*The Annals of Statistics*
2018, Vol. 46, No. 3, 1352–1382
https://doi.org/10.1214/17-AOS1587


# DISTRIBUTED TESTING AND ESTIMATION UNDER SPARSE HIGH DIMENSIONAL MODELS

BY HEATHER BATTEY[*,†,1], JIANQING FAN[†,‡,2], HAN LIU[†], JUNWEI LU[†] AND ZIWEI ZHU[†]

*Imperial College London*[*], *Princeton University*[†] *and Fudan University*[‡]

This paper studies hypothesis testing and parameter estimation in the context of the divide-and-conquer algorithm. In a unified likelihood-based framework, we propose new test statistics and point estimators obtained by aggregating various statistics from $k$ subsamples of size $n/k$, where $n$ is the sample size. In both low dimensional and sparse high dimensional settings, we address the important question of how large $k$ can be, as $n$ grows large, such that the loss of efficiency due to the divide-and-conquer algorithm is negligible. In other words, the resulting estimators have the same inferential efficiencies and estimation rates as an oracle with access to the full sample. Thorough numerical results are provided to back up the theory.

**1. Introduction.** In recent years, the field of statistics has developed apace in response to the opportunities and challenges spawned from the "data revolution", which marked the dawn of an era characterized by the availability of enormous datasets. An extensive toolkit of methodology is now in place for addressing a wide range of high dimensional problems, whereby the number of unknown parameters, $d$, is much larger than the number of observations, $n$. However, many modern datasets are instead characterized by $n$ and $d$ both large. The latter presents intimidating practical challenges resulting from storage and computational limitations, as well as numerous statistical challenges [Fan, Han and Liu (2014)]. It is important that statistical methodology targeting modern application areas does not lose sight of the practical burdens associated with manipulating such large scale datasets. In this vein, incisive new algorithms have been developed for exploiting modern computing architectures and recent advances in distributed computing. These algorithms enjoy computational or communication efficiency and facilitate data handling and storage, but come with a statistical overhead if inappropriately tuned.

With increased mindfulness of the algorithmic difficulties associated with large datasets, the statistical community has witnessed a surge in recent activity in the

Received September 2015; revised December 2016.

[1]Supported in part by the NIH Grant 2R01-GM072611-11 as a postdoctoral fellow at Princeton University.

[2]Supported in part by NSF Grants DMS-1206464 and DMS-1406266, and NIH Grants 2R01-GM072611-11.

*MSC2010 subject classifications.* Primary 62F05, 62F10; secondary 62F12.

*Key words and phrases.* Divide and conquer, debiasing, massive data, thresholding.

statistical analysis of various divide-and-conquer (DC) algorithms, which randomly partition the $n$ observations into $k$ subsamples of size $n_k = n/k$, construct statistics based on each subsample, and aggregate them in a suitable way. In splitting the dataset, a single, very large scale estimation or testing problem with computational complexity $O(\gamma(n))$, for a given function $\gamma(\cdot)$ that depends on the underlying problem, is transformed into $k$ smaller problems with computational complexity $O(\gamma(n/k))$ on each machine. What gets lost in this process are the interactions of split subsamples in each machine. They are not recoverable without additional rounds of communication or without additional communication between the machines. Since every additional split of the dataset incurs some efficiency loss, it is of significant practical interest to derive a theoretical upper bound on the number of subsamples $k$ that delivers the same asymptotic statistical performance as the practically unavailable "oracle" procedure based on the full sample.

We develop communication efficient generalizations of the Wald and Rao's score tests for the sparse high dimensional scheme, as well as communication efficient estimators for the parameters of the sparse high dimensional and low dimensional linear and generalized linear models. In all cases, we give the upper bound on $k$ for preserving the statistical error of the analogous full sample procedure. While hypothesis testing in a low dimensional context is straightforward, in the sparse high dimensional setting, nuisance parameters introduce a nonnegligible bias, causing classical low dimensional theory to break down. In our high dimensional Wald construction, the phenomenon is remedied through a debiasing of the estimator, which gives rise to a test statistic with tractable limiting distribution, as documented in the $k = 1$ (no sample split) setting in Zhang and Zhang (2014) and van de Geer et al. (2014). For the high dimensional analogue of Rao's score statistic, the incorporation of a correction factor increases the convergence rate of higher order terms, thereby vanquishing the effect of the nuisance parameters. The approach is introduced in the $k = 1$ setting in Ning and Liu (2017), where the test statistic is shown to possess a tractable limit distribution. However, the computation complexity for the debiased estimators increases by an order of magnitude, due to solving $d$ high dimensional regularization problems. This motivates us to appeal to the divide-and-conquer strategy.

We develop the theory and methodology for DC versions of these tests. In the case of $k = 1$, each of the above test statistics can be decomposed into a dominant term with tractable limit distribution and a negligible remainder term. The DC extension requires delicate control of these remainder terms to ensure the error accumulation remains sufficiently small so as not to materially contaminate the leading term. We obtain an upper bound on the number of permitted subsamples, $k$, subject to a statistical guarantee. More specifically, we find that the theoretical upper bound on the number of subsamples guaranteeing the same inferential or estimation efficiency as the whole-sample procedure is $k = o((s \log d)^{-1}\sqrt{n})$ in the linear model, where $s$ is the sparsity of the parameter vector. In the generalized

linear model, the scaling is $k = o(((s \vee s_1) \log d)^{-1} \sqrt{n})$, where $s_1$ is the sparsity of the inverse information matrix.

For sparse high dimensional estimation problems, we use the same debiasing technique introduced in the high dimensional testing problems to obtain a thresholded divide-and-conquer estimator that achieves the full sample minimax rate. The appropriate scaling is found to be $k = O(\sqrt{n/(s^2 \log d)})$ for the estimation of the sparse parameter vector in the high dimensional linear model and $k = O(\sqrt{n/((s \vee s_1)^2 \log d)})$ for the high dimensional generalized linear model. Moreover, we find that the loss incurred by the divide-and-conquer strategy, as quantified by the distance between the DC estimator and the full sample estimator, is negligible in comparison to the statistical error of the full sample estimator provided that $k$ is not too large. In the context of estimation, the optimal scaling of $k$ with $n$ and $d$ is also developed for the low dimensional linear and generalized linear model. This theory is of independent interest. It also allows us to study a refitted estimation procedure under a minimal signal strength assumption.

1.1. *Related literature.* A partial list of references covering DC algorithms from a statistical perspective is Chen and Xie (2014), Zhang, Duchi and Wainwright (2015), Kleiner et al. (2014), Liu and Ihler (2014) and Zhao, Cheng and Liu (2016). The closest works to ours are Zhang, Duchi and Wainwright (2015), Lee et al. (2017) and Rosenblatt and Nadler (2016). Zhang, Duchi and Wainwright (2015) consider the distributed estimator for kernel ridge regression. In the context of $d < n$, Zhang, Duchi and Wainwright (2015) propose the distributed estimator by averaging the kernel ridge regression estimators for each data split. They obtain an explicit upper bound on the number of splits yielding the minimax optimal rates for the mean squared error. However, it is not straightforward to generalize their estimator to the high dimensional setting. In an independent work, Lee et al. (2017) propose the same debiasing approach of van de Geer et al. (2014) to allow aggregation of local estimates on distributed data splits in the context of sparse high dimensional linear and generalized linear models. Though using different techniques of proofs, the conclusions of Lee et al. (2017) in terms of the optimal choice of tuning parameter scaling and the upper bound on the permissible number of sample splits is of the same order. Our work differs from theirs in two aspects: (1) our work also contributes to the distributed testing in sparse high dimensional models and (2) we propose a refitted distributed estimator which has the oracle rate. Our results on hypothesis testing reveal a different phenomenon to that found in estimation, as we observe through the different requirements on the scaling of $k$. On the estimation side, our results also differ from those of Lee et al. (2017) in that our additional refitting step allows us to achieve the oracle rate. Rosenblatt and Nadler (2016) consider the distributed empirical risk minimization for $M$-estimators. They require the dimension of the interest parameter to satisfy the scaling condition $d/n \to \kappa \in (0, 1)$, which rules out the $d \gg n$ case. They

quantify the accuracy loss over the full sample estimator in terms of the number of splits.

1.2. *Organization of the paper.* The rest of the paper is organized as follows. Section 2 collects notation and details of a generic likelihood-based framework. Section 3 covers testing, providing high dimensional DC analogues of the Wald test (Section 3.1) and Rao's score test (Section 3.2), in each case deriving a tractable limit distribution for the corresponding test statistic under standard assumptions. Section 4 covers distributed estimation, proposing an aggregated estimator of the unknown parameters of linear and generalized linear models in low dimensional and sparse high dimensional scenarios, as well as a refitting procedure that improves the estimation rate, with the same scaling, under a minimal signal strength assumption. Section 5 provides numerical experiments to back up the developed theory. In Section 6, we discuss our results together with remaining future challenges. Proofs of our main results are collected in Section 7, while the statement and proofs of a number of technical lemmas are deferred to the Supplementary Material [Battey et al. (2018)].

**2. Background and notation.** We first collect the general notation, before providing a formal statement of our statistical problems. More specialized notation is introduced in context.

2.1. *Generic notation.* We adopt the common convention of using boldface letters for vectors only, while regular font is used for both matrices and scalars. $|\cdot|$ denotes both absolute value and cardinality of a set, with the context ensuring no ambiguity. For $\boldsymbol{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$, and $1 \leq q \leq \infty$, we define $\|\boldsymbol{x}\|_q = (\sum_{j=1}^d |x_j|^q)^{1/q}$, $\|\boldsymbol{x}\|_0 = |\operatorname{supp}(\boldsymbol{x})|$, where $\operatorname{supp}(\boldsymbol{x}) = \{j : x_j \neq 0\}$. Write $\|\boldsymbol{x}\|_\infty = \max_{1\leq j\leq d} |x_j|$, while for a matrix $M = [M_{jk}]$, let $\|M\|_{\max} = \max_{j,k} |M_{jk}|$, $\|M\|_1 = \sum_{j,k} |M_{jk}|$. For any matrix $M$, we use $\boldsymbol{M}_\ell$ to index the transposed $\ell$th row of $M$ and $[\boldsymbol{M}]_\ell$ to index the $\ell$th column. The sub-Gaussian norm of a scalar random variable $X$ is defined as $\|X\|_{\psi_2} = \sup_{q\geq 1} q^{-1/2}(\mathbb{E}|X|^q)^{1/q}$. For a random vector $\boldsymbol{X} \in \mathbb{R}^d$, its sub-Gaussian norm is defined as $\|\boldsymbol{X}\|_{\psi_2} = \sup_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \|\langle \boldsymbol{X}, \boldsymbol{x}\rangle\|_{\psi_2}$, where $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$. Let $I_d$ denote the $d \times d$ identity matrix; when the dimension is clear from the context, we omit the subscript. We also denote the Hadamard product of two matrices $A$ and $B$ as $A \circ B$ and $(A \circ B)_{jk} = A_{jk}B_{jk}$ for any $j, k$. $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$ denotes the canonical basis for $\mathbb{R}^d$. For a vector $\boldsymbol{v} \in \mathbb{R}^d$ and a set of indices $\mathcal{S} \subseteq \{1, \ldots, d\}$, $\boldsymbol{v}_\mathcal{S}$ is the vector of length $|\mathcal{S}|$ whose components are $\{v_j : j \in \mathcal{S}\}$. Additionally, for a vector $\boldsymbol{v}$ with $j$th element $v_j$, we use the notation $\boldsymbol{v}_{-j}$ to denote the remaining vector when the $j$th element is removed. With slight abuse of notation, we write $\boldsymbol{v} = (v_j, \boldsymbol{v}_{-j})$ when we wish to emphasize the dependence of $\boldsymbol{v}$ on $v_j$ and $\boldsymbol{v}_{-j}$ individually. The gradient of a function $f(\boldsymbol{x})$ is denoted by $\nabla f(\boldsymbol{x})$, while $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})$ denotes the gradient

of $f(\boldsymbol{x}, \boldsymbol{y})$ with respect to $\boldsymbol{x}$, and $\nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}, \boldsymbol{y})$ denotes the matrix of cross partial derivatives with respect to the elements of $\boldsymbol{x}$ and $\boldsymbol{y}$. For a scalar $\eta$, we simply write $f'(\eta) := \nabla_\eta f(\eta)$ and $f''(\eta) := \nabla^2_{\eta\eta} f(\eta)$. For a random variable $X$ and a sequence of random variables, $\{X_n\}$, we write $X_n \rightsquigarrow X$ when $\{X_n\}$ converges weakly to $X$. If $X$ is a random variable with standard distribution, say $F_X$, we simply write $X_n \rightsquigarrow F_X$. Given $a, b \in \mathbb{R}$, let $a \vee b$ and $a \wedge b$ denote the maximum and minimum of $a$ and $b$. We also make use of the notation $a_n \lesssim b_n$ ($a_n \gtrsim b_n$) if $a_n$ is less than (greater than) $b_n$ up to a constant, and $a_n \asymp b_n$ if $a_n$ is the same order as $b_n$.

2.2. *General likelihood-based framework.* Let $(\boldsymbol{X}_1^T, Y_1)^T, \ldots, (\boldsymbol{X}_n^T, Y_n)^T$ be $n$ i.i.d. copies of the random vector $(\boldsymbol{X}^T, Y)^T$, whose realizations take values in $\mathbb{R}^d \times \mathcal{Y}$. Write the collection of these $n$ i.i.d. random couples as $\mathcal{D} = \{(\boldsymbol{X}_1^T, Y_1)^T, \ldots, (\boldsymbol{X}_n^T, Y_n)^T\}$ with $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $X = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T \in \mathbb{R}^{n\times d}$. Conditional on $\boldsymbol{X}_i$, we assume $Y_i$ is distributed as $F_{\boldsymbol{\beta}^*}$ for all $i \in \{1, \ldots, n\}$, where $F_{\boldsymbol{\beta}^*}$ is a known distribution parameterized by a sparse $d$-dimensional vector $\boldsymbol{\beta}^*$ and has a density or mass function $f_{\boldsymbol{\beta}^*}$. We thus define the negative log-likelihood function, $\ell_n(\boldsymbol{\beta})$, as

$$\ell_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \ell_i(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^n \log f_{\boldsymbol{\beta}}(Y_i|\boldsymbol{X}_i). \tag{2.1}$$

We use $J^* = J(\boldsymbol{\beta}^*)$ to denote the information matrix and $\Theta^*$ to denote $(J^*)^{-1}$, where $J(\boldsymbol{\beta}) = \mathbb{E}[\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta})]$.

For testing problems, our goal is to test $H_0 : \beta_v^* = \beta_v^H$ for a specific fixed index $v \in \{1, \ldots, d\}$. We partition $\boldsymbol{\beta}^*$ as $\boldsymbol{\beta}^* = (\beta_v^*, \boldsymbol{\beta}_{-v}^{*T})^T \in \mathbb{R}^d$, where $\boldsymbol{\beta}_{-v}^* \in \mathbb{R}^{d-1}$ is a vector of nuisance parameters and $\beta_v^*$ is the parameter of interest. To handle the curse of dimensionality, we exploit a penalized M-estimator defined as

$$\widehat{\boldsymbol{\beta}}^\lambda = \operatorname*{argmin}_{\boldsymbol{\beta}} \{\ell_n(\boldsymbol{\beta}) + \mathcal{P}_\lambda(\boldsymbol{\beta})\}, \tag{2.2}$$

with $\mathcal{P}_\lambda(\boldsymbol{\beta})$ a sparsity inducing penalty function with a regularization parameter $\lambda$. Examples of $\mathcal{P}_\lambda(\boldsymbol{\beta})$ include the convex $\ell_1$ penalty, $\mathcal{P}_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1 = \lambda\sum_{j=1}^d |\beta_j|$ which, in the context of the linear model, gives rise to the Lasso estimator [Tibshirani (1996)],

$$\widehat{\boldsymbol{\beta}}^\lambda_{\text{Lasso}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{2n}\|\boldsymbol{Y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}. \tag{2.3}$$

Other penalties include folded concave penalties such as the smoothly clipped absolute deviation (SCAD) penalty [Fan and Li (2001)] and minimax concave MCP penalty [Zhang (2010)], which eliminate the estimation bias and attain the oracle rates of convergence [Loh and Wainwright (2013), Wang, Liu and Zhang (2014)].

The SCAD penalty is defined as

$$\mathcal{P}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{d} p_\lambda(\beta_j) \qquad \text{where}$$
$$p_\lambda(t) = \int_0^{|t|} \left\{ \lambda \mathbf{1}(z \le \lambda) + \frac{a\lambda - z}{a-1} \mathbf{1}(z > \lambda) \right\} dz, \tag{2.4}$$

for a given parameter $a > 0$ and MCP penalty is given by

$$\mathcal{P}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{d} p_\lambda(\beta_j) \qquad \text{where } p_\lambda(t) = \lambda \int_0^{|t|} \left(1 - \frac{z}{\lambda b}\right)_+ dz, \tag{2.5}$$

where $b > 0$ is a fixed parameter. The only requirement we have on $\mathcal{P}_\lambda(\boldsymbol{\beta})$ is that it induces an estimator satisfying the following condition.

CONDITION 2.1. For any $\delta \in (0, 1)$, if $\lambda \asymp \sqrt{\log(d/\delta)/n}$,

$$\mathbb{P}\big(\|\widehat{\boldsymbol{\beta}}^\lambda - \boldsymbol{\beta}^*\|_1 > Csn^{-1/2}\sqrt{\log(d/\delta)}\big) \le \delta, \tag{2.6}$$

where $s$ is the sparsity of $\boldsymbol{\beta}^*$, i.e., $s = \|\boldsymbol{\beta}^*\|_0$.

Condition 2.1 is crucial for the theory developed in Sections 3 and 4. Under suitable conditions on the design matrix $X$, it holds for the Lasso, SCAD and MCP. See Bühlmann and van de Geer (2011), Fan and Li (2001), Zhang (2010), respectively, and Zhang and Zhang (2012).

The DC algorithm randomly and evenly partitions $\mathcal{D}$ into $k$ disjoint subsets $\mathcal{D}_1, \ldots, \mathcal{D}_k$, so that $\mathcal{D} = \bigcup_{j=1}^k \mathcal{D}_j$, $\mathcal{D}_j \cap \mathcal{D}_\ell = \varnothing$ for all $j, \ell \in \{1, \ldots, k\}$, and $|\mathcal{D}_1| = |\mathcal{D}_2| = \cdots = |\mathcal{D}_k| = n_k = n/k$, where it is implicitly assumed that $n$ can be divided evenly. Let $\mathcal{I}_j \subset \{1, \ldots, n\}$ be the index set corresponding to the elements of $\mathcal{D}_j$. Then for an arbitrary $n \times d$ matrix $A$, $A^{(j)} = [A_{i\ell}]_{i \in \mathcal{I}_j, 1 \le \ell \le d}$. For an arbitrary estimator $\widehat{\tau}$, we write $\widehat{\tau}(\mathcal{D}_j)$ when the estimator is constructed based only on $\mathcal{D}_j$. Finally, we write $\ell_{n_k}^{(j)}(\boldsymbol{\beta}) = \sum_{i \in \mathcal{I}_j} \ell_i(\boldsymbol{\beta})$ to denote the negative log-likelihood function based on $\mathcal{D}_j$.

While the results of this paper hold in a general likelihood-based framework, for simplicity we state conditions at the population level for the generalized linear model (GLM) with a canonical link. A much more general set of statements appear in the auxiliary lemmas upon which our main results are based. Under GLM with the canonical link, the response follows the distribution,

$$f_n(\boldsymbol{Y}; X, \boldsymbol{\beta}^*) = \prod_{i=1}^{n} f(Y_i; \eta_i^*) = \prod_{i=1}^{n} \left\{ c(Y_i) \exp\left[ \frac{Y_i \eta_i^* - b(\eta_i^*)}{\phi} \right] \right\}, \tag{2.7}$$

where $\eta_i^* = \boldsymbol{X}_i^T \boldsymbol{\beta}^*$. The negative log-likelihood corresponding to (2.7) is given, up to an affine transformation, by

$$\ell_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} -Y_i \boldsymbol{X}_i^T \boldsymbol{\beta} + b(\boldsymbol{X}_i^T \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} -Y_i \eta_i + b(\eta_i) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}), \tag{2.8}$$

and the gradient and Hessian of $\ell_n(\boldsymbol{\beta})$ are respectively

$$\nabla \ell_n(\boldsymbol{\beta}) = -\frac{1}{n} X^T \big(\boldsymbol{Y} - \boldsymbol{\mu}(X\boldsymbol{\beta})\big) \quad \text{and} \quad \nabla^2 \ell_n(\boldsymbol{\beta}) = \frac{1}{n} X^T D(X\boldsymbol{\beta}) X,$$

where $\boldsymbol{\mu}(\boldsymbol{\beta}) = (b'(\eta_1), \ldots, b'(\eta_n))^T$ and $D(\boldsymbol{\beta}) = \operatorname{diag}\{b''(\eta_1), \ldots, b''(\eta_n)\}$. In this setting, $J(\boldsymbol{\beta}) = \mathbb{E}[b''(\boldsymbol{X}_1^T \boldsymbol{\beta}) \boldsymbol{X}_1 \boldsymbol{X}_1^T]$ and $J^* = \mathbb{E}[b''(\boldsymbol{X}_1^T \boldsymbol{\beta}^*) \boldsymbol{X}_1 \boldsymbol{X}_1^T]$.

**3. Divide-and-conquer hypothesis tests.** In the context of the two classical testing frameworks, the Wald and Rao's score tests, our objective is to construct a test statistic $\overline{S}_n$ with low communication cost and a tractable limiting distribution $F$. From this statistic, we define a test of size $\alpha$ of the null hypothesis, $H_0 : \beta_v^* = \beta_v^H$, against the alternative, $H_1 : \beta_v^* \neq \beta_v^H$, as a partition of the sample space described by

$$T_n^\alpha = \begin{cases} 0 & \text{if } |\overline{S}_n| \leq F^{-1}(1-\alpha/2), \\ 1 & \text{if } |\overline{S}_n| > F^{-1}(1-\alpha/2), \end{cases} \tag{3.1}$$

for a two-sided test.

3.1. *Two divide-and-conquer Wald-type constructions.* For the high dimensional linear model, Zhang and Zhang (2014), van de Geer et al. (2014) and Javanmard and Montanari (2014) propose methods for debiasing the Lasso estimator with a view to constructing high dimensional analogues of Wald statistics and confidence intervals for low dimensional coordinates. As pointed out by Zhang and Zhang (2014), the debiased estimator does not impose the minimum signal condition used in establishing oracle properties of regularized estimators [Fan and Li (2001), Fan and Lv (2011), Loh and Wainwright (2015), Wang, Liu and Zhang (2014), Zhang and Zhang (2012)], and hence has wider applicability than those inferences based on the oracle properties. The method of van de Geer et al. (2014) is appealing in that it accommodates a general penalized likelihood-based framework, while the Javanmard and Montanari (2014) approach is appealing in that it optimizes asymptotic variance and requires a weaker condition than van de Geer et al. (2014) in the specific case of the linear model. We consider the DC analogues of Javanmard and Montanari (2014) and van de Geer et al. (2014) in Sections 3.1.1 and 3.1.2, respectively.

3.1.1. *Lasso-based Wald test for the linear model.* The linear model assumes

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}^* + \varepsilon_i, \tag{3.2}$$

where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. with $\mathbb{E}(\varepsilon_i) = 0$ and variance $\sigma^2$. For concreteness, we focus on a Lasso-based method, but our procedure is also valid when other pilot estimators are used. We describe a modification of the bias correction method introduced in Javanmard and Montanari (2014) as a means to testing hypotheses on low dimensional coordinates of $\boldsymbol{\beta}^*$ via pivotal test statistics.

On each subset $\mathcal{D}_j$, we compute the debiased estimator of $\boldsymbol{\beta}^*$ as in Javanmard and Montanari (2014) as

$$\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j) = \widehat{\boldsymbol{\beta}}^{\lambda}_{\text{Lasso}}(\mathcal{D}_j) + \frac{1}{n_k} M^{(j)} \big(X^{(j)}\big)^T \big(Y^{(j)} - X^{(j)} \widehat{\boldsymbol{\beta}}^{\lambda}_{\text{Lasso}}(\mathcal{D}_j)\big), \tag{3.3}$$

where the superscript $d$ is used to indicate the debiased version of the estimator, $M^{(j)} = (\mathbf{m}_1^{(j)}, \ldots, \mathbf{m}_d^{(j)})^T$ and $\mathbf{m}_v$ is the solution of

$$\begin{aligned} \mathbf{m}_v^{(j)} = \operatorname*{argmin}_{\mathbf{m}} \mathbf{m}^T \widehat{\Sigma}^{(j)} \mathbf{m} \qquad & \text{s.t. } \big\| \widehat{\Sigma}^{(j)} \mathbf{m} - \boldsymbol{e}_v \big\|_\infty \le \vartheta_1, \\ & \big\| X^{(j)} \mathbf{m} \big\|_\infty \le \vartheta_2. \end{aligned} \tag{3.4}$$

The choice of tuning parameters $\vartheta_1$ and $\vartheta_2$ is discussed in Javanmard and Montanari (2014) and Zhao, Cheng and Liu (2016) and they suggest to choose $\vartheta_1 \asymp \sqrt{\log d/n}$, $\vartheta_2 n^{-1/2} = o(1)$. In the context of our DC procedure, $\vartheta_1$ and $\vartheta_2$ rely on $k$ and should be chosen as $\vartheta_1 \asymp \sqrt{k \log d/n}$, $\vartheta_2 n^{-1/2} = o(1)$, as quantified in Theorem 3.3. Above, $\widehat{\Sigma}^{(j)} = {n_k}^{-1} \sum_{i \in \mathcal{I}_j} \boldsymbol{X}_i^{(j)} \boldsymbol{X}_i^{(j)T}$ is the sample covariance based on $\mathcal{D}_j$, whose population counterpart is $\Sigma = \mathbb{E}(\boldsymbol{X}_1 \boldsymbol{X}_1^T)$ and $M^{(j)}$ is its regularized inverse. The second term in (3.3) is a bias correction term, while $\sigma^2 \mathbf{m}_v^{(j)T} \widehat{\Sigma}^{(j)} \mathbf{m}_v^{(j)} / n_k$ is shown in Javanmard and Montanari (2014) to be the variance of the $v$th component of $\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$. The parameter $\vartheta_1$, which tends to zero, controls the bias of the debiased estimator (3.3) and the optimization in (3.4) minimizes the variance of the resulting estimator.

Solving $d$ optimization problems in (3.4) increases an order of magnitude of computation complexity even for $k = 1$. Thus, it is necessary to appeal to the divide-and-conquer strategy to reduce the computation burden. This gives rise to the question how large $k$ can be in order to maintain the same statistical properties as the whole sample one ($k = 1$).

Because our DC procedure gives rise to smaller samples, $\widehat{\Sigma}$ is singular. This singularity does not pose a statistical problem but it does make the optimization problem ill-posed. To overcome the singularity in $\widehat{\Sigma}$ and the resulting instability of the algorithm, we propose a change of variables. More specifically, noting that $M^{(j)}$ is not required explicitly, but rather the product $M^{(j)}(X^{(j)})^T$, we

propose

$$\boldsymbol{b}_v^{(j)} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \frac{\boldsymbol{b}^{(j)T}\boldsymbol{b}^{(j)}}{n_k} \qquad \text{s.t.} \left\| \frac{X^{(j)T}\boldsymbol{b}^{(j)}}{n_k} - \boldsymbol{e}_v \right\|_\infty \le \vartheta_1, \quad \left\|\boldsymbol{b}^{(j)}\right\|_\infty \le \vartheta_2, \tag{3.5}$$

from which we construct $M^{(j)}(X^{(j)})^T = B^T$, where $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d)$. The algorithm in equation (3.5) is crucial to the success of our procedure in practice.

The following conditions on the data generating process and the tail behavior of the design vectors are imposed in Javanmard and Montanari (2014). Both conditions are used to derive the theoretical properties of the DC Wald test statistic based on the aggregated debiased estimator, $\overline{\boldsymbol{\beta}}^d = k^{-1}\sum_{j=1}^k \widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$.

CONDITION 3.1. $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ are i.i.d. and $\Sigma$ satisfies $0 < C_{\min} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le C_{\max}$.

CONDITION 3.2. The rows of $X$ are sub-Gaussian with $\|\boldsymbol{X}_i\|_{\psi_2} \le \kappa$, $i = 1, \ldots, n$.

Note that under the two conditions above, there exists a constant $\kappa_1 > 0$ such that $\|\boldsymbol{X}_1\Sigma^{-1/2}\|_{\psi_2} \le \kappa_1$. Without loss of generality, we set $\kappa_1 = \kappa$. Our first main theorem provides the relative scaling of the various tuning parameters involved in the construction of $\overline{\boldsymbol{\beta}}^d$.

THEOREM 3.3. *Suppose Conditions 2.1, 3.1 and 3.2 are fulfilled. Suppose* $\mathbb{E}[\varepsilon_1^4] < \infty$ *and choose* $\vartheta_1$, $\vartheta_2$ *and* $k$ *such that* $\vartheta_1 \asymp \sqrt{k\log d/n}$, $\vartheta_2 n^{-1/2} = o(1)$ *and* $k = o((s\log d)^{-1}\sqrt{n})$. *For any* $v \in \{1, \ldots, d\}$, *we have*

$$\sqrt{n}\frac{1}{k}\sum_{j=1}^k \frac{\widehat{\beta}_v^d(\mathcal{D}_j) - \beta_v^*}{\widehat{Q}_v^{(j)}} \rightsquigarrow N\big(0, \sigma^2\big), \tag{3.6}$$

*where* $\widehat{Q}_v^{(j)} = (\mathbf{m}_v^{(j)T}\widehat{\Sigma}^{(j)}\mathbf{m}_v^{(j)})^{1/2}$.

Theorem 3.3 entertains the prospect of a divide-and-conquer Wald statistic of the form

$$\overline{S}_n = \sqrt{n}\frac{1}{k}\sum_{j=1}^k \frac{\widehat{\beta}_v^d(\mathcal{D}_j) - \beta_v^H}{\overline{\sigma}(\mathbf{m}_v^{(j)T}\widehat{\Sigma}^{(j)}\mathbf{m}_v^{(j)})^{1/2}} \tag{3.7}$$

for $\beta_v^*$, where $\overline{\sigma}$ is an estimator for $\sigma$ based on the $k$ subsamples. On the left-hand side of equation (3.7), we suppress the dependence on $v$ to simplify notation. As

an estimator for $\sigma$, a simple suggestion with the same computational complexity is $\overline{\sigma}$ where

$$\overline{\sigma}^2 = \frac{1}{k}\sum_{j=1}^{k} \widehat{\sigma}^2(\mathcal{D}_j) \quad \text{and} \quad \widehat{\sigma}^2(\mathcal{D}_j) = \frac{1}{n_k}\sum_{i\in\mathcal{I}_j} \big(Y_i^{(j)} - \boldsymbol{X}_i^{(j)T}\widehat{\boldsymbol{\beta}}^{\lambda}_{\text{Lasso}}(\mathcal{D}_j)\big)^2. \tag{3.8}$$

One can use the refitted cross-validation procedure of Fan, Guo and Hao (2012) to reduce the bias of the estimate. In Lemma 3.4, we show that with the scaling of $k$ and $\lambda$ required for the weak convergence results of Theorem 3.3, consistency of $\overline{\sigma}^2$ is also achieved.

LEMMA 3.4. *Suppose* $\mathbb{E}[\varepsilon_i|\boldsymbol{X}_i] = 0$ *for all* $i \in \{1,\dots,n\}$. *Then with* $\lambda \asymp \sqrt{k\log d/n}$ *and* $k = o(\sqrt{n}(s\log d)^{-1})$, $|\overline{\sigma}^2 - \sigma^2| = o_{\mathbb{P}}(1)$.

With Lemma 3.4 and Theorem 3.3 at hand, we establish in Corollary 3.5 the asymptotic distribution of $\overline{S}_n$ under the null hypothesis $H_0 : \beta_v^* = \beta_v^H$. This holds for each component $v \in \{1,\dots,d\}$.

COROLLARY 3.5. *Suppose Conditions 3.1 and 3.2 are fulfilled,* $\mathbb{E}[\varepsilon_1^4] < \infty$, *and* $\lambda$, $\vartheta_1$ *and* $\vartheta_2$ *are chosen as* $\lambda \asymp \sqrt{k\log d/n}$, $\vartheta_1 \asymp \sqrt{k\log d/n}$ *and* $\vartheta_2 n^{-1/2} = o(1)$. *Then provided* $k = o((s\log d)^{-1}\sqrt{n})$, *under* $H_0 : \beta_v^* = \beta_v^H$, *we have*

$$\lim_{n\to\infty} \big|\mathbb{P}(\overline{S}_n \le t) - \Phi(t)\big| = 0, \tag{3.9}$$

*where* $\Phi(\cdot)$ *is the c.d.f. of a standard normal distribution.*

3.1.2. *Wald test in the likelihood-based framework.* An alternative route to debiasing the Lasso estimator of $\boldsymbol{\beta}^*$ is the one proposed in van de Geer et al. (2014). Their so-called desparsified estimator of $\boldsymbol{\beta}^*$ is more general than the debiased estimator of Javanmard and Montanari (2014) in that it accommodates generic estimators of the form (2.2) as pilot estimators, but the latter optimizes the variance of the resulting estimator. The desparsified estimator for subsample $\mathcal{D}_j$ is

$$\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j) = \widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \widehat{\Theta}^{(j)}\nabla\ell_{n_k}^{(j)}\big(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)\big), \tag{3.10}$$

where $\widehat{\Theta}^{(j)}$ is a regularized inverse of the Hessian matrix of second-order derivatives of $\ell_{n_k}^{(j)}(\boldsymbol{\beta})$ at $\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j)$, denoted by $\widehat{J}^{(j)} = \nabla^2\ell_{n_k}^{(j)}(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j))$. We will make this explicit in due course. The estimator resembles the classical one-step estimator [Bickel (1975)], but now in the high dimensional setting via regularized inverse of the Hessian matrix $\widehat{J}^{(j)}$, which reduces to the empirical covariance of the design matrix in the case of the linear model. From equation (3.10), the aggregated debiased estimator over the $k$ subsamples is defined as $\overline{\boldsymbol{\beta}}^d = k^{-1}\sum_{j=1}^{k}\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$.

We now use the nodewise Lasso [Meinshausen and Bühlmann (2006)] to approximately invert $\widehat{J}^{(j)}$ via $L_1$-regularization. The basic idea is to find the regularized invert row by row via a penalized $L_1$-regression, which is the same as

regressing the variable $X_v$ on $\boldsymbol{X}_{-v}$ but expressed in the sample covariance form. For each row $v \in 1, \ldots, d$, consider the optimization

$$\widehat{\boldsymbol{\kappa}}_v(\mathcal{D}_j) = \underset{\boldsymbol{\kappa} \in \mathbb{R}^{d-1}}{\operatorname{argmin}} \big( \widehat{J}^{(j)}_{vv} - 2\widehat{J}^{(j)}_{v,-v}\boldsymbol{\kappa} + \boldsymbol{\kappa}^T \widehat{J}^{(j)}_{-v,-v}\boldsymbol{\kappa} + 2\lambda_v \|\boldsymbol{\kappa}\|_1 \big), \tag{3.11}$$

where $\widehat{J}^{(j)}_{v,-v}$ denotes the $v$th row of $\widehat{J}^{(j)}$ without the $(v, v)$th diagonal element, and $\widehat{J}^{(j)}_{-v,-v}$ is the principal submatrix without the $v$th row and $v$th column. Introduce

$$\widehat{C}^{(j)} := \begin{pmatrix} 1 & -\widehat{\kappa}_{1,2}(\mathcal{D}_j) & \ldots & -\widehat{\kappa}_{1,d}(\mathcal{D}_j) \\ -\widehat{\kappa}_{2,1}(\mathcal{D}_j) & 1 & \ldots & -\widehat{\kappa}_{2,d}(\mathcal{D}_j) \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\kappa}_{d,1}(\mathcal{D}_j) & -\widehat{\kappa}_{d,2}(\mathcal{D}_j) & \ldots & 1 \end{pmatrix} \tag{3.12}$$

and $\widehat{\Xi}^{(j)} = \operatorname{diag}(\widehat{\tau}_1(\mathcal{D}_j), \ldots, \widehat{\tau}_d(\mathcal{D}_j))$, where $\widehat{\tau}_v(\mathcal{D}_j)^2 = \widehat{J}^{(j)}_{vv} - \widehat{J}^{(j)}_{v,-v}\widehat{\boldsymbol{\kappa}}_v(\mathcal{D}_j)$. $\widehat{\Theta}^{(j)}$ in equation (3.10) is given by

$$\widehat{\Theta}^{(j)} = \big(\widehat{\Xi}^{(j)}\big)^{-2}\widehat{C}^{(j)}, \tag{3.13}$$

and we define $\widehat{\boldsymbol{\Theta}}^{(j)}_v$ as the transposed $v$th row of $\widehat{\Theta}^{(j)}$.

Theorem 3.8 establishes the limit distribution of the term

$$\overline{S}_n = \sqrt{n}\frac{1}{k}\sum_{j=1}^{k} \frac{\widehat{\beta}^d_v(\mathcal{D}_j) - \beta^H_v}{\sqrt{\Theta^*_{vv}}} \tag{3.14}$$

for any $v \in \{1, \ldots, d\}$ under the null hypothesis $H_0 : \beta^*_v = \beta^H_v$. This provides the basis for the statistical testing based on divide and conquer. We need the following condition. Recall that $J^* = \mathbb{E}[\nabla_{\boldsymbol{\beta}\boldsymbol{\beta}}\ell_n(\boldsymbol{\beta}^*)]$ and consider the generalized linear model (2.7).

CONDITION 3.6. (i) $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ are i.i.d., $0 < C_{\min} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le C_{\max}$, $\lambda_{\min}(J^*) \ge L_{\min} > 0$, $\|J^*\|_{\max} < U_1 < \infty$. (ii) For some constant $M < \infty$, $\max_{1\le i\le n} |\boldsymbol{X}_i^T\boldsymbol{\beta}^*| \le M$ and $\max_{1\le i\le n} \|\boldsymbol{X}_i\|_\infty \le M$. (iii) There exist finite constants $U_2, U_3 > 0$ such that $b''(\eta) < U_2$ and $b'''(\eta) < U_3$ for all $\eta \in \mathbb{R}$.

The same assumptions appear in van de Geer et al. (2014). In the case of the Gaussian GLM, the condition on $\lambda_{\min}(J^*)$ reduces to the requirement that the covariance of the design has a minimal eigenvalue bounded away from zero, which is a standard assumption. We require $\|J^*\|_{\max} < \infty$ to control the estimation error of different functionals of $J^*$. The restriction in (ii) on the covariates and the projection of the covariates are imposed for technical simplicity; it can be extended to the case of exponential tails [see Fan and Song (2010)]. Note that $\operatorname{Var}(Y_i) = \phi b''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)$ where $\phi$ is the dispersion parameter in (2.7), so $b''(\eta) < U_2$ essentially implies an upper bound on the variance of the response. In

fact, Lemma E.2 shows that $b''(\eta) < U_2$ can guarantee that the response is sub-Gaussian. $b'''(\eta) < U_3$ is used to derive the Lipschitz property of $b''(\boldsymbol{X}_i^T\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ as shown in Lemma E.5. We emphasize that no requirement in Condition 3.6 is specific to the divide- and-conquer framework.

The assumption of bounded design in (ii) can be relaxed to the sub-Gaussian design. However, the price to pay is that the allowable number of subsets $k$ is smaller than the bounded case, which means we need a larger sub-sample size. To be more precise, the order of maximum $k$ for the sub-Gaussian design has an extra factor, which is a polynomial of $\sqrt{\log d}$, compared to the order for the bounded design. This logarithmic factor comes from different Lipschitz properties of $b''(\boldsymbol{X}_i^T\boldsymbol{\beta})$ in the two designs, which is fully explained in Lemma E.5 in the Supplementary Material [Battey et al. (2018)]. In the following theorems, we only present results for the case of bounded design for technical simplicity.

In addition, recalling that $\Theta^* = (J^*)^{-1}$, where $J^* := J(\boldsymbol{\beta}^*) = \mathbb{E}[\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}}\ell_n(\boldsymbol{\beta}^*)]$, we impose Condition 3.7 on $\Theta^*$ and its estimator $\widehat{\Theta}$.

CONDITION 3.7. (i) $\min_{1\le v\le d}\Theta^*_{vv} > \theta_{\min} > 0$. (ii) $\max_{1\le i\le n}\|\boldsymbol{X}_i^T\boldsymbol{\Theta}^*\|_\infty \le M$. (iii) For $v = 1,\dots,d$, whenever $\lambda_v \asymp \sqrt{k\log d/n}$ in (3.11), we have

$$\mathbb{P}\big(\|\widehat{\boldsymbol{\Theta}}_v - \boldsymbol{\Theta}^*_v\|_1 \ge Cs_1\sqrt{\log d/n}\big) \le d^{-1},$$

where $C$ is a constant and $s_1$ is such that $\|\boldsymbol{\Theta}^*_v\|_0 \lesssim s_1$ for all $v \in \{1,\dots,d\}$.

Part (i) of Corollary 3.7 ensures that the variances of each component of the debiased estimator exist, guaranteeing the existence of the Wald statistic. Parts (ii) and (iii) are imposed directly for technical simplicity. Results of this nature have been established under a similar set of assumptions in van de Geer et al. (2014) and Negahban et al. (2009) for convex penalties and in Wang, Liu and Zhang (2014) and Loh and Wainwright (2015) for folded concave penalties.

As a step towards deriving the limit distribution of the proposed divide-and-conquer Wald statistic in the GLM framework, we establish the asymptotic behavior of the aggregated debiased estimator $\overline{\beta}^d_v$ for every given $v \in [d]$.

THEOREM 3.8. *Under Conditions 2.1, 3.6 and 3.7, with* $\lambda \asymp \sqrt{k\log d/n}$, *we have*

$$\overline{\beta}^d_v - \beta^*_v = -\frac{1}{k}\sum_{j=1}^{k}\widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla\ell^{(j)}_{n_k}(\boldsymbol{\beta}^*) + o_{\mathbb{P}}(n^{-1/2}) \tag{3.15}$$

*for any* $k \ll d$ *satisfying* $k = o(((s\vee s_1)\log d)^{-1}\sqrt{n})$, *where* $\widehat{\boldsymbol{\Theta}}_v^{(j)}$ *is the transposed* $v$*th row of* $\widehat{\Theta}^{(j)}$.

The proof of Theorem 3.8 shows that for the Wald test procedure, the divide-and-conquer estimator $\overline{\beta}_v^d$ is asymptotically as efficient as the full sample estimator $\widehat{\beta}_v$, that is,

$$\lim_{n\to\infty} \frac{\mathrm{Var}(\overline{\beta}_v^d)}{\mathrm{Var}(\widehat{\beta}_v^d)} - 1 = 0.$$

A corollary of Theorem 3.8 provides the asymptotic distribution of the Wald statistic in equation (3.14) under the null hypothesis.

COROLLARY 3.9. *Let $\overline{S}_n$ be as in equation (3.14), with $\Theta^*_{vv}$ replaced with an estimator $\widetilde{\Theta}_{vv}$. Then under the conditions of Theorem 3.8 and $H_0 : \beta_v^* = \beta_v^H$, provided $|\widetilde{\Theta}_{vv} - \Theta_{vv}| = o_{\mathbb{P}}(1)$ under the scaling $k = o(((s \vee s_1)\log d)^{-1}\sqrt{n})$, we have*

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left|\mathbb{P}(\overline{S}_n \le t) - \Phi(t)\right| = 0.$$

REMARK 3.10. Although Theorem 3.8 and Corollary 3.9 are stated only for the GLM, their proofs are in fact an application of two more general results. Further details are available in Lemmas E.7 and E.8 in the Supplementary Material [Battey et al. (2018)].

We return to the issue of estimating $\Theta^*_{vv}$ in Section 4, where we introduce a consistent estimator of $\Theta^*_{vv}$ that preserves the scaling of Theorem 3.8 and Corollary 3.9.

3.2. *Divide-and-conquer score test.* In this section, we use $\nabla_v f(\boldsymbol{\beta})$ and $\nabla_{-v} f(\boldsymbol{\beta})$ to denote, respectively, the partial derivative of $f$ with respect to $\beta_v$ and the partial derivative vector of $f$ with respect to $\boldsymbol{\beta}_{-v}$. $\nabla^2_{vv} f(\boldsymbol{\beta})$, $\nabla^2_{v,-v} f(\boldsymbol{\beta})$, $\nabla^2_{-v,v} f(\boldsymbol{\beta})$ and $\nabla^2_{-v,-v} f(\boldsymbol{\beta})$ are analogously defined.

In the low dimensional setting (where $d$ is fixed), Rao's score test of $H_0 : \beta_v^* = \beta_v^H$ against $H_1 : \beta_v^* \neq \beta_v^H$ is based on $\nabla_v \ell_n(\beta_v^H, \widetilde{\boldsymbol{\beta}}_{-v})$, where $\widetilde{\boldsymbol{\beta}}_{-v}$ is a constrained maximum likelihood estimator of $\boldsymbol{\beta}^*_{-v}$, constructed as $\widetilde{\boldsymbol{\beta}}_{-v} = \mathrm{argmin}_{\boldsymbol{\beta}_{-v}} \ell_n(\beta_v^H, \boldsymbol{\beta}_{-v}) = \mathrm{argmax}_{\boldsymbol{\beta}_{-v}}\{-\ell_n(\beta_v^H, \boldsymbol{\beta}_{-v})\}$. If $H_0$ is false, imposing the constraint postulated by $H_0$ significantly violates the first-order conditions from M-estimation with high probability; this is the principle underpinning the classical score test. Under regularity conditions, it can be shown [e.g., Cox and Hinkley (1974)] that

$$\sqrt{n}\big(\nabla_v \ell_n\big(\beta_v^H, \widetilde{\boldsymbol{\beta}}_{-v}\big)\big) J^{*-1/2}_{v|-v} \rightsquigarrow N(0,1),$$

where $J^*_{v|-v}$ is given by $J^*_{v|-v} = J^*_{v,v} - \boldsymbol{J}^*_{v,-v} J^{*-1}_{-v,-v} \boldsymbol{J}^*_{-v,v}$, with $J^*_{v,v}$, $\boldsymbol{J}^*_{v,-v}$, $J^*_{-v,-v}$ and $\boldsymbol{J}^*_{-v,v}$ the partitions of the information matrix $J^* = J(\boldsymbol{\beta}^*)$,

$$J(\boldsymbol{\beta}) = \begin{pmatrix} J_{v,v} & \boldsymbol{J}_{v,-v} \\ \boldsymbol{J}_{-v,v} & J_{-v,-v} \end{pmatrix} = \begin{pmatrix} \mathbb{E}\nabla^2_{v,v}\ell_n(\boldsymbol{\beta}) & \mathbb{E}\nabla^2_{v,-v}\ell_n(\boldsymbol{\beta}) \\ \mathbb{E}\nabla^2_{-v,v}\ell_n(\boldsymbol{\beta}) & \mathbb{E}\nabla^2_{-v,-v}\ell_n(\boldsymbol{\beta}) \end{pmatrix}. \tag{3.16}$$

The problems associated with the use of the classical score statistic in the presence of a high dimensional nuisance parameter are brought to light by Ning and Liu (2017), who propose a remedy via the decorrelated score. The problem stems from the inversion of the matrix $J^*_{-v,-v}$ in high dimensions. The decorrelated score is defined as

$$S(\beta_v^*, \boldsymbol{\beta}_{-v}^*) = \nabla_v \ell_n(\beta_v^*, \boldsymbol{\beta}_{-v}^*) - \boldsymbol{w}^{*T} \nabla_{-v} \ell_n(\beta_v^*, \boldsymbol{\beta}_{-v}^*), \tag{3.17}$$

where $\boldsymbol{w}^{*T} = \boldsymbol{J}^*_{v,-v} J^{*-1}_{-v,-v}$. For a regularized estimator $\widehat{\boldsymbol{w}}$ of $\boldsymbol{w}^*$, to be defined below, we consider the score estimator

$$\widehat{S}(\beta_v^*, \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}) = \nabla_v \ell_n(\beta_v^*, \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}) - \widehat{\boldsymbol{w}}^T \nabla_{-v} \ell_n(\beta_v^*, \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}). \tag{3.18}$$

Hence, provided $\boldsymbol{w}^*$ is sufficiently sparse to avoid excessive noise accumulation, we are able to achieve consistency of $\widehat{\boldsymbol{w}}$ under the high dimensional setting, ultimately giving rise to a tractable limit distribution of a suitable rescaling of $\widehat{S}(\beta_v^*, \widehat{\boldsymbol{\beta}}_{-v}^{\lambda})$. Since $\beta_v^*$ is restricted under the null hypothesis, $H_0 : \beta_v^* = \beta_v^H$, the statistic in (3.18) is accessible once $H_0$ is imposed. As Ning and Liu (2017) point out, $\boldsymbol{w}^*$ is the solution to

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}} \mathbb{E}\big[\nabla_v \ell_n(\beta_v^H, \boldsymbol{\beta}_{-v}^*) - \boldsymbol{w}^T \nabla_{-v} \ell_n(\beta_v^H, \boldsymbol{\beta}_{-v}^*)\big]^2$$

under $H_0 : \beta_v^* = \beta_v^H$.

Our divide-and-conquer score statistic under $H_0 : \beta_v^* = \beta_v^H$ is

$$\overline{S}(\beta_v^H) = \frac{1}{k} \sum_{j=1}^{k} \widehat{S}^{(j)}(\beta_v^H, \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}(\mathcal{D}_j)), \tag{3.19}$$

where

$$\widehat{S}^{(j)}(\beta_v, \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}(\mathcal{D}_j)) = \nabla_v \ell_{n_k}^{(j)}(\beta_v, \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}(\mathcal{D}_j)) - \widehat{\boldsymbol{w}}(\mathcal{D}_j)^T \nabla_{-v} \ell_{n_k}^{(j)}(\beta_v, \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}(\mathcal{D}_j))$$

and we estimate $\boldsymbol{w}^*$ using the Dantzig selector of Candes and Tao (2007)

$$\widehat{\boldsymbol{w}}(\mathcal{D}_j) = \operatorname*{argmin}_{\boldsymbol{w}} \|\boldsymbol{w}\|_1, \qquad \text{s.t.}$$

$$\big\|\nabla^2_{-v,v} \ell_{n_k}^{(j)}(\widehat{\beta}_v^{\lambda}(\mathcal{D}_j), \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}(\mathcal{D}_j)) - \boldsymbol{w}^T \nabla^2_{-v,-v} \ell_{n_k}^{(j)}(\widehat{\beta}_v^{\lambda}(\mathcal{D}_j), \widehat{\boldsymbol{\beta}}_{-v}^{\lambda}(\mathcal{D}_j))\big\|_{\infty} \le \mu.$$

THEOREM 3.11. *Let $\widehat{J}_{v|-v}$ be a consistent estimator of $J^*_{v|-v}$ and*

$$S^{(j)}(\beta_v^H, \boldsymbol{\beta}_{-v}^*) = \nabla_v \ell_{n_k}^{(j)}(\beta_v^H, \boldsymbol{\beta}_{-v}^*) - \boldsymbol{w}^{*T} \nabla_{-v} \ell_{n_k}^{(j)}(\beta_v^H, \boldsymbol{\beta}_{-v}^*).$$

*Suppose $\|\boldsymbol{w}^*\|_1 \lesssim s_1$ and Conditions 2.1 and 3.6 are fulfilled. Then under $H_0 : \beta_v^* = \beta_v^H$ with $\lambda \asymp \mu \asymp \sqrt{k \log d / n}$,*

$$\sqrt{n}\overline{S}(\beta_v^H) = \sqrt{n} \frac{1}{k} \sum_{j=1}^{k} S^{(j)}(\beta_v^H, \boldsymbol{\beta}_{-v}^*) + o_{\mathbb{P}}(1) \quad \textit{and}$$

$$\lim_{n \to \infty} \sup_{t \in \mathbb{R}} \big|\mathbb{P}\big(\sqrt{n} \cdot \overline{S}(\beta_v^H) \widehat{J}_{v|-v}^{-1/2} \le t\big) - \Phi(t)\big| = 0,$$

*for any* $k \ll d$ *satisfying* $k = o(((s \vee s_1)\log d)^{-1}\sqrt{n})$, *where* $\overline{S}(\beta_v^H)$ *is defined in equation* (3.19).

REMARK 3.12. By the definition of $\boldsymbol{w}^*$ and the block matrix inversion formula for $\Theta^* = (J^*)^{-1}$, sparsity of $\boldsymbol{w}^*$ is implied by sparsity of $\Theta^*$ as assumed in van de Geer et al. (2014) and Condition 3.7 of Section 3.1.2. In turn, $\|\boldsymbol{w}^*\|_0 \lesssim s_1$ implies $\|\boldsymbol{w}^*\|_1 \lesssim s_1$ provided that the elements of $\boldsymbol{w}^*$ are bounded.

REMARK 3.13. Although Theorem 3.11 is stated in the penalized GLM setting, the result holds more generally; further details are available in Lemma E.13 in the Supplementary Material [Battey et al. (2018)].

To maintain the same computational complexity, an estimator of the conditional information needs to be constructed using a DC procedure. For this, we propose to use

$$\overline{J}_{v|-v} = k^{-1}\sum_{j=1}^{k}\big(\nabla^2_{v,v}\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v}) - \overline{\boldsymbol{w}}^T\nabla^2_{-v,v}\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v})\big),$$

where the divide-and-conquer estimator $\overline{\beta}^d_v = k^{-1}\sum_{j=1}^k \widehat{\beta}^d_v(\mathcal{D}_j)$, $\overline{\boldsymbol{\beta}}_{-v} = k^{-1}\times \sum_{j=1}^k \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j)$ and $\overline{\boldsymbol{w}} = k^{-1}\sum_{j=1}^k \widehat{\boldsymbol{w}}(\mathcal{D}_j)$. Note that for certain $v$, the communication cost for calculating $\overline{J}_{v|-v}$ is not high, since all the involved quantities $\{\nabla^2_{v,v}\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v})\}_{j=1}^k$, $\{\nabla^2_{-v,v}\ell^{(j)}_{n_k}(\overline{\beta}^d_v, \overline{\boldsymbol{\beta}}_{-v})\}_{j=1}^k$ and $\{\widehat{\boldsymbol{w}}(\mathcal{D}_j)\}_{j=1}^k$ are scalars, $d$-dimensional vectors and $d$-dimensional vectors, respectively. The communication cost is thus of order $O(kd)$. We do not communicate the entire huge Hessian matrix here.

LEMMA 3.14. *Suppose* $\|\boldsymbol{w}^*\|_1 = O(s_1)$ *and Conditions* 2.1 *and* 3.6 *are fulfilled. Then for any* $k \ll d$ *satisfying* $k = o(((s \vee s_1)\log d)^{-1}\sqrt{n})$, $|\overline{J}_{v|-v} - J^*_{v|-v}| = o_{\mathbb{P}}(1)$.

By Lemma 3.14, we show that $\overline{J}_{v|-v}$ is consistent and can be applied to Theorem 3.11.

**4. Accuracy of distributed estimation.** This section focuses on high dimensional ($d \gg n$) divide-and-conquer estimators for linear and generalized linear models. As explained below Theorem 3.8 in Section 3, the efficiency loss from the divide-and-conquer process is asymptotically zero. This motivates us to consider $\|\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d\|$, the loss incurred by the divide-and-conquer strategy in comparison with the practically unavailable full sample debiased estimator $\widehat{\boldsymbol{\beta}}^d$, where $\|\cdot\|$ is certain norm. Indeed, it turns out that, for $k$ not too large, $\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d$ appears only

as a higher order term in the decomposition of $\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*$ and thus $\|\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d\|$ is negligible compared to the statistical error, $\|\widehat{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|$. In other words, the divide-and-conquer errors are statistically negligible.

Compared with calculating the full sample debiased Lasso estimator, our proposed DC strategy enjoys computational advantages since it is highly parallel and each subsample problem has a much smaller scale than the full sample problem given a suitably large $k$. However, relative to just the full sample penalized M-estimator (e.g., Lasso), distributed point estimation does not entail a computational gain like distributed testing, since our distributed algorithm requires debiasing each component of the Lasso estimator, and hence brings high expense of computation. The bottleneck of computation of our DC procedure comes from the $d$ extra debiasing steps. To mitigate this problem, we can actually debias each component of $\widehat{\boldsymbol{\beta}}$ in a parallel fashion. According to the optimization procedures (3.4) and (3.11), debiasing one component of the Lasso estimator is entirely independent of the debiasing of another component. Therefore, as long as each branch computer in the cluster shares the sub-dataset $\mathcal{D}_j$ and the Lasso estimator $\widehat{\boldsymbol{\beta}}^{(j)}$, they can work in parallel and collectively return to a central server all the components of the debiased Lasso estimator. This parallelization reduces the time complexity significantly.

When the minimum signal strength is sufficiently strong, thresholding $\overline{\boldsymbol{\beta}}^d$ achieves exact support recovery, motivating a refitting procedure based on the low dimensional selected variables. As a means to understanding the theoretical properties of this refitting procedure, as well as for independent interest, we develop new theories and methodologies for the low dimensional ($d < n$) linear and generalized linear models in Appendices A and B in the Supplementary Material [Battey et al. (2018)], respectively. We show that simple averaging of low dimensional OLS or GLM estimators (denoted uniformly as $\widehat{\boldsymbol{\beta}}^{(j)}$, without superscript $d$ as debiasing is not necessary) suffices to preserve the statistical error, that is, achieving the same statistical accuracy as the estimator based on the full sample. This is because, in contrast to the high dimensional setting, parameters are not penalized in the low dimensional case. With $\overline{\boldsymbol{\beta}}$ the average of $\widehat{\boldsymbol{\beta}}^{(j)}$ over the $k$ machines and $\widehat{\boldsymbol{\beta}}$ the full sample counterpart ($k = 1$), we derive the rate of convergence of $\|\overline{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2$. Refitted estimation using only the selected covariates allows us to eliminate the $\log d$ term in the statistical rate of convergence of the estimator under high dimensional settings. We present theoretical results on the refitting estimation as corollaries to the low dimensional regression results in Appendices A and B in the Supplementary Material [Battey et al. (2018)].

4.1. *The high dimensional linear model.* Recall that the high dimensional DC estimator is $\overline{\boldsymbol{\beta}}^d = k^{-1}\sum_{j=1}^k \widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$, where $\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$ for $1 \le j \le k$ is the debiased estimator defined in (3.3). We also denote the debiased Lasso estimator using the

entire dataset as $\widehat{\boldsymbol{\beta}}^d = \widehat{\boldsymbol{\beta}}^d(\bigcup_{j=1}^k \mathcal{D}_j)$. The following lemma shows that not only is $\overline{\boldsymbol{\beta}}^d$ asymptotically normal, it approximates the full sample estimator $\widehat{\boldsymbol{\beta}}^d$ so well that it has the same statistical error as $\widehat{\boldsymbol{\beta}}^d$ provided the number of subsamples $k$ is not too large.

LEMMA 4.1. *Consider the linear model* (3.2). *Under the Conditions* 3.1 *and* 3.2, *if* $\lambda$, $\vartheta_1$ *and* $\vartheta_2$ *are chosen as* $\lambda \asymp \sqrt{k \log d/n}$, $\vartheta_1 \asymp \sqrt{k \log d/n}$ *and* $\vartheta_2 n^{-1/2} = o(1)$, *we have with probability* $1 - c/d$,

$$\|\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d\|_\infty \le C\frac{sk\log d}{n} \quad \text{and} \quad \|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty \le C\left(\sqrt{\frac{\log d}{n}} + \frac{sk\log d}{n}\right). \tag{4.1}$$

REMARK 4.2. The term $\sqrt{\log d/n}$ in (4.1) is the estimation error of $\|\widehat{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty$, while the term $(sk\log d)/n$ is the rate of the distance between the divide-and-conquer estimator and the full sample estimator. Lemma 4.1 does not rely on any specific choice of $k$. However, in order for the aggregated estimator $\overline{\boldsymbol{\beta}}^d$ to attain the same $\|\cdot\|_\infty$ norm estimation error as the full sample Lasso estimator, $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}$, the required scaling is $k = O(\sqrt{n/(s^2\log d)})$. This is a weaker scaling requirement than that of Theorem 3.3 because the latter entails a guarantee of asymptotic normality, which is a stronger result. It is for the same reason that our estimation results only require $O(\cdot)$ scaling while those for testing require $o(\cdot)$ scaling.

Rosenblatt and Nadler (2016) show that in the high dimensional regime where $d/n_k \to \kappa \in (0, 1)$, the divide-and-conquer procedure suffers from first-order accuracy loss. This seems a contradiction to our result, since our dimension is even higher than their context, but we have no first-order accuracy loss while averaging debiased estimators based on subsamples, as long as we have an appropriate number of data splits. In fact, in the high dimensional sparse linear regression, the intrinsic dimension is the sparsity $s$ rather than $d$, which is regarded instead as the ambient dimension. The sparsity assumption changes the original high dimensional problem to be an intrinsically low dimensional one, and thus allows us to escape from any first-order accuracy loss of the divide- and-conquer procedure. Given $s = o(n_k)$, we can treat high dimensional sparse linear regression approximately as the classical linear regression setting where $d = o(n_k)$. Hence we expect no first-order accuracy loss from the divide-and-conquer procedure here.

Although $\overline{\boldsymbol{\beta}}^d$ achieves the same rate as the Lasso estimator under the infinity norm, it cannot achieve the minimax rate in $\ell_2$ norm since it is not a sparse estimator. To obtain an estimator with the $\ell_2$ minimax rate, we sparsify $\overline{\boldsymbol{\beta}}^d$ by hard thresholding. For any $\boldsymbol{\beta} \in \mathbb{R}^d$, define the hard thresholding operator $\mathcal{T}_\nu$ such that the $j$th entry of $\mathcal{T}_\nu(\boldsymbol{\beta})$ is

$$\big[\mathcal{T}_\nu(\boldsymbol{\beta})\big]_j = \boldsymbol{\beta}_j \mathbb{1}\{|\boldsymbol{\beta}_j| \ge \nu\} \qquad \text{for } 1 \le j \le d. \tag{4.2}$$

According to (4.1), if $\boldsymbol{\beta}_j^* = 0$, we have $|\overline{\boldsymbol{\beta}}_j^d| \le C(\sqrt{\log d/n} + sk\log d/n)$ with high probability. The following theorem characterizes the estimation error, $\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_2$, and divide-and-conquer error, $\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \mathcal{T}_\nu(\widehat{\boldsymbol{\beta}}^d)\|_2$, of the thresholded estimator $\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d)$.

THEOREM 4.3. *Under the linear model* (3.2), *suppose Conditions* 3.1 *and* 3.2 *are fulfilled and choose* $\lambda \asymp \sqrt{k\log d/n}$, $\vartheta_1 \asymp \sqrt{k\log d/n}$ *and* $\vartheta_2 n^{-1/2} = o(1)$. *Take the parameter of the hard threshold operator in* (4.2) *as* $\nu = C_0\sqrt{\log d/n}$ *for some sufficiently large constant* $C_0$. *If the number of subsamples satisfies* $k = O(\sqrt{n/(s^2\log d)})$, *for large enough* $d$ *and* $n$, *we have with probability* $1 - c/d$,

$$\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \mathcal{T}_\nu(\widehat{\boldsymbol{\beta}}^d)\|_2 \le C\frac{s^{3/2}k\log d}{n}, \qquad \|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_\infty \le C\sqrt{\frac{\log d}{n}} \quad \text{and}$$

$$\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_2 \le C\sqrt{\frac{s\log d}{n}}.$$

REMARK 4.4. In fact, in the proof of Theorem 4.3, we show that if the thresholding parameter $\nu$ satisfies $\nu \ge \|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty$, we have $\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_2 \le 2\sqrt{2s}\cdot\nu$; it is for this reason that we choose $\nu \asymp \sqrt{\log d/n}$. Unfortunately, the constant is difficult to choose in practice. In the following paragraphs, we propose a practical method to select the tuning parameter $\nu$.

Let $(M^{(j)}X^{(j)T})_\ell$ denote the transposed $\ell$th row of $M^{(j)}X^{(j)T}$. Inspection of the proof of Theorem 3.3 reveals that the leading term of $\sqrt{n}\|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty$ satisfies

$$T_0 = \max_{1\le\ell\le d}\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\frac{1}{\sqrt{n_k}}\big(M^{(j)}X^{(j)T}\big)_\ell^T\boldsymbol{\varepsilon}^{(j)}.$$

Chernozhukov, Chetverikov and Kato (2013) propose the Gaussian multiplier bootstrap to estimate the quantile of $T_0$. Let $\{\xi_i\}_{i=1}^n$ be i.i.d. standard normal random variable independent of $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$. Consider the statistic

$$W_0 = \max_{1\le\ell\le d}\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\frac{1}{\sqrt{n_k}}\big(M^{(j)}X^{(j)T}\big)_\ell^T\big(\widehat{\boldsymbol{\varepsilon}}^{(j)}\circ\boldsymbol{\xi}^{(j)}\big),$$

where $\widehat{\boldsymbol{\varepsilon}}^{(j)} \in \mathbb{R}^{n_k}$ is an estimator of $\boldsymbol{\varepsilon}^{(j)}$ such that for any $i \in \mathcal{I}_j$, $\widehat{\varepsilon}_i^{(j)} = Y_i^{(j)} - \boldsymbol{X}_i^{(j)}\widehat{\boldsymbol{\beta}}(\mathcal{D}_j)$, and $\boldsymbol{\xi}^{(j)}$ is a subvector of $\{\xi_i\}_{i=1}^n$ with indices in $\mathcal{I}_j$. Recall that "$\circ$" denotes the Hadamard product. The $\alpha$-quantile of $W_0$ conditioning on $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$ is defined as $c_{W_0}(\alpha) = \inf\{t \mid \mathbb{P}(W_0 \le t \mid \boldsymbol{Y}, X) \ge \alpha\}$. We estimate $c_{W_0}(\alpha)$ by Monte Carlo, and thus choose $\nu_0 = c_{W_0}(\alpha)/\sqrt{n}$. This choice ensures

$$\|\mathcal{T}_{\nu_0}(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}(\sqrt{s\log d/n}),$$

which coincides with the $\ell_2$ convergence rate of the Lasso.

REMARK 4.5. Lemma 4.1 and Theorem 4.3 show that if the number of subsamples satisfies $k = o(\sqrt{n/(s^2 \log d)})$, $\|\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d\|_\infty = o_{\mathbb{P}}(\sqrt{\log d/n})$ and $\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \mathcal{T}_\nu(\widehat{\boldsymbol{\beta}}^d)\|_2 = o_{\mathbb{P}}(\sqrt{s \log d/n})$, and thus the error incurred by the divide-and-conquer procedure is negligible compared to the statistical minimax rate. The reason for this contraction phenomenon is that $\overline{\boldsymbol{\beta}}^d$ and $\widehat{\boldsymbol{\beta}}^d$ share the same leading term in their Taylor expansions around $\boldsymbol{\beta}^*$. The difference between them is only the difference of two remainder terms which has a smaller order than the leading term. We uncover a similar phenomenon in the low dimensional case covered in Appendix A in the Supplementary Material [Battey et al. (2018)]. However, in the low dimensional case, $\ell_2$ norm consistency is automatic while the high dimensional case requires an additional thresholding step to guarantee sparsity and, consequently, $\ell_2$ norm consistency.

4.2. *The high dimensional generalized linear model.* We generalize the DC estimation of the linear model to GLM. Recall that $\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$ is the de-biased estimator defined in (3.10) and the aggregated estimator is $\overline{\boldsymbol{\beta}}^d = k^{-1}\sum_{j=1}^k \widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j)$. We still denote $\widehat{\boldsymbol{\beta}}^d = \widehat{\boldsymbol{\beta}}^d(\bigcup_{j=1}^k \mathcal{D}_j)$. The next lemma bounds the error incurred by splitting the sample and the statistical rate of convergence of $\overline{\boldsymbol{\beta}}^d$ in terms of the infinity norm.

LEMMA 4.6. *Consider the generalized linear model* (2.7) *with canonical link. Under Conditions* 2.1, 3.6 *and* 3.7, *for* $\widehat{\boldsymbol{\beta}}^\lambda$ *with* $\lambda \asymp \sqrt{k \log d/n}$, *we have with probability* $1 - c/d$, *there exists a constant* $C > 0$, *such that*

$$\|\overline{\boldsymbol{\beta}}^d - \widehat{\boldsymbol{\beta}}^d\|_\infty \leq C\frac{(s \vee s_1)k \log d}{n},$$

$$\|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty \leq C\left(\sqrt{\frac{\log d}{n}} + \frac{(s \vee s_1)k \log d}{n}\right).$$

REMARK 4.7. The term $\sqrt{\log d/n}$ in above is the estimation error of $\|\widehat{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty$, while the error term $(s \vee s_1)k \log d/n$ is attributable to the distance between $\overline{\boldsymbol{\beta}}^d$ and $\widehat{\boldsymbol{\beta}}^d$.

Applying a similar thresholding step as in the linear model, we quantify the $\ell_2$-norm estimation error, $\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_2$ and the distance between the divide-and-conquer estimator and full sample estimator $\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \mathcal{T}_\nu(\widehat{\boldsymbol{\beta}}^d)\|_2$.

THEOREM 4.8. *For the GLM* (2.7), *under Conditions* 2.1–3.7, *choose* $\lambda \asymp \sqrt{k \log d/n}$ *and* $\lambda_\nu \asymp \sqrt{k \log d/n}$. *Take the parameter of the hard threshold operator in* (4.2) *as* $\nu = C_0\sqrt{\log d/n}$ *for some sufficiently large constant* $C_0$. *If the*

*number of subsamples satisfies* $k = O(\sqrt{n/((s \vee s_1)^2 \log d)})$, *for large enough* $d$ *and* $n$, *we have with probability* $1 - c/d$

$$\begin{aligned} &\|\mathcal{T}_v(\overline{\boldsymbol{\beta}}^d) - \mathcal{T}_v(\widehat{\boldsymbol{\beta}}^d)\|_2 \le C\frac{(s \vee s_1)s^{1/2}k \log d}{n}, \\ &\|\mathcal{T}_v(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_\infty \le C\sqrt{\frac{\log d}{n}} \quad \textit{and} \\ &\|\mathcal{T}_v(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_2 \le C\sqrt{s \log d/n}. \end{aligned} \tag{4.3}$$

REMARK 4.9. As in the case of the linear model, Theorem 4.8 reveals that the loss incurred by the divide-and-conquer procedure is negligible compared to the statistical minimax estimation error provided $k = o(\sqrt{n/(s_1 \vee s)^2 s \log d})$.

A similar proof strategy to that of Theorem 4.8 allows us to construct an estimator of $\Theta^*_{vv}$ that achieves the required consistency with the scaling of Corollary 3.9. Our estimator is $\widetilde{\Theta}_{vv} := [\mathcal{T}_\zeta(\overline{\Theta})]_{vv}$, where $\overline{\Theta} = k^{-1}\sum_{j=1}^k \widehat{\Theta}^{(j)}$ and $\mathcal{T}_\zeta(\cdot)$ is the thresholding operator defined in equation (4.2) with $\zeta = C_1\sqrt{\log d/n}$ for some sufficiently large constant $C_1$.

COROLLARY 4.10. *Under the conditions and scaling of Theorem* 3.8, $|\widetilde{\Theta}_{vv} - \Theta^*_{vv}| = o_{\mathbb{P}}(1)$.

Substituting this estimator in Corollary 3.9 delivers a practically implementable test statistic based on $k = o(((s \vee s_1)\log d)^{-1}\sqrt{n})$ subsamples.

REMARK 4.11. Notice that point estimation requires less stringent scaling of $k$ than hypothesis testing in both the linear and generalized linear models. This is because the testing and estimation require different rates for the higher order term $\boldsymbol{\Delta}$ in the decomposition

$$\sqrt{n}(\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*) = \boldsymbol{Z} + \boldsymbol{\Delta},$$

where $\boldsymbol{Z}$ is the leading term contributing to the asymptotic normality of $\sqrt{n}(\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*)$. For hypothesis testing, we need $\|\boldsymbol{\Delta}/\sqrt{n}\|_\infty = o_P(1/\sqrt{n})$ to guarantee the asymptotic normality. For estimation, we need $\|\boldsymbol{\Delta}/\sqrt{n}\|_\infty = o_P(\sqrt{\log d/n})$ to match the minimax rate of $\|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty$. Therefore, the number of splits $k$ for testing is more stringent by a factor of $1/\sqrt{\log d}$ than in estimation.

**5. Simulations.** In this section, we illustrate and validate our theoretical findings through simulations. For hypothesis testing, we use QQ plots to compare the distribution of $p$-values for divide-and-conquer test statistics to their theoretical

uniform distribution. We also investigate the estimated type I error and power of the divide-and-conquer tests. For estimation, we validate our claim in the beginning of Section 4 that the loss incurred by the divide-and-conquer strategy is negligible compared with the statistical error of the corresponding full sample estimator in the high dimensional case. Specifically, we compare the performance of the divide-and-conquer thresholding estimator of Section 4.1 with the full sample Lasso and the average Lasso over subsamples. An analogous empirical verification of the theory is performed for the low dimensional case as well; we put it in Appendices C and D of the Supplementary Material [Battey et al. (2018)].

5.1. *Results on hypothesis testing*. We explore the probability of rejection of a null hypothesis of the form $H_0 : \beta_1^* = 0$ when data $(Y_i, \boldsymbol{X}_i)_{i=1}^n$ are generated according to the linear model

$$\boldsymbol{Y}_i = \boldsymbol{X}_i^T \boldsymbol{\beta}^* + \varepsilon_i, \qquad \varepsilon_i \sim N\big(0, \sigma_\varepsilon^2\big),$$

where $\sigma_\varepsilon^2 = 1$ and

$$\boldsymbol{\beta}^* = \big(\beta_1^*, \underbrace{0, \ldots, 0}_{d-s-1}, \underbrace{1, \ldots, 1}_{s}\big)^T,$$

where $d = 5000$ and $s = 3$. In each Monte Carlo replication, we split the initial sample of size $n$ into $k$ subsamples of size $n/k$. In particular, we choose $n = 5000$ and $k \in \{1, 2, 5, 10, 20, 25, 40, 50, 100, 200, 500\}$. The number of Monte Carlo replications is 500. Using $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}$ as a preliminary estimator of $\boldsymbol{\beta}^*$, we construct Wald and Rao's score test statistics as described in Sections 3.1.2 and 3.2, respectively.

Panels (A) and (B) of Figure 1 are QQ plots of the $p$-values of the divide-and-conquer Wald and score test statistics under the null hypothesis against the theoretical quantiles of of the uniform [0, 1] distribution for eight different values of $k$. For both test constructions, the distributions of the $p$-values are close to uniform and remain so as we split the data set. When $k \geq 100$, the distribution of the corresponding $p$-values deviates from the uniform distribution visibly, as expected from the theory developed in Sections 3.1.2 and 3.2. Panel (A) of Figure 2 shows that, for both test constructions, when the number of splits $k \leq 50$, the empirical level of the test is close to both the nominal $\alpha = 0.05$ level and the level of the full sample oracle OLS estimator which knows the true support of $\boldsymbol{\beta}^*$. On the other hand, the type I error increases dramatically when $k$ is larger than 50. This is consistent with asymptotic normality of the test statistics we established when $k$ is controlled appropriately. Panel (B) of Figure 2 displays the power of the test for two different signal strengths, $\beta_1^* = 0.05$ and 0.06. We see that the power for the Score and Wald tests improves when the signal strength goes from 0.05 to 0.06. In addition, we find that the power is high regardless of how large $k$ is. However, Figure 2(A) shows that the Type I error is large when $k$ is large, which makes the

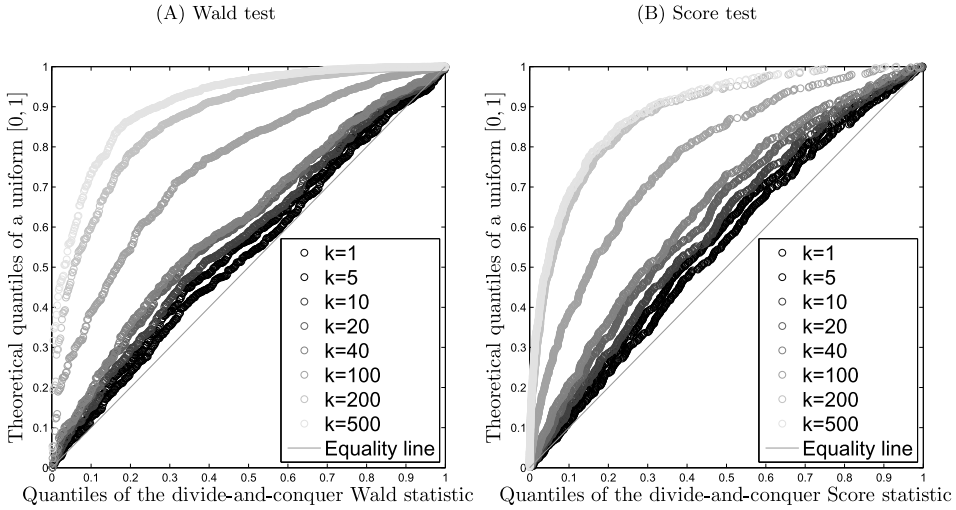


FIG. 1. *QQ plots of the p-values of the Wald* (A) *and score* (B) *divide-and-conquer test statistics against the theoretical quantiles of the uniform* [0, 1] *distribution under the null hypothesis*.

tests invalid. Therefore, these results illustrate that the Type I and II errors are controllable when the number of splits $k$ is relatively small. We also record the wall time for computation for these $k$'s in Table 1. The wall time is computed by taking the maximal time taken for each split and averaged over replications.

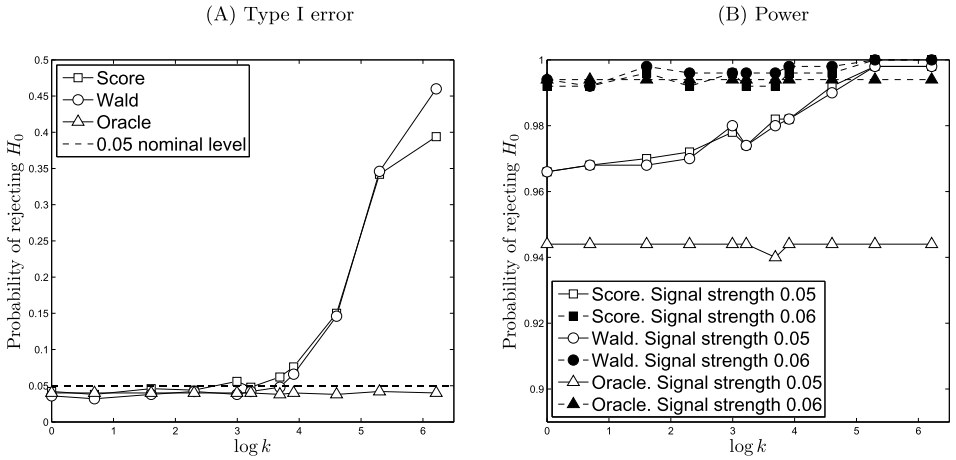


FIG. 2. (A) *Estimated probabilities of type I error for the Wald and score tests as a function of* $k$. (B) *Estimated power with signal strength* 0.05 *and* 0.06 *for the Wald, and score tests as a function of* $k$.

TABLE 1
*Computation time for the divide-and-conquer testing and estimation, where $k = 1$ corresponds to the nonsplitting case and $k > 1$ corresponds to the distributed case*

| $k$ | 1 | 2 | 5 | 10 | 20 | 25 | 50 | 100 | 200 |
|---|---|---|---|---|---|---|---|---|---|
| Score test (s) | 364.39 | 73.22 | 35.09 | 23.61 | 23.56 | 20.78 | 24.13 | 37.53 | 64.67 |
| Wald test (s) | 426.23 | 68.95 | 19.66 | 10.09 | 6.70 | 5.71 | 3.88 | 2.60 | 1.91 |
| $\mathcal{T}_\nu(\overline{\beta}^d)$ ($10^3$ s) | 61.50 | 30.00 | 7.92 | 6.58 | 4.48 | 2.94 | 2.64 | 2.11 | 1.66 |
| Split Lasso (s) | 89.18 | 32.02 | 34.57 | 6.47 | 4.87 | 4.16 | 2.56 | 1.92 | 2.64 |

5.2. *Results on estimation.* In this section, we turn our attention to experimental validation of our divide-and-conquer estimation theory, focusing first on the low dimensional case and then on the high dimensional case.

5.2.1. *The high dimensional linear model.* We now consider the same setting of Section 5.1 with $n = 5000$, $d = 5000$ and $\beta_j^* = 10$ for all $j$ in the support of $\boldsymbol{\beta}^*$. In this context, we analyze the performance of the thresholded averaged debiased estimator of Section 4.1. Figure 3(A) depicts the average over 100 Monte Carlo replications of $\|\boldsymbol{b} - \boldsymbol{\beta}^*\|_2$ for three different estimators: debiased divide and conquer $\boldsymbol{b} = \mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d)$, the Lasso estimator based on the whole sample $\boldsymbol{b} = \widehat{\boldsymbol{\beta}}_{\text{Lasso}}$ and the estimator obtained by naïvely averaging the Lasso estimators from the $k$ subsamples $\boldsymbol{b} = \overline{\boldsymbol{\beta}}_{\text{Lasso}}$. The parameter $\nu$ is taken as $\nu = \sqrt{\log d/n}$ in the specification of $\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d)$. As expected, the performance of $\overline{\boldsymbol{\beta}}_{\text{Lasso}}$ deteriorates sharply as $k$ increases. $\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d)$ outperforms $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}$ as long as $k$ is not too large. This is expected because, for sufficiently large signal strength, both $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}$ and $\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d)$

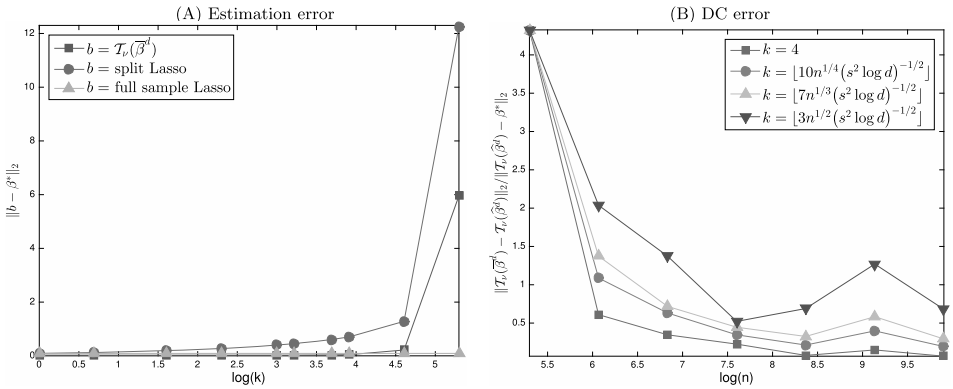


FIG. 3. (A) *Statistical error of the DC estimator, split Lasso and the full sample Lasso for* $k \in \{1, 2, 5, 10, 20, 25, 40, 50, 100, 200\}$ *when* $n = 5000$, $d = 5000$. (B) *Euclidean norm difference between the DC thresholded debiased estimator and its full sample analogue.*

recover the correct support, however, $\mathcal{T}_v(\overline{\boldsymbol{\beta}}^d)$ is unbiased for those $\beta_j^*$ in the support of $\boldsymbol{\beta}^*$, while $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}$ is biased. Figure 3(B) shows the error incurred by the divide-and-conquer procedure $\|\mathcal{T}_v(\overline{\boldsymbol{\beta}}^d) - \mathcal{T}_v(\widehat{\boldsymbol{\beta}}^d)\|_2$ relative to the statistical error of the full sample estimator, $\|\mathcal{T}_v(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_2$, for four different scalings of $k$. We observe that, with $k \asymp O(\sqrt{n/s^2 \log d})$, the relative error incurred by the divide-and-conquer procedure can hardly converge. This is consistent with Theorem 4.3. Given the lower bound of statistical error of the full sample Lasso estimator $\widehat{\boldsymbol{\beta}}$, From Theorem 4.3, we derive that

$$\frac{E\|\mathcal{T}_v(\overline{\boldsymbol{\beta}}^d) - \mathcal{T}_v(\widehat{\boldsymbol{\beta}}^d)\|_2^2}{E\|\mathcal{T}_v(\widehat{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_2^2} \leq \frac{s^2 k^2 \log d}{n}.$$

When $k \asymp O(\sqrt{n/s^2 \log d})$, the right-hand side is an $O(1)$ term. Therefore, the line with inverted triangles in Figure 3(B) implies that the statistical error rate we developed in Theorem 4.3 is tight. We also record the wall time for estimation computation for these $k$'s in Table 1. The wall time is computed by taking the maximal time taken for each splits and averaged over replications. We notice that the computation time decreases with $k$ at first due to the parallel algorithm. However, for the score test and split Lasso, the time becomes increasing when $k$ is large; this is because the computation time to aggregate results from different splits is no longer negligible for very large $k$'s.

**6. Discussion.** With the advent of the data revolution comes the need to modernize the classical statistical tool kit. For very large scale datasets, distribution of data across multiple machines is the only practical way to overcome storage and computational limitations. It is thus essential to build aggregation procedures for conducting inference based on the combined output of multiple machines. We successfully achieve this objective, deriving divide-and-conquer analogues of the Wald and score statistics and providing statistical guarantees on their performance as the number of sample splits grows to infinity with the full sample size. Tractable limit distributions of each DC test statistic are derived. These distributions are valid as long as the number of subsamples, $k$, does not grow too quickly. In particular, $k = o(((s \vee s_1) \log d)^{-1}\sqrt{n})$ is required in a general likelihood-based framework. If $k$ grows faster than $((s \vee s_1) \log d)^{-1}\sqrt{n}$, remainder terms become nonnegligible and contaminate the tractable limit distribution of the leading term. When attention is restricted to the linear model, a faster growth rate of $k = o((s \log d)^{-1}\sqrt{n})$ is allowed.

The divide-and-conquer strategy is also successfully applied to estimation of regression parameters. We obtain the rate of the loss incurred by the divide-and-conquer strategy. Based on this result, we derive an upper bound on the number of subsamples for preserving the statistical error. For low dimensional models,

simple averaging is shown to be effective in preserving the statistical error, so long as $k = O(n/d)$ for the linear model and $k = O(\sqrt{n}/d)$ for the generalized linear model. For high dimensional models, the debiased estimator used in the Wald construction is also successfully employed, achieving the same statistical error as the Lasso-based on the full sample, so long as $k = O(\sqrt{n/s^2 \log d})$.

Our contribution advances the understanding of distributed inference in the presence of large scale and distributed data, but there is still a great deal of work to be done in the area. We focus here on the fundamentals of hypothesis testing and estimation in the divide-and-conquer setting. Beyond this, there is a whole tool kit of statistical methodology designed for the single sample setting, whose split sample asymptotic properties are yet to be understood.

**7. Proofs.** In this section, we present the proofs of the main theorems appearing in Sections 3 and 4. The statements and proofs of several auxiliary lemmas appear in the Supplementary Material [Battey et al. (2018)]. To simplify notation, we take $\beta_v^H = 0$ without loss of generality.

7.1. *Proofs for Section 3.1.* The proof of Theorem 3.3, relies on the following lemma, which bounds the probability that optimization problems in (3.4) are feasible.

LEMMA 7.1. *Assume* $\Sigma = \mathbb{E}(\boldsymbol{X}_i \boldsymbol{X}_i^T)$ *satisfies* $C_{\min} < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_{\max}$ *as well as* $\|\Sigma^{-1/2}\boldsymbol{X}_1\|_{\psi_2} = \kappa$, *then we have*

$$\mathbb{P}\Big(\max_{j=1,\ldots,k} \big\|M^{(j)}\widehat{\Sigma}^{(j)} - I\big\|_{\max} \leq a\sqrt{\frac{\log d}{n}}\Big) \geq 1 - 2kd^{-c_2},$$

*where* $c_2 = \frac{a^2 C_{\min}}{24e^2\kappa^4 C_{\max}} - 2$.

PROOF. The proof is an application of the union bound in Lemma 6.2 of Javanmard and Montanari (2014). □

PROOF OF THEOREM 3.3. For $1 \leq j \leq k$, let $\sqrt{n_k}(\widehat{\boldsymbol{\beta}}^d(\mathcal{D}_j) - \boldsymbol{\beta}^*) = \boldsymbol{Z}^{(j)} + \boldsymbol{\Delta}^{(j)}$, where $\boldsymbol{Z}^{(j)} = \frac{1}{\sqrt{n_k}} M^{(j)} X^{(j)T} \boldsymbol{\varepsilon}^{(j)}$. From Theorem F.1, we know that as long as $\boldsymbol{m}_v^{(j)T}\widehat{\boldsymbol{\Sigma}}^{(j)}\boldsymbol{m}_v^{(j)} \geq c > 0$ holds uniformly for $j = 1, \ldots, d$,

$$\overline{\boldsymbol{\Delta}} := \sqrt{n}\frac{1}{k}\sum_{j=1}^{k} \frac{\Delta_v^{(j)}}{\widehat{Q}^{(j)}} = o_{\mathbb{P}}(1).$$

Then we define

$$\overline{V}_n := \sqrt{n}\frac{1}{k}\sum_{j=1}^{k} \frac{Z_v^{(j)}}{\widehat{Q}^{(j)}} = \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j} \xi_{iv}^{(j)} \qquad \text{where } \xi_{iv}^{(j)} := \frac{\mathbf{m}_v^{(j)T}\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}}{(n\mathbf{m}_v^{(j)T}\widehat{\Sigma}^{(j)}\mathbf{m}_v^{(j)})^{1/2}}.$$

We now establish the asymptotic normality of $\overline{V}_n$ by verifying the requirements of the Lindeberg–Feller central limit theorem [e.g., Kallenberg (1997), Theorem 4.12]. By the fact that $\varepsilon_i$ is independent of $X$ for all $i$ and $\mathbb{E}[\varepsilon_i] = 0$,

$$\begin{aligned}\mathbb{E}(\xi_{iv}) &= \mathbb{E}\big(\mathbb{E}(\xi_{iv}^{(j)}|X)\big) = \mathbb{E}\big\{\mathbb{E}\big[\mathbf{m}_v^{(j)T}\boldsymbol{X}_i^{(j)}\varepsilon_i^{(j)}/\big(n\mathbf{m}_v^{(j)T}\widehat{\Sigma}^{(j)}\mathbf{m}_v^{(j)}\big)^{1/2}|X\big]\big\} \\ &= \mathbb{E}\big\{\big(n\mathbf{m}_v^{(j)T}\widehat{\Sigma}^{(j)}\mathbf{m}_v^{(j)}\big)^{-1/2}\mathbf{m}_v^{(j)T}\boldsymbol{X}_i^{(j)}\mathbb{E}\big(\varepsilon_i^{(j)}\big)\big\} = 0.\end{aligned}$$

By independence of $\{\varepsilon_i\}_{i=1}^n$ and the definition of $\widehat{\Sigma}^{(j)}$, we also have

$$\begin{aligned}\operatorname{Var}(\overline{V}_n|X) &= \sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\operatorname{Var}(\xi_{iv}^{(j)}|X) \\ &= \frac{1}{n}\sum_{j=1}^{k}\big(\mathbf{m}_v^{(j)T}\widehat{\Sigma}^{(j)}\mathbf{m}_v^{(j)}\big)^{-1}\sum_{i\in\mathcal{I}_j}\mathbf{m}_v^{(j)T}\boldsymbol{X}_i^{(j)}\boldsymbol{X}_i^{(j)T}\mathbf{m}_v^{(j)}\operatorname{Var}\big(\varepsilon_i^{(j)}|X\big) \\ &= \sigma^2.\end{aligned}$$

Therefore, we have

$$\operatorname{Var}(\overline{V}_n) = \mathbb{E}\big(\operatorname{Var}(\overline{V}_n|X)\big) + \operatorname{Var}\big(\mathbb{E}(\overline{V}_n|X)\big) = \sigma^2.$$

It only remains to verify the Lindeberg condition, that is,

$$\lim_{k\to\infty}\lim_{n_k\to\infty}\frac{1}{\sigma^2}\sum_{j=1}^{k}\sum_{i\in\mathcal{I}_j}\mathbb{E}\big[\big(\xi_{iv}^{(j)}\big)^2\mathbb{1}\big\{\big|\xi_{iv}^{(j)}\big| > \varepsilon\sigma\big\}\big] = 0 \qquad \forall\varepsilon > 0, \tag{7.1}$$

whose verification is relegated to the Appendix E of the Supplementary Material [Battey et al. (2018)]. Finally, we reach the conclusion by Slutsky's theorem. □

PROOF OF COROLLARY 3.5. Let $\mathcal{F}_n := \{\boldsymbol{m}_v^{(j)T}\widehat{\boldsymbol{\Sigma}}^{(j)}\boldsymbol{m}_v^{(j)} \geq c > 0, j = 1,\dots,k\}$, where $n$ is the total sample size. According to Theorem 3.3, when $\mathcal{F}_n$ holds, we have

$$\lim_{n\to\infty}\mathbb{P}(\overline{S}_n \leq t \mid \mathbf{X}) - \Phi(t) = 0.$$

From the proof of Lemma 13 in Javanmard and Montanari (2014), $\lim_{n\to\infty}\mathbb{P}(\mathcal{F}_n) = 1$. For any $t \in \mathbb{R}$ and $\delta > 0$, by applying dominating convergence Theorem to $\mathbb{1}\{|\mathbb{P}(\overline{S}_n \leq t \mid \mathbf{X}) - \Phi(t)| > \delta$ and $\mathcal{F}_n$ holds$\}$, we have

$$\lim_{n\to\infty}\mathbb{P}\big(\big|\mathbb{P}(\overline{S}_n \leq t \mid \mathbf{X}) - \Phi(t)\big| > \delta\big) = 0.$$

According to the dominate convergence theorem, since $\mathbb{P}(\overline{S}_n \leq t \mid \mathbf{X}) \in [0,1]$, we have

$$\lim_{n\to\infty}\mathbb{P}(\overline{S}_n \leq t) = \lim_{n\to\infty}\mathbb{E}\big[\mathbb{P}(\overline{S}_n \leq t \mid \mathbf{X})\big] = \mathbb{E}\Big[\lim_{n\to\infty}\mathbb{P}(\overline{S}_n \leq t \mid \mathbf{X})\Big] = \Phi(t).$$

Therefore, we complete the proof of the corollary. □

The proofs of Theorem 3.8 and Corollary 3.9 are stated as an application of Lemmas E.7 and E.8 in the Supplementary Material [Battey et al. (2018)], which apply under a more general set of requirements. We present the proof of Theorem 3.8 below and defer Corollary 3.9 to Appendix E in the Supplementary Material.

PROOF OF THEOREM 3.8. We verify (A1)–(A4) of Lemma E.7. For (A1), decompose the object of interest as

$$\frac{1}{n_k}\|X^{(j)}\widehat{\Theta}^{(j)}\|_{\max} = \frac{1}{n_k}\|X^{(j)}(\widehat{\Theta}^{(j)} - \Theta^*)\|_{\max} + \frac{1}{n_k}\|X^{(j)}\Theta^*\|_{\max} = \Delta_1 + \Delta_2,$$

where $\Delta_1$ can be further decomposed and bounded by

$$\begin{aligned}\frac{1}{n_k}\|X^{(j)}(\widehat{\Theta}^{(j)} - \Theta^*)\|_{\max} &= \frac{1}{n_k}\max_{1\le i\le n}\max_{1\le v\le d}[|\boldsymbol{X}_i^{(j)T}(\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*)|] \\ &\le \frac{1}{n_k}\max_{1\le i\le n}\|\boldsymbol{X}_i\|_\infty \max_{1\le v\le d}\|\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*\|_1.\end{aligned}$$

We have

$$\mathbb{P}(\Delta_1 > q/2) \le \mathbb{P}\Big(\max_{1\le v\le d}\|\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*\|_1 > \frac{n}{kM}\frac{q}{2}\Big) < \psi$$

and by Condition 3.7, $\psi = o(d^{-1}) = o(k^{-1})$ for any $q \ge 2CMs_1(k/n)^{3/2}\cdot\sqrt{\log d}$, a fortiori for $q$ a constant. Since $\boldsymbol{X}_i$ is sub-Gaussian, a matching probability bound can easily be obtained for $\Delta_2$, thus we obtain

$$\mathbb{P}(n_k^{-1}\|X^{(j)}\widehat{\Theta}^{(j)}\|_{\max}) \le 2\psi$$

for $\psi = o(k^{-1})$. (A2) and (A3) of Lemma E.7 are applications of Lemmas E.3 and E.4, respectively. To establish (A4), observe that

$$(\widehat{\boldsymbol{\Theta}}_v^{(j)T}\nabla^2\ell_{n_k}^{(j)}(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)) - \boldsymbol{e}_v) = \Delta_1 + \Delta_2 + \Delta_3,$$

where $\Delta_1 = (\widehat{\boldsymbol{\Theta}}_v^{(j)} - \boldsymbol{\Theta}_v^*)^T\nabla^2\ell_{n_k}^{(j)}(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j))$, $\Delta_2 = \boldsymbol{\Theta}_v^{*T}(\nabla^2\ell_{n_k}^{(j)}(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)) - \nabla^2\ell_{n_k}^{(j)}(\boldsymbol{\beta}^*))$ and $\Delta_3 = \boldsymbol{\Theta}_v^{*T}\nabla^2\ell_{n_k}^{(j)}(\boldsymbol{\beta}^*) - \boldsymbol{e}_v$. We thus consider $|\Delta_\ell(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)|$ for $\ell = 1, 2, 3$,

$$\begin{aligned}&|\Delta_2(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)| \\ &\quad= \Big|\frac{1}{n_k}\sum_{i\in\mathcal{I}_j}\boldsymbol{\Theta}_v^{*T}\boldsymbol{X}_i\boldsymbol{X}_i^T(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)[b''(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j)) - b''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)]\Big| \\ &\quad\le U_3\max_{1\le i\le n}|\boldsymbol{\Theta}_v^{*T}\boldsymbol{X}_i|\frac{1}{n_k}\|X(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)\|_2^2.\end{aligned}$$

$\mathbb{P}(\|X(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)\|_2^2 \gtrsim n^{-1}sk\log(d/\delta)) < \delta$ by Lemma E.4, thus $\mathbb{P}(|\Delta_2\cdot(\widehat{\boldsymbol{\beta}}^\lambda(\mathcal{D}_j) - \boldsymbol{\beta}^*)| > t) < \delta$ for $t \asymp MU_3n^{-1}sk\log(d/\delta)$. Invoking Hölder's

inequality, Hoeffding's inequality and Condition 2.1, we also obtain, for $t \asymp n^{-1}sk\log(d/\delta)$,

$$\mathbb{P}\big(|\Delta_3(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*)| > t\big) \le \mathbb{P}\bigg(\bigg\|\boldsymbol{\Theta}_v^{*T}\bigg(\frac{1}{n_k}\sum_{i\in\mathcal{I}_j} b''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\boldsymbol{X}_i\boldsymbol{X}_i^T\bigg) - \boldsymbol{e}_v\bigg\|_{\max} \|\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*\|_1 > t\bigg).$$

Therefore, $\mathbb{P}(|\Delta_2(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*)| > t) < 2\delta$. Finally, with $t \asymp n^{-1}(s \vee s_1)k \times \log(d/\delta)$,

$$\mathbb{P}\big(|\Delta_1(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*)| > t\big) \le \mathbb{P}\bigg(\frac{1}{n_k}\bigg\|\sum_{i\in\mathcal{I}_j} X_i^T(\widehat{\boldsymbol{\Theta}}_v - \boldsymbol{\Theta}_v)b''(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j))\bigg\|_2 \|X^{(j)^T}(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*)\|_2 > t\bigg),$$

hence $\mathbb{P}(|\Delta_1(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*)| > t) < 2\delta$. This follows because, by Lemma E.4,

$$\mathbb{P}\bigg(\bigg\|\frac{1}{n_k}X^{(j)}(\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j) - \boldsymbol{\beta}^*)\bigg\|_2 \gtrsim n^{-1/2}\sqrt{sk\log(d/\delta)}\bigg) < \delta$$

and by Lemma C.4 of Ning and Liu (2017),

$$\mathbb{P}\bigg(\bigg\|\frac{1}{n_k}\sum_{i\in\mathcal{I}_j} X_i^T(\widehat{\boldsymbol{\Theta}}_v - \boldsymbol{\Theta}_v)b''(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_j))\bigg\|_2 \gtrsim n^{-1/2}\sqrt{s_1k\log(d/\delta)}\bigg) < \delta. \qquad \square$$

7.2. *Proofs for theorems in Section 3.2.* The proof of Theorem 3.11 relies on several preliminary lemmas, collected in Appendix E in the Supplementary Material [Battey et al. (2018)]. Without loss of generality, we set $H_0 : \beta_v^* = 0$ to ease notation.

PROOF OF THEOREM 3.11. Since $\overline{S}(0) = k^{-1}\sum_{j=1}^k \widehat{S}^{(j)}(0, \widehat{\boldsymbol{\beta}}^{\lambda}_{-v}(\mathcal{D}_j))$, and (B1)–(B4) of Condition E.9 in the Supplementary Material [Battey et al. (2018)] are fulfilled under Conditions 3.6 and 2.1 by Lemma E.10 (see Appendix E in the Supplementary Material). The proof is now simply an application of Lemma E.13 in the Supplementary Material with $\beta_v^* = 0$ under the restriction of the null hypothesis. $\square$

PROOF OF LEMMA 3.14. The proof is an application of Lemma E.16 in the Supplementary Material [Battey et al. (2018)], noting that (B1)–(B5) of Condition E.9 in the Supplementary Material are fulfilled under Conditions 3.6 and 2.1 by Lemmas E.10 and E.11 in the Supplementary Material. $\square$

7.3. *Proofs for theorems in Section 4.* Recall from Section 2 that for an arbitrary matrix $M$, $\boldsymbol{M}_\ell$ denotes the transposed $\ell$th row of $M$ and $[\boldsymbol{M}]_\ell$ denotes the $\ell$th column of $M$.

PROOF OF THEOREM 4.3. By Lemma 4.1 and $k = O(\sqrt{n/(s^2 \log d)})$, there exists a sufficiently large $C_0$ such that for the event $\mathcal{E} := \{\|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty \leq C_0\sqrt{\log d/n}\}$, we have $\mathbb{P}(\mathcal{E}) \geq 1 - c/d$. We choose $\nu = C_0\sqrt{\log d/n}$, which implies that, under $\mathcal{E}$, we have $\nu \geq \|\overline{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty$.

Let $\mathcal{S}$ be the support of $\boldsymbol{\beta}^*$. The derivations in the remainder of the proof hold on the event $\mathcal{E}$. Observe $\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d_{\mathcal{S}^c}) = \mathbf{0}$ as $\|\overline{\boldsymbol{\beta}}^d_{\mathcal{S}^c}\|_\infty \leq \nu$. For $j \in \mathcal{S}$, if $|\boldsymbol{\beta}^*_j| \geq 2\nu$, we have $|\overline{\boldsymbol{\beta}}^d_j| \geq |\beta^*_j| - \nu \geq \nu$, and thus $|\mathcal{T}_\nu(\overline{\beta}^d_j) - \beta^*_j| = |\overline{\beta}^d_j - \beta^*_j| \leq \nu$. While if $|\beta^*_j| < 2\nu$, $|\mathcal{T}_\nu(\overline{\beta}^d_j) - \beta^*_j| \leq |\beta^*_j| \vee |\overline{\beta}^d_j - \beta^*_j| \leq 2\nu$. Therefore, on the event $\mathcal{E}$,

$$\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_2 = \|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d_{\mathcal{S}}) - \boldsymbol{\beta}^*_{\mathcal{S}}\|_2 \leq 2\sqrt{s}\nu \quad \text{and}$$
$$\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \boldsymbol{\beta}^*\|_\infty = \|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d_{\mathcal{S}}) - \boldsymbol{\beta}^*_{\mathcal{S}}\|_\infty \leq 2\nu.$$

The statement of the theorem follows because $\nu = C_0\sqrt{\log d/n}$ and $\mathbb{P}(\mathcal{E}) \geq 1 - c/d$. Following the same reasoning, on the event $\mathcal{E}' := \mathcal{E} \cup \{\|\widehat{\boldsymbol{\beta}}^d - \boldsymbol{\beta}^*\|_\infty \leq C_0\sqrt{\log d/n}\} \cup \{\|\widehat{\boldsymbol{\beta}}^d - \overline{\boldsymbol{\beta}}^d\|_\infty \leq C_0 sk \log d/n\}$, we have

$$\begin{aligned}\|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d) - \mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d)\|_2 &= \|\mathcal{T}_\nu(\overline{\boldsymbol{\beta}}^d_{\mathcal{S}}) - \mathcal{T}_\nu(\widehat{\boldsymbol{\beta}}^d_{\mathcal{S}})\|_2 \\ &\leq \|\overline{\boldsymbol{\beta}}^d_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}^d_{\mathcal{S}}\|_2 \leq \sqrt{s}\|\overline{\boldsymbol{\beta}}^d_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}^d_{\mathcal{S}}\|_\infty \leq Cs^{3/2}k \log d/n.\end{aligned}$$

As Lemma 4.1 also gives $\mathbb{P}(\mathcal{E}') \geq 1 - c/d$, the proof is complete. □

PROOF OF COROLLARY 4.10. By an analogous proof strategy to that of Theorem 4.8, $|[\mathcal{T}_\zeta(\overline{\Theta})]_{vv} - \Theta^*_{vv}| = O_p(\sqrt{n^{-1}\log d}) = o_{\mathbb{P}}(1)$ under the conditions of the Corollary provided $k = o(((s \vee s_1)\log d)^{-1}\sqrt{n})$. □

**Acknowledgements.** The authors thank Weichen Wang, Jason Lee and Yuekai Sun for helpful comments.

## SUPPLEMENTARY MATERIAL

**Supplement to "Distributed testing and estimation under sparse high dimensional models"** (DOI: 10.1214/17-AOS1587SUPP; .pdf). We put all technical lemmas, proofs and low dimensional results in the supplementary materials for reference.

## REFERENCES


BATTEY, H., FAN, J., LIU, H., LU, J. and ZHU, Z. (2018). Supplement to "Distributed testing and estimation under sparse high dimensional models." DOI:10.1214/17-AOS1587SUPP.

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434. MR0386168

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when $p$ is much larger than $n$. *Ann. Statist.* **35** 2313–2351. MR2382644

CHEN, X. and XIE, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statist. Sinica* **24** 1655–1684. MR3308656

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London. MR0370837

FAN, J., GUO, S. and HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 37–65. MR2885839

FAN, J., HAN, F. and LIU, H. (2014). Challenges of big data analysis. *Nat. Sci. Rev.* **1** 293–314.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484. MR2849368

FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

KALLENBERG, O. (1997). *Foundations of Modern Probability*. Springer, New York. MR1464694

KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 795–816. MR3248677

LEE, J. D., LIU, Q., SUN, Y. and TAYLOR, J. E. (2017). Communication-efficient sparse regression. *J. Mach. Learn. Res.* **18** Paper No. 5. MR3625709

LIU, Q. and IHLER, A. T. (2014). Distributed estimation, information loss and exponential families. In *Advances in Neural Information Processing Systems* 27 (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger, eds.) 1098–1106. MIT Press, Cambridge, MA.

LOH, P.-L. and WAINWRIGHT, M. J. (2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems* 26 (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 476–484. Curran Associates Inc., Red Hook, NY.

LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized $M$-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616. MR3335800

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

NEGAHBAN, S. N., YU, B., WAINWRIGHT, M. J. and RAVIKUMAR, P. (2009). A unified framework for high-dimensional analysis of $M$-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems* 22 (Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta, eds.) 1348–1356. Curran Associates Inc., Red Hook, NY.

NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45** 158–195. MR3611489

ROSENBLATT, J. D. and NADLER, B. (2016). On the optimality of averaging in distributed statistical learning. *Inf. Inference* **5** 379–404. MR3609865

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

WANG, Z., LIU, H. and ZHANG, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Statist.* **42** 2164–2201. MR3269977

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHANG, Y., DUCHI, J. and WAINWRIGHT, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.* **16** 3299–3340. MR3450540

ZHANG, C.-H. and ZHANG, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593. MR3025135

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

ZHAO, T., CHENG, G. and LIU, H. (2016). A partially linear framework for massive heterogeneous data. *Ann. Statist.* **44** 1400–1437. MR3519928

H. BATTEY
DEPARTMENT OF MATHEMATICS
IMPERIAL COLLEGE LONDON
LONDON, SW7 2AZ
UNITED KINGDOM
AND
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: h.battey@imperial.ac.uk
hbattey@princeton.edu

J. FAN
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
AND
SCHOOL OF DATA SCIENCE
FUDAN UNIVERSITY
SHANGHAI
CHINA
E-MAIL: jqfan@princeton.edu

B. LIU
J. LU
Z. ZHU
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: hanliu@princeton.edu
junweil@princeton.edu
ziweiz@princeton.edu