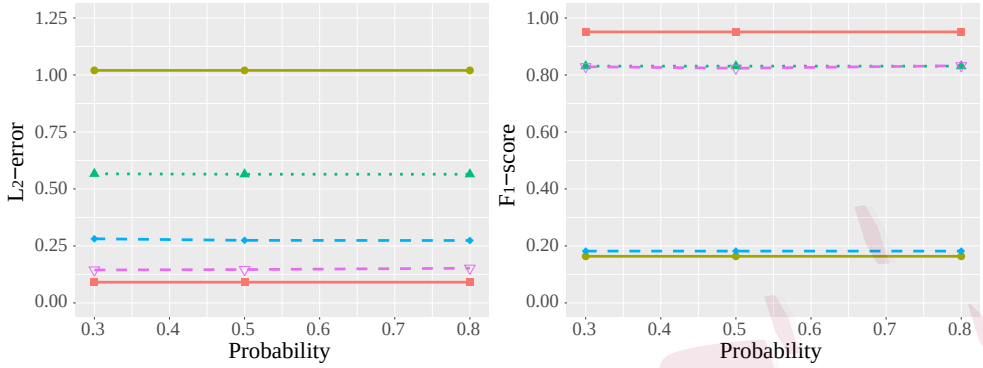

Figure 3: The ℓ_2-errors (left panel) and F_1-scores (right panel) of the Pooled (—■—), Local (—●—), Avg (···▲···), subGD (- -◆- -), and our estimates (- -▽- -) under different probabilities of network connection.

4.4 Effect of heterogeneity

In this subsection, we assess the effect of covariate heterogeneity. We fix the total sample size at $N = 2000$ and set the number of nodes to $m = 20$, so each node contains $n = 100$ observations. To introduce heterogeneity, each node considers covariate design from one of two Gaussian designs with probabilities $1 - \pi$ and π: the first with $\sigma_j^2 = 1$ and $\rho = 0.1$, and the second with $\sigma_j^2 = 3$ and $\rho = 0.3$. Using the standard definition of the SNR, these correspond to SNR $= 0.869$ and SNR $= 0.557$, repetitively. We vary π over $\{0, 0.2, 0.4, 0.6, 0.8\}$. Larger values of π correspond to higher heterogeneity across nodes.

The resulting performance is summarized in Figure 4. As heterogeneity increases, the ℓ_2-error for all methods rises modestly, and the F_1-score shows

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

a mild decrease. Our method remains the most competitive decentralized method. Overall, the results indicate that the competing methods remain notably stable even under substantial variation in covariate distributions. This kind of robustness is, in fact, the main benefit of the least squares formulation adopted in our method.

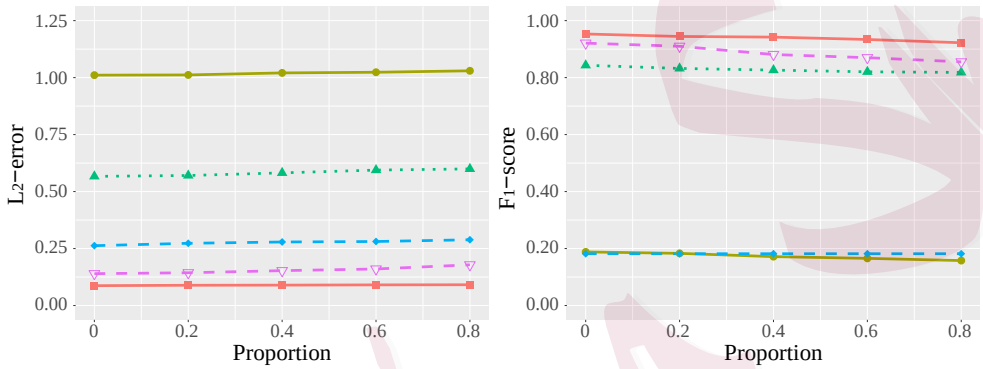

Figure 4: The ℓ_2-errors (left panel) and F_1-scores (right panel) of the Pooled (—■—), Local (—●—), Avg (···▲···), subGD (- - ◆ - -), and our estimates (- - ▽ - -) under different proportions of hetergenous covariates.

4.5 Effect of dimension, sparsity, and probability of sign flips

We investigate the impact of the dimension of the covariate p, sparsity, and probability of sign flips on the performance of different estimation methods. Results are summarized in Figures 5 to 7. We can find that as long as the dimension, the sparsity, and the probability of sign flips are not too large, the performance of our proposed decentralized estimate is quite close to that of the pooled estimate, and our proposal outperforms the rest of the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

decentralized competitors greatly. Accordingly, all the conclusions from the previous subsections hold for a large range of configurations of dimension, sparsity, and probability of sign flips and apply well to many real-world applications. For moderate values of s, the condition in Theorem 2 remains valid. By the definition of the F_1-score, when the proportion of correctly identified support does not change a lot, the F_1-score increases with s, which explains the observed upward trend in our estimate's F_1-score as s grows. In particular, we further report the false positive rate and true positive rate under different sparsity in Tables 1 and 2, respectively. As shown in Table 1, the FPR values of all methods increase with s, indicating that support recovery becomes more challenging as s increases. Table 2 shows that the TPR values of all methods decrease with s, further confirming the increased difficulty in support recovery with larger s.

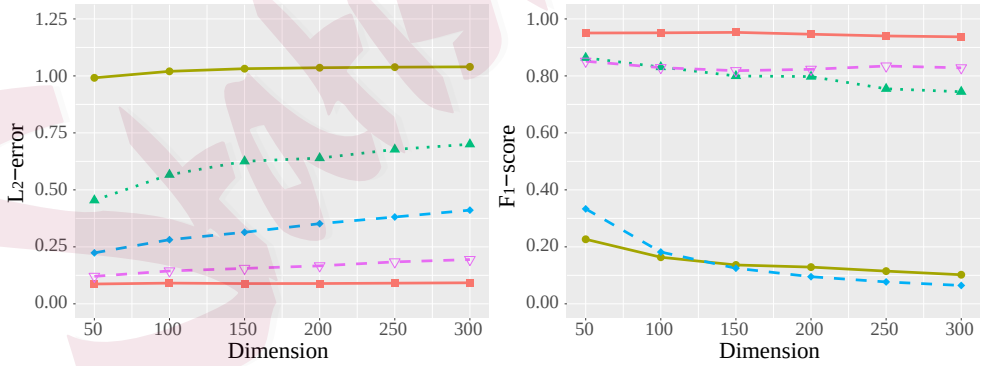

Figure 5: The ℓ_2-errors (left panel) and F_1-scores (right panel) of the Pooled (—■—), Local (—●—), Avg (···▲···), subGD (- - ◆ - -), and our estimates (- - ▽ - -) under different dimensions of the covariate.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

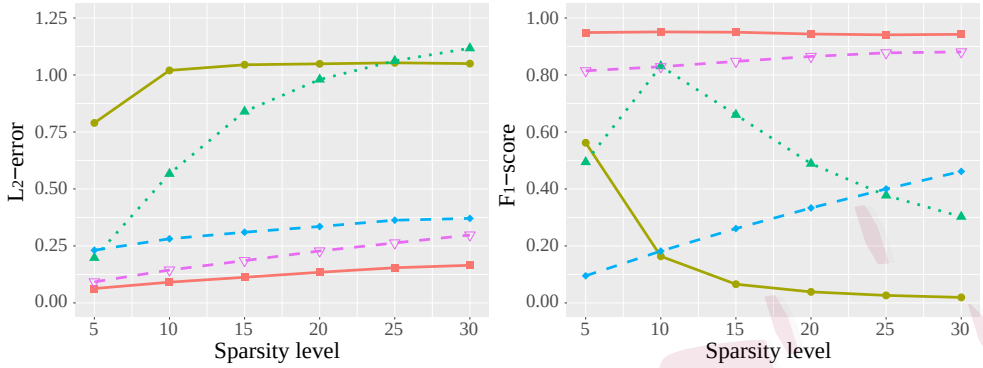

Figure 6: The ℓ_2-errors (left panel) and F_1-scores (right panel) of the Pooled (—■—), Local (—●—), Avg (···▲···), subGD (- -◆- -), and our estimates (- -▽- -) under different sparsity levels.

Table 1: False Positive Rate (FPR) under varying s.

s	Pooled	Local	Avg	subGD	Ours
5	0.0048	0.0060	0.1135	1.0000	0.0285
10	0.0112	0.0015	0.0296	1.0000	0.0498
15	0.0214	0.0008	0.0155	1.0000	0.0666
20	0.0321	0.0006	0.0121	1.0000	0.0808
25	0.0365	0.0006	0.0113	1.0000	0.0956
30	0.0523	0.0005	0.0103	1.0000	0.1069

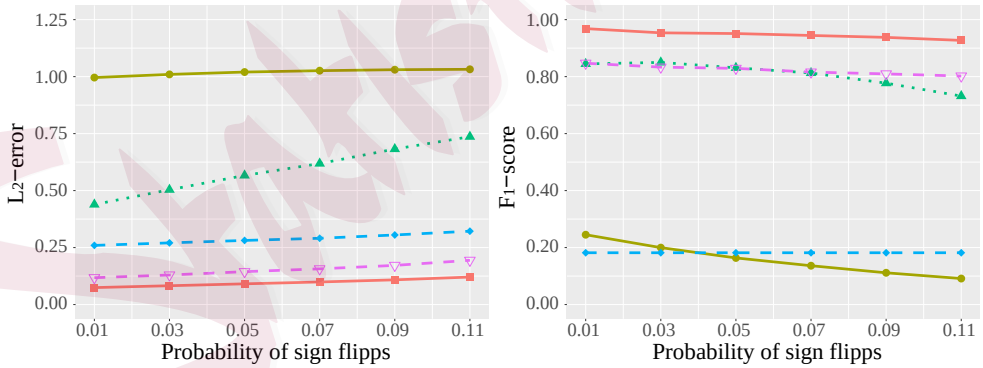

Figure 7: The ℓ_2-errors (left panel) and F_1-scores (right panel) of the Pooled (—■—), Local (—●—), Avg (···▲···), subGD (- -◆- -), and our estimates (- -▽- -) under different probabilities of sign flipps.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

Table 2: True Positive Rate (TPR) under varying s.

s	Pooled	Local	Avg	subGD	Ours
5	1.0000	0.5293	1.0000	1.0000	1.0000
10	1.0000	0.1112	0.9045	1.0000	1.0000
15	1.0000	0.0400	0.5653	1.0000	1.0000
20	1.0000	0.0227	0.3590	1.0000	1.0000
25	1.0000	0.0151	0.2590	1.0000	0.99997
30	1.0000	0.0105	0.1912	1.0000	0.99983

5. Conclusion

We propose an efficient decentralized algorithm for 1-bit CS over decentralized networks, which converges linearly to the statistically optimal solution with a straightforward implementation. To address the nondifferentiable sign function, we introduce a novel least squares formulation for 1-bit CS. This formulation imposes mild restrictions on the distribution of sensing vectors, allowing practitioners to select their preferred sensing vectors adaptively. In scenarios where communication is limited to neighborhoods, we develop a generalized ADMM algorithm to solve the resulting novel penalized least squares problem. Our focus in this work is on the ℓ_1-norm penalty. Notably, the generalized ADMM algorithm is versatile and can accommodate various advanced non-convex penalties through a simple linear approximation to the penalty (Zou and Li, 2008). In many applications, there is interest in recovering multiple signals that share common features while retaining individual innovations. Future research may explore personalized 1-bit com-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

pressive sensing by imposing group structure penalties to address these needs. Additionally, enhancing privacy by incorporating differential privacy techniques is a promising direction. Our current sign-flips setting already offers a degree of privacy protection. A key limitation of the current study is its focus on the regime $p < n$. Future work may extend the methodology to high-dimensional settings by incorporating structural assumptions such as an upper bound of signal sparsity to constrain the model space (Ji et al., 2023).

Supplementary Materials

The online supplementary materials provide additional simulation studies, an application to EEG signal reconstruction, and practical guidance for implementing the generalized ADMM algorithm, including choices of initialization, iteration number, and tuning parameters. Complete technical details are provided, including derivations of the algorithmic updates and full proofs of all theoretical results, covering key lemmas, propositions, and main theorems.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

Acknowledgments

Liping Zhu is the corresponding author. The authors contribute equally to this paper; their names are listed alphabetically. Liping Zhu is supported by the National Key R&D Program of China (2023YFA1008702) and the National Natural Science Foundation of China (12225113 and 12171477). Zhengtian Zhu is supported by the National Natural Science Foundation of China (12501411), and partially supported by National Key R&D Program of China (2024YFA1015700).

References

Arjoune, Y., N. Kaabouch, H. El Ghazi, and A. Tamtaoui (2018, July). A performance comparison of measurement matrices in compressive sensing. *International Journal of Communication Systems 31*(10), e3576.

Boufounos, P. T. and R. G. Baraniuk (2008). 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pp. 16–21.

Boyd, S. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning 3*(1), 1–122.

Candes, E. J., J. K. Romberg, and T. Tao (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences 59*(8), 1207–1223.

Candes, E. J. and T. Tao (2005). Decoding by linear programming. *IEEE Transactions on Information Theory 51*(12), 4203–4215.

Chang, T.-H., M. Hong, and X. Wang (2015). Multi-agent distributed optimization via inexact consensus admm. *IEEE Transactions on Signal Processing 63*(2), 482–497.

Chen, C. and L. Zhu (2023). Distributed Decoding From Heterogeneous 1-Bit Compressive Measurements. *Journal of Computational and Graphical Statistics 32*(3), 884–894.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics* (1 ed.). Wiley series in probability and statistics. Probability and statistics section. New York and Chichester: Wiley.

Dai, D.-Q., L. Shen, Y. Xu, and N. Zhang (2016). Noisy 1-bit compressive sensing: models and algorithms. *Applied and Computational Harmonic Analysis 40*(1), 1–32.

Deng, W. and W. Yin (2016, March). On the Global and Linear Convergence of the Generalized Alternating Direction Method of Multipliers. *Journal of Scientific Computing 66*(3), 889–916.

Diaconis, P. and D. Freedman (1984). Asymptotics of graphical projection pursuit. *The Annals of Statistics 12*(3), 793–815.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory 52*(4), 1289–1306.

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

Genzel, M. and P. Jung (2019). Recovering structured data from superimposed non-linear measurements. *IEEE Transactions on Information Theory 66*(1), 453–477.

Gu, Y., J. Fan, L. Kong, S. Ma, and H. Zou (2018, July). ADMM for high-dimensional sparse penalized quantile regression. *Technometrics 60*(3), 319–331.

Hall, P. and K.-C. Li (1993, June). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics 21*(2), 867–889.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations* (1st ed.). Chapman & Hall/CRC monographs on statistics & applied probability. Boca Raton: Chapman & Hall/CRC.

Hector, E. C. and P. X.-K. Song (2020). Doubly distributed supervised learning and inference with high-dimensional correlated outcomes. *Journal of Machine Learning Research 21*(173), 1–35.

Hector, E. C. and P. X.-K. Song (2021, April). A Distributed and Integrated Method of Moments for High-Dimensional Correlated Data Analysis. *Journal of the American Statistical Association 116*(534), 805–818.

Huang, J., Y. Jiao, X. Lu, and L. Zhu (2018, January). Robust decoding from 1-bit compres-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

sive sampling with ordinary and regularized least squares. *SIAM Journal on Scientific Computing 40*(4), A2062–A2086.

Ito, K. and B. Jin (2015). *Inverse Problems: Tikhonov Theory and Algorithms*. World Scientific.

Jacques, L., K. Degraux, and C. de Vleeschouwer (2013). Quantized iterative hard thresholding: Bridging 1-bit and high-resolution quantized compressed sensing. *arXiv preprint arXiv:1305.1786*.

Ji, Y., G. Scutari, Y. Sun, and H. Honnappa (2023, August). Distributed (atc) gradient descent for high dimension sparse regression. *IEEE Transactions on Information Theory 69*(8), 5253–5276.

Knudson, K., R. Saab, and R. Ward (2016). One-bit compressive sensing with norm estimation. *IEEE Transactions on Information Theory 62*(5), 2748–2758.

Laska, J. N., Z. Wen, W. Yin, and R. G. Baraniuk (2011). Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements. *IEEE Transactions on Signal Processing 59*(11), 5289–5301.

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

Li, K.-C. and N. Duan (1989). Regression Analysis Under Link Violation. *The Annals of Statistics 17*(3), 1009–1052.

Li, Z., W. Shi, and M. Yan (2019, September). A Decentralized Proximal-Gradient Method With Network Independent Step-Sizes and Separated Convergence Rates. *IEEE Transactions on Signal Processing 67*(17), 4494–4506.

Ling, Q. and Z. Tian (2010, July). Decentralized Sparse Signal Recovery for Compressive Sleeping Wireless Sensor Networks. *IEEE Transactions on Signal Processing 58*(7), 3816–3827.

Liu, W., D. Gong, and Z. Xu (2016). One-bit compressed sensing by greedy algorithms. *Numerical Mathematics: Theory, Methods and Applications 9*(2), 169–184.

Liu, W., X. Mao, and X. Zhang (2022). Fast and Robust Sparsity Learning Over Networks: A Decentralized Surrogate Median Regression Approach. *IEEE Transactions on Signal Processing 70*, 797–809.

Ma, S. and J. Huang (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association 112*(517), 410–423.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

Ma, Y. and L. Zhu (2012, March). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association 107*(497), 168–179.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing*. Elsevier.

Maly, J. and L. Palzer (2019). Analysis of hard-thresholding for distributed compressed sensing with one-bit measurements. *Information and Inference: A Journal of the IMA 9*(2), 455–471.

Nedic, A. and A. Ozdaglar (2009, January). Distributed Subgradient Methods for Multi-Agent Optimization. *IEEE Transactions on Automatic Control 54*(1), 48–61.

Qiao, N. and C. Chen (2024). Fast and robust low-rank learning over networks: A decentralized matrix quantile regression approach. *Journal of Computational and Graphical Statistics 33*(4), 1214–1223.

Rani, M., S. B. Dhok, and R. B. Deshmukh (2018). A systematic review of compressive sensing: Concepts, implementations and applications. *IEEE Access 6*, 4875–4894.

Shannon, C. (1949, January). Communication in the Presence of Noise. *Proceedings of the IRE 37*(1), 10–21.

Shi, W., Q. Ling, G. Wu, and W. Yin (2015). A Proximal Gradient Algorithm for Decentralized Composite Optimization. *IEEE Transactions on Signal Processing 63*(22), 6013–6023.

Shi, W., Q. Ling, K. Yuan, G. Wu, and W. Yin (2014, April). On the Linear Convergence of the ADMM in Decentralized Consensus Optimization. *IEEE Transactions on Signal Processing 62*(7), 1750–1761.

Stein, C. M. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics 9*(6), 1135–1151.

Sun, Y., P. Babu, and D. P. Palomar (2017, February). Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning. *IEEE Transactions on Signal Processing 65*(3), 794–816.

Tang, L., L. Zhou, and P. X.-K. Song (2020, March). Distributed simultaneous inference in generalized linear models via confidence distribution. *Journal of Multivariate Analysis 176*, 104567.

Tao, S., D. Boley, and S. Zhang (2016, January). Local Linear Convergence of ISTA and FISTA on the LASSO Problem. *SIAM Journal on Optimization 26*(1), 313–336.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory 55*(5), 2183–2202.

Wang, H., B. Li, and C. Leng (2009, June). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(3), 671–683.

Wang, H. and C. Li (2018, June). Distributed Quantile Regression Over Sensor Networks. *IEEE Transactions on Signal and Information Processing over Networks 4*(2), 338–348.

Wang, H., L. Xia, and C. Li (2019, April). Distributed online quantile regression over networks with quantized communication. *Signal Processing 157*, 141–150.

Xia, Y. (2007, December). A constructive approach to the estimation of dimension reduction directions. *Annals of Statistics 35*(6), 2654–2690.

Yan, M., Y. Yang, and S. Osher (2012). Robust 1-bit compressive sensing using adaptive outlier pursuit. *IEEE Transactions on Signal Processing 60*(7), 3868–3875.

Yaskov, P. (2014, September). Lower bounds on the smallest eigenvalue of a sample covariance matrix. *Electronic Communications in Probability 19*, 1–10.

Yuan, K., Q. Ling, and W. Yin (2016, January). On the convergence of decentralized gradient descent. *SIAM Journal on Optimization 26*(3), 1835–1854.

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

Zhang, J., K. You, and T. Başar (2019, June). Distributed Discrete-Time Optimization in Multiagent Networks Using Only Sign of Relative State. *IEEE Transactions on Automatic Control 64*(6), 2352–2367.

Zhou, L., X. She, and P. X.-K. Song (2023). Distributed empirical likelihood approach to integrating unbalanced datasets. *Statistica Sinica 33*, 2209–2231.

Zhu, L.-P. and L.-X. Zhu (2009). On distribution-weighted partial least squares with diverging number of highly correlated predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 525–548.

Zhu, L.-P., L.-X. Zhu, and Z.-H. Feng (2010, December). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Associa-*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0329

tion 105(492), 1455–1466.

Zhu, Y. (2017, January). An Augmented ADMM Algorithm With Application to the Generalized Lasso Problem. *Journal of Computational and Graphical Statistics 26*(1), 195–204.

Zou, H. and R. Li (2008, August). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics 36*(4), 1509–1533.

School of Statistics & Data Science, Southeast University, Nanjing 211189, China

E-mail: cychen.stats@outlook.com

Institute of Statistics and Big Data, Renmin University of China, Beijing 100875, China

E-mail: zhu.liping@ruc.edu.cn

School of Mathematical Sciences, School of Economics and Management, Key Laboratory of Intelligent Computing and Applications (Tongji University), Ministry of Education, Tongji University, Shanghai 200092, China

E-mail: zhengtianzhu@tongji.edu.cn