

# Distributed simultaneous inference in generalized linear models via confidence distribution

Lu Tang<sup>a</sup>, Ling Zhou<sup>b</sup>, Peter X.-K. Song<sup>c,\*</sup>

<sup>a</sup> Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA

<sup>b</sup> Center of Statistical Research, Southwestern University of Finance and Economics, Chengdu, Sichuan, China

<sup>c</sup> Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

## ARTICLE INFO

### Article history:

Received 31 January 2019

Received in revised form 9 November 2019

Accepted 9 November 2019

Available online 28 November 2019

### AMS 2010 subject classifications:

primary 62H15

secondary 62F12

### Keywords:

Bias correction

Confidence distribution

Inference

Lasso

Meta-analysis

Parallel computing

## ABSTRACT

We propose a distributed method for simultaneous inference for datasets with sample size much larger than the number of covariates, i.e.,  $N \gg p$ , in the generalized linear models framework. When such datasets are too big to be analyzed entirely by a single centralized computer, or when datasets are already stored in distributed database systems, the strategy of divide-and-combine has been the method of choice for scalability. Due to partition, the sub-dataset sample sizes may be uneven and some possibly close to  $p$ , which calls for regularization techniques to improve numerical stability. However, there is a lack of clear theoretical justification and practical guidelines to combine results obtained from separate regularized estimators, especially when the final objective is simultaneous inference for a group of regression parameters. In this paper, we develop a strategy to combine bias-corrected lasso-type estimates by using confidence distributions. We show that the resulting combined estimator achieves the same estimation efficiency as that of the maximum likelihood estimator using the centralized data. As demonstrated by simulated and real data examples, our divide-and-combine method yields nearly identical inference as the centralized benchmark.

© 2019 Elsevier Inc. All rights reserved.

## 1. Introduction

We consider simultaneous inference for the generalized linear model (GLM) under the situation where data are stored on distributed computer clusters instead of a centralized location. The use of distributed storage can be due to either large data volume or protection of individual-level sensitive data from leaving data-owning entities. Such distributed data presents great challenges in statistical analyses because the entire dataset cannot be loaded once to a single processor for computation [9]. In the advent of cloud storage and computing, the method of divide-and-combine, also known as divide-and-conquer [1], has become the state-of-the-art in big data analytics to effectively improve scalability. Divide-and-combine is a computational procedure that divides the data into relatively independent, smaller and computable batches, processes them in parallel and combines the separate results. However, not all existing statistical methods are directly parallelizable. Some complicated methods require special treatment in order to be adapted to the parallel computing architecture; see for examples, parallel matrix factorization by randomized matrix approximation in [22], scalable bootstrap by bag of little bootstraps in [16], divide-and-combine-type kernel ridge regression in [39],

\* Corresponding author.

E-mail address: [pxsong@umich.edu](mailto:pxsong@umich.edu) (P.X.-K. Song).

and communication efficient lasso regression in [17], among others. In this paper, we consider simultaneous inference for the GLM using divide-and-combine. While both the sample size  $N$  and the number of covariates  $p$  may be large in practice, here we focus on the case when  $N \gg p$ , but  $p$  is not small and can vary from hundreds to thousands. The  $N$  observations are split into  $K$  mutually independent sub-datasets.

Meta-analysis is an example of divide-and-combine that combines summary statistics from independent studies, see for examples [14,28,29]. The classical fixed-effect meta-analysis uses inverse variance weighted average to combine separate point estimates. Raw data can be processed locally and only summary quantities are communicated between machines to reduce cost of data transfer [17]. In the development of distributed algorithms for statistical inference, one question arises naturally: are the proposed divide-and-combine estimators and the maximum likelihood estimator (MLE) obtained from the centralized data asymptotically equivalent, leading to comparable statistical inferences? Lin and Zeng [20] showed that such a meta-estimator asymptotically achieves the Fisher's efficiency; in other words, it follows asymptotically the same distribution as the centralized MLE. The Fisher's efficiency has also been established for a combined estimator by [19] through aggregating estimating equations under a relatively strong condition that  $K$  is of order  $O(n^r)$  where  $r < 1/3$  and  $n$  is the sample size of a sub-dataset. Recently, Battey et al. [2] proposed test statistics and point estimators in the context of the divide-and-combine, where the method of hypothesis testing is only developed for low dimensional parameters, and the combined estimator takes a simple form of an arithmetic average over sub-datasets. Different from [2], we consider simultaneous inference for all parameters and use the inverse of variance-covariance matrices to combine estimates.

Although the overall sample size is large, it is reduced  $K$  times in the sub-datasets due to data partition. The sample size reduction and potentially unbalanced sample sizes across sub-datasets may cause numerical instability in the search for the MLE, especially in overfitted models when most of covariates are unimportant among all covariates that are included in the analysis. As shown in Section 4, coverage probabilities of confidence intervals obtained by the classical meta-analysis method deviate drastically from the nominal level as  $K$  increases. This motivates the use of regularized regression to overcome such numerical instability. For regularized estimators, such as lasso [31] and SCAD [10], constructing confidence intervals is analytically challenging because: (i) sparse estimators usually do not have a tractable limiting distribution, and (ii) the oracle property [10] relying on knowledge of the truly non-zero parameters is not applicable to statistical inference since the oracle is unknown in practice.

When penalized regression is applied on each sub-dataset, variable selection procedures will choose different sets of important covariates by different tuning schemes. Such misaligned selection prohibits any weighting approaches from combining the separate results; both dimensionality and meaning of the estimates across sub-datasets may be very different. Chen and Xie [4] proposed a majority-voting method to combine the estimates of the covariates most frequently identified by the lasso across the sub-datasets. Unfortunately, this method does not provide inference for the combined estimator, and it is sensitive to the choice of inclusion criterion. To fill in this gap, we propose a new approach along the lines of the post-selection inference developed for the penalized estimator by [13] and [40], which allows us to combine bias-corrected lasso estimators obtained from sub-datasets.

In this paper, we use the confidence distribution approach [36] to combine results from the separate analyses of sub-datasets. The confidence distribution, originally proposed by Fisher [11] and later formally formulated by Efron [8], has recently attracted renewed attention in the statistical literature; see for examples, [26,36] and references therein. An advantage of the confidence distribution approach is that it provides a unified framework for combining distributions of estimators, so statistical inference with the combined estimator can be established in a straightforward and mathematically rigorous fashion. Specifically related to divide-and-combine, Xie et al. [37] developed a robust meta-analysis-type approach through confidence distribution, and Liu et al. [21] proposed to combine the confidence distribution functions in the same way as combining likelihood functions for inference, and showed their estimator achieves the Fisher's efficiency. The step of combining via confidence distribution theory requires well-defined asymptotic joint distributions of all model parameters of interest, which, in the current literature, are only available for  $p$  less than  $n$ , the sample size of one sub-dataset under equal data split. Here, we consider the scenarios where  $p$  and  $K$  can both diverge to infinity with rates slower than  $N$ . Our new contribution is two-fold: (i) the combined estimator achieves asymptotically the Fisher's efficiency; that is, it is asymptotically as efficient as the MLE obtained from the direct analysis on the full data; and (ii) the distributed procedure is scalable and parallelizable to address very large sample sizes through easy and fast parallel algorithmic implementation. The latter presents a desirable numerical recipe to handle the case when the centralized data analysis is time consuming and CPU demanding, or even numerically prohibitive.

This paper is organized as follows. Section 2 focuses on the asymptotics of the bias-corrected lasso estimator in sub-datasets. Section 3 presents the confidence distribution method to combine results from multiple regularized regressions. Section 4 provides extensive simulation results, and Section 5 illustrates our method by a real data. We conclude in Section 6. We provide key technical details in the Appendix A and defer complete proofs and supporting information to the Supplementary Material.

## 2. Distributed penalized regressions for sub-datasets

For GLM, the systematic component is specified by the mean of a response  $y_i$  that is related to a  $p$ -dimensional vector of covariates  $\mathbf{x}_i$  by a known monotonic canonical link function  $g(\cdot)$  in the form  $\mu_i = E(y_i) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$ , for subject

$i \in \{1, \dots, N\}$ . The random component is specified by the conditional density of  $\mathbf{Y} = (y_1, \dots, y_N)^T$  given  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^T$ . The variance of the response takes the form of  $\text{var}(y_i) = \phi v(\mu_i)$  where  $\phi$  is the dispersion parameter and  $v(\cdot)$  is the unit variance function [23]. The associated likelihood function is given by  $\mathcal{L}_N(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}) = \prod_{i=1}^N \exp[\{y_i \theta_i - b(\theta_i)\} / \phi + c(y_i, \phi)]$ , where  $\dot{b}(\cdot) = g^{-1}(\cdot)$  and the canonical parameters have the form  $\theta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ , with  $\boldsymbol{\beta}$  being the  $p$ -element vector of regression parameters of interest.

The centralized MLE solution,  $\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} \mathcal{L}_N(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X})$ , in general has no closed-form expression, except for the Gaussian linear model, and is often obtained numerically by certain iterative algorithms such as Newton–Raphson. Thus, it is not trivial to establish exact parallel algorithms that only require a single passing of each sub-dataset, and still achieve the same efficiency as the centralized MLE. Sample partition naturally results in  $K$  sub-datasets, each with size  $n_k$ , and  $\sum_{k=1}^K n_k = N$ .

This section focuses on deriving the regularized estimator and confidence distribution for a *single sub-dataset* of sample size  $n_k$  for a specific  $k$ . Since the method in this section is general to all sub-datasets, for ease of exposition, we suppress  $k$  unless otherwise noted. We start by deriving the asymptotic properties of lasso regularized regression, as our divide-and-combine procedure is dependent on the asymptotic results. The regularization plays an important role in stabilizing numerical performance on the divided datasets, which will be shown in later sections. We use lasso [31] in the development of this paper. With little effort, other types of regularization, such as SCAD [10] or elastic net [42], may be adopted in our proposed procedure.

## 2.1. Lasso in generalized linear models

The lasso estimator is obtained by maximizing the following penalized log-likelihood function with respect to the regression parameters  $\boldsymbol{\beta}$  subject to a normalizing constant,

$$\begin{aligned} PL(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}) &\stackrel{\text{def}}{=} \frac{1}{n} \mathcal{L}_n(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}) - \lambda \|\boldsymbol{\beta}\|_1 \\ &\propto \frac{1}{n\phi} \sum_{i=1}^n \{y_i \mathbf{x}_i^T \boldsymbol{\beta} - b(\mathbf{x}_i^T \boldsymbol{\beta})\} - \lambda \|\boldsymbol{\beta}\|_1, \end{aligned}$$

where  $\lambda$  is a nonnegative tuning parameter, and  $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$  is the  $\ell_1$ -norm of the regression coefficient vector  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ . Let  $\hat{\boldsymbol{\beta}}_\lambda = \arg \max_{\boldsymbol{\beta}} PL(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X})$  be a lasso estimator of  $\boldsymbol{\beta}$  at a given tuning parameter  $\lambda \geq 0$ . Solution  $\hat{\boldsymbol{\beta}}_\lambda$  may be obtained by coordinate descent via Donoho and Johnstone [7]’s soft-thresholding approach, with the tuning parameter being determined by, say, cross-validation [25].

## 2.2. Confidence distribution for bias-corrected lasso estimator

To combine multiple lasso estimators obtained from separate sub-datasets, we need to overcome the issue of misalignment: the sets of selected covariates with non-zero estimates in the model are different across sub-datasets. Our solution is based on bias-corrected lasso estimators. The bias correction enables us not only to obtain non-zero estimates of all regression coefficients, but also, more importantly, to establish the joint distribution of regularized estimators. The latter is critical for us to utilize the confidence distribution to combine estimators, which will be described in Section 3.

Denote the score function by  $\mathbf{S}_n(\boldsymbol{\beta}) = \frac{1}{n\phi} \sum_{i=1}^n \{y_i - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})\} \mathbf{x}_i$ . It is known that the lasso estimator,  $\hat{\boldsymbol{\beta}}_\lambda$ , satisfies the following Karush–Kuhn–Tucker (KKT) condition:  $\mathbf{S}_n(\hat{\boldsymbol{\beta}}_\lambda) - \lambda \hat{\boldsymbol{\kappa}} = \mathbf{0}$ , where subdifferentials  $\hat{\boldsymbol{\kappa}} = (\hat{\kappa}_1, \dots, \hat{\kappa}_p)^T$  satisfy  $\max_j |\hat{\kappa}_j| \leq 1$ , and  $\hat{\kappa}_j = \text{sign}(\hat{\beta}_{\lambda,j})$  if  $\hat{\beta}_{\lambda,j} \neq 0$ . The first-order Taylor expansion of  $\mathbf{S}_n(\hat{\boldsymbol{\beta}}_\lambda)$  in the KKT condition at the true value  $\boldsymbol{\beta}_0$  leads to  $-\dot{\mathbf{S}}_n(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0) + \lambda \hat{\boldsymbol{\kappa}} \approx \mathbf{S}_n(\boldsymbol{\beta}_0)$ . It follows that  $\boldsymbol{\beta}_\lambda^c - \boldsymbol{\beta}_0 \approx \{-\dot{\mathbf{S}}_n(\boldsymbol{\beta}_0)\}^{-1} \mathbf{S}_n(\boldsymbol{\beta}_0)$ , where  $\boldsymbol{\beta}_\lambda^c$  is a bias-corrected lasso estimator [13]:

$$\boldsymbol{\beta}_\lambda^c \stackrel{\text{def}}{=} \hat{\boldsymbol{\beta}}_\lambda + \{-\dot{\mathbf{S}}_n(\boldsymbol{\beta}_0)\}^{-1} \lambda \hat{\boldsymbol{\kappa}} = \hat{\boldsymbol{\beta}}_\lambda + \{-\dot{\mathbf{S}}_n(\boldsymbol{\beta}_0)\}^{-1} \mathbf{S}_n(\hat{\boldsymbol{\beta}}_\lambda). \quad (1)$$

The second equality in (1) follows directly from the KKT condition and the definition of the sensitivity matrix  $-\dot{\mathbf{S}}_n(\boldsymbol{\beta}) = \frac{1}{n\phi} \sum_{i=1}^n v(\mu_i) \mathbf{x}_i \mathbf{x}_i^T$ , which is assumed to be a positive-definite Hessian matrix, and  $v(\cdot)$  is the variance function. For now, let us first consider the case when  $p < n$ . We show in Theorem 1 that under some regularity conditions,  $\boldsymbol{\beta}_\lambda^c$  is asymptotically normally distributed, namely,

$$n^{1/2}(\boldsymbol{\beta}_\lambda^c - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)), \text{ as } n \rightarrow \infty, \quad (2)$$

where  $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0) = [E\{-\dot{\mathbf{S}}_n(\boldsymbol{\beta}_0)\}]^{-1}$ . Based on the joint asymptotic normality in (2), following [36], we form the asymptotic confidence distribution density function of  $\boldsymbol{\beta}_0$  as  $h_n(\boldsymbol{\beta}_0) \propto \exp[-\frac{n}{2}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_\lambda^c)^T \{\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\}^{-1}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_\lambda^c)]$ . Replacing  $\boldsymbol{\beta}_0$  in (1) by the sparse lasso estimator  $\hat{\boldsymbol{\beta}}_\lambda$ , we obtain

$$\hat{\boldsymbol{\beta}}_\lambda^c = \hat{\boldsymbol{\beta}}_\lambda + \{-\dot{\mathbf{S}}_n(\hat{\boldsymbol{\beta}}_\lambda)\}^{-1} \mathbf{S}_n(\hat{\boldsymbol{\beta}}_\lambda). \quad (3)$$

Likewise, replacing  $\beta_0$  by  $\hat{\beta}_\lambda$  in the asymptotic covariance in (2) leads to a “data-driven” asymptotic confidence density

$$\hat{h}_n(\beta_0) \propto \exp \left[ -\frac{n}{2} (\beta_0 - \hat{\beta}_\lambda^c)^T \{ -\dot{S}_n(\hat{\beta}_\lambda) \} (\beta_0 - \hat{\beta}_\lambda^c) \right]. \quad (4)$$

It is worth pointing out that this bias-corrected estimator in (3) is equivalent to a one-step Newton–Raphson updated estimator of the lasso estimator. In the GLM framework, we have  $\hat{h}_n(\beta_0) \propto \exp[-\frac{1}{2\phi}(\beta_0 - \hat{\beta}_\lambda^c)^T \{ \mathbf{X}^T \mathbf{P}_n(\hat{\beta}_\lambda) \mathbf{X} \} (\beta_0 - \hat{\beta}_\lambda^c)]$ , where  $\mathbf{P}_n(\beta) = \text{diag}\{v(\mu_1), \dots, v(\mu_n)\}$  is the diagonal matrix of the variance functions. When the dispersion parameter  $\phi$  is unknown, e.g., in the linear regression setting, we use a root- $n$  consistent estimator  $\hat{\phi} = (n - \|\hat{\beta}_\lambda\|_0)^{-1} \sum_{i=1}^n d(y_i, \hat{\mu}_i)$ , where  $\|\mathbf{x}\|_0$  is the number of non-zero entries of vector  $\mathbf{x}$ ,  $\hat{\mu}_i = g^{-1}(\mathbf{x}_i^T \hat{\beta}_\lambda)$ , and  $d(\cdot, \cdot)$  is the unit deviance function; refer to [27, Chapter 2] for details.

### 2.3. Examples

**Example 1.** Gaussian linear model. Assume  $y_i$  follows a normal distribution with mean  $\mu_i = \mathbf{x}_i^T \beta$ , variance function  $v(\mu_i) = 1$ , and link function  $g(x) = x$ . The score function takes the form  $\mathbf{S}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \{y_i - \mathbf{x}_i^T \beta\} \mathbf{x}_i / \phi$ . The confidence density function  $\hat{h}_n(\beta_0)$  in (4) is obtained by plugging in the bias-corrected estimator  $\hat{\beta}_\lambda^c = \hat{\beta}_\lambda + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{X} \hat{\beta}_\lambda)$ . Here  $\mathbf{P}_n(\hat{\beta}_\lambda) = \mathbf{I}_n$ .

**Example 2.** Binomial logistic model. Assume  $y_i$  follows a Bernoulli distribution with probability of success  $\mu_i \in (0, 1)$ , variance function  $\hat{v}_i = \hat{\mu}_i(1 - \hat{\mu}_i)$ , link function  $g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i}) = \mathbf{x}_i^T \beta$  and  $\phi = 1$ . Similarly, we obtain its confidence density  $\hat{h}_n(\beta_0)$  with  $\hat{\beta}_\lambda^c = \hat{\beta}_\lambda + \{\mathbf{X}^T \mathbf{P}_n(\hat{\beta}_\lambda) \mathbf{X}\}^{-1} \mathbf{X}^T (\mathbf{Y} - \hat{\boldsymbol{\mu}})$ , where  $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_n)^T$ ,  $\hat{\mu}_i = \exp(\mathbf{x}_i^T \hat{\beta}_\lambda) / (1 + \exp(\mathbf{x}_i^T \hat{\beta}_\lambda))$  and  $\mathbf{P}_n(\hat{\beta}_\lambda) = \text{diag}(\hat{v}_1, \dots, \hat{v}_n)$ .

**Example 3.** Poisson log-linear model. Assume  $y_i$  follows a Poisson distribution with mean  $\mu_i$ , variance function  $v(\mu_i) = \mu_i$ , link function  $g(\mu_i) = \log(\mu_i) = \mathbf{x}_i^T \beta$  and  $\phi = 1$ . We can obtain  $\hat{h}_n(\beta_0)$  with  $\hat{\beta}_\lambda^c = \hat{\beta}_\lambda + \{\mathbf{X}^T \mathbf{P}_n(\hat{\beta}_\lambda) \mathbf{X}\}^{-1} \mathbf{X}^T (\mathbf{Y} - \hat{\boldsymbol{\mu}})$ , where  $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_n)^T$ ,  $\hat{\mu}_i = \hat{v}_i = \exp(\mathbf{x}_i^T \hat{\beta}_\lambda)$  and  $\mathbf{P}_n(\hat{\beta}_\lambda) = \text{diag}(\hat{v}_1, \dots, \hat{v}_n)$ .

### 2.4. Large sample property

From here on, we bring back the subscript  $k$  to denote a quantity concerning the  $k$ th sub-dataset as the results will be carried forward to Section 3 where we discuss the combination step. Let  $\underline{\sigma}(M)$  and  $\overline{\sigma}(M)$  denote the minimum and maximum singular values of a matrix  $M$ , respectively. Let  $c_{\min}$  and  $c_{\max}$  be the minimum and maximum across the set of constants  $c_k, k \in \{1, \dots, K\}$ . Denote the signal set by  $\mathcal{A}_{0,k} = \{j : \beta_{0,k,j} \neq 0\}$  and the non-signal set by  $\mathcal{A}_{0,k}^c = \{j : \beta_{0,k,j} = 0\}$ , where  $\beta_{0,k} = (\beta_{0,k,1}, \dots, \beta_{0,k,p})^T$  is the true coefficient. Here we allow  $p \geq n_k$  and  $p$  may diverge to infinity. To establish large-sample properties for  $\hat{\beta}_{\lambda,k}^c$  given in (3) based on the  $k$ th sub-dataset  $(\mathbf{Y}_k, \mathbf{X}_k)$ , and subsequently the combined estimator in Section 3 across all sub-datasets, we postulate the following regularity conditions:

(C1) Assume the score function is unbiased, namely,  $E[\{\mathbf{Y}_k - g^{-1}(\mathbf{X}_k^T \beta_{0,k})\} \mathbf{X}_k / \phi_k] = 0$ .

(C2) Assume  $0 < b_k \leq \underline{\sigma}(n_k^{-1/2} \mathbf{X}_k) \leq \overline{\sigma}(n_k^{-1/2} \mathbf{X}_k) \leq B_k$  for constants  $b_k$  and  $B_k$ , and  $\|\mathbf{X}_k\|_\infty \leq D_k$  for some  $D_k > 0$ , where  $\|\mathbf{X}\|_\infty = \max_{i,j} |x_{i,j}|$ .

(C3) For some  $\psi_{0,k} > 0$ , for all  $\beta$  satisfying  $\|\beta_{\mathcal{A}_{0,k}^c}\|_1 \leq 3\|\beta_{\mathcal{A}_{0,k}}\|_1$ , it holds that  $\|\beta_{\mathcal{A}_{0,k}}\|_1^2 \leq \|\beta\|_2^2 s_{0,k} / \psi_{0,k}^2$ , where  $s_{0,k}$  is the number of true signals in  $\beta_{0,k}$  and  $\|\beta\|_2^2 = \beta^T \beta$ . In addition, assume  $\lambda_k = O(\sqrt{\log p / n_k})$  and  $s_{0,k} = o\{(n_k/p)^{1/2} / \log p\}$ .

(C4) Assume the same underlying true parameters  $\beta_0 = \beta_{0,k}, k \in \{1, \dots, K\}$ . Denote the common signal set, non-signal set, and number of signals as  $\mathcal{A}_0 = \mathcal{A}_{0,k}, \mathcal{A}_0^c = \mathcal{A}_{0,k}^c$  and  $s_0 = s_{0,k}$ , respectively, for all  $k$ . Further, assume  $0 < b_{\min} < B_{\max} < \infty$  and  $\psi_{0,\min} > 0$ .

Conditions (C1) and (C2) are two mild regularity conditions widely used in the literature; see for example [21]. It follows from condition (C2) that,  $C \geq \max_{\mu \in \Omega_\delta} v(\mu) \geq \min_{\mu \in \Omega_\delta} v(\mu) \geq c > 0$  with  $\Omega_\delta = \{g^{-1}(\mathbf{x}^T \beta) : \|\mathbf{x}^T \beta - \mathbf{x}^T \beta_{0,k}\|_1 < \delta, \mathbf{x} \in \mathbb{R}^p\}$  for some positive constants  $\delta, c$ , and  $C$ . Condition (C3) is the compatibility condition required to ensure the convergence of lasso estimator in terms of both  $\ell_1$  and  $\ell_2$  norm [3]. When  $p = O(n_k^\delta)$  with  $\delta \in [0, 1)$ , condition (C3) states that  $s_{0,k}$  must be of the order of  $o(n_k^{(1-\delta)/2} / \log p)$  in the GLM, which is slightly stronger than order  $s_0 = o(n_k^{1/2} / \log p)$ , a usual condition required in the linear model; see for examples [40] and [13] and detailed discussion therein. Condition (C4) is the model homogeneity assumption as well as the uniformly bounded assumption across  $K$  sub-datasets, which is required to combine results, as considered in Theorem 2.

**Theorem 1.** Under conditions (C1)–(C3), for  $p = O(n_k^\delta)$ ,  $\delta \in [0, 1)$ , and any fixed integer  $q$ , let  $\mathbf{H}$  be a matrix of rank  $q$  with  $\overline{\sigma}(\mathbf{H}) < \infty$ . Then the fixed-length bias-corrected estimator  $\hat{\gamma}_{\lambda,k} = \mathbf{H} \hat{\beta}_{\lambda,k}^c$ , with  $\hat{\beta}_{\lambda,k}^c$  given in (3), is consistent and asymptotically normally distributed, namely,  $n_k^{1/2}(\hat{\gamma}_{\lambda,k} - \gamma_{0,k}) \xrightarrow{d} \mathcal{N}(0, \mathbf{J}_{\gamma,k}(\beta_{0,k}))$ , as  $n_k \rightarrow \infty$ , where  $\gamma_{0,k} = \mathbf{H} \beta_{0,k}$ , and  $\mathbf{J}_{\gamma,k}(\beta_{0,k}) = E\{-\mathbf{H} \dot{S}_{n_k}^{-1}(\beta_{0,k}) \mathbf{H}^T\}$ .

**Theorem 1** may be viewed as an extension of the covariate-wise asymptotic result in [13] to the joint asymptotic distribution on  $\hat{\gamma}_{\lambda_k, k}$ , a fixed-length sub-vector of  $\hat{\beta}_{\lambda_k, k}^c$ . Matrix  $\mathbf{H}$  chosen under a target subset of parameters allows to perform a joint inference, and univariate inference is a special case with  $q = 1$ . We emphasize the need of a joint asymptotic distribution in order to use the method of confidence distribution in (4) to combine results in Section 3. This is a critical step to yield a combined estimator and related inference. **Corollary 1** establishes the validity of the bias-corrected lasso estimator  $\hat{\gamma}_{\lambda_{cv}, k}$  with  $\lambda_{cv}, k$  being selected via the commonly used R-fold cross-validation procedure. It shows that such  $\lambda_{cv}, k$  satisfies a sufficient condition required by **Theorem 1**, and can be tuned locally within individual sub-datasets. In effect, when the sample sizes  $\{n_k\}_{k=1}^K$  are balanced, a single tuning parameter  $\lambda$  is needed. However, to synchronously tune a common  $\lambda$  across  $K$  sub-datasets will introduce additional overhead cost in communication. Thus, we keep parameter tuning separate. More discussion is given in **Remark 5** in Section 3. A brief proof of **Theorem 1** is given in the **Appendix A**, and the complete proofs of **Theorem 1** and **Corollary 1** are given in the Supplementary Material.

**Corollary 1.** Under the same conditions of **Theorem 1**, for any finite integer  $R$ ,  $\hat{\gamma}_{\lambda_{cv}, k, k}$  has the same asymptotic distribution with the tuning parameter  $\lambda_{cv}, k$  being obtained from  $R$ -fold cross-validation using the  $k$ th sub-dataset.

**Remark 1.** The procedure based on **Theorem 1** for the construction of the confidence density remains valid when the adaptive lasso estimator [41] is used to replace  $\hat{\beta}_\lambda$  in (3) and (4). An adaptive lasso estimator is obtained by  $\hat{\beta}_\lambda = \arg \max_{\beta} \frac{1}{n\phi} \sum_{i=1}^n \{y_i \mathbf{x}_i^T \beta - b(\mathbf{x}_i^T \beta)\} - \lambda \sum_{j=1}^p \hat{w}_j |\beta_j|$ , where the weights  $\{\hat{w}_j\}_{j=1}^p$  are given by  $\hat{w}_j = (|\hat{\beta}_j^{ini}|)^{-\xi}$ , with an initial root- $n$  consistent estimate  $\hat{\beta}^{ini}$  of  $\beta$  and some suitable constant  $\xi > 0$ , which is typically set to 1.

**Remark 2.** Collinearity is often encountered in high-dimensional data analysis where some of the covariates are highly correlated. One solution is to construct the confidence distribution in (4) by using the KKT condition of the elastic net estimator [42]. Another remedy to improve numerical stability is to use a ridge-type estimator by adding a ridge term  $\tau \mathbf{I}_p$ , where  $\tau > 0$ , to stabilize the matrix inverse of  $-\hat{\mathbf{S}}_n(\beta_0)$ , i.e.,  $\{-\hat{\mathbf{S}}_n(\beta_0) + \tau \mathbf{I}_p\}^{-1}$ .

### 3. Combined estimation and inference

We now consider a full data of size  $N$  being partitioned into  $K$  sub-datasets,  $\{(\mathbf{Y}_k, \mathbf{X}_k)\}_{k=1}^K$ , each with size  $n_k$ , and  $N = \sum_{k=1}^K n_k$ . Here, both  $p$  and  $K$  are allowed to diverge along with  $N$ . Let  $n_{\inf} = \inf_{k \in \{1, \dots, K\}} n_k$  be the sample size infimum as  $N$  and  $K$  grow. Consider a target parameter set  $\gamma = \mathbf{H}\beta$ , where  $q = \dim(\gamma)$  is fixed. At a rate  $p = O(n_{\inf}^\delta)$ ,  $\delta \in [0, 1)$ , we obtain  $\hat{\gamma}_k = \arg \max_{\gamma \in \mathbf{H}\beta} \mathcal{L}_{n_k}(\beta; \mathbf{Y}_k, \mathbf{X}_k)$ , where  $\mathcal{L}_{n_k}(\beta; \mathbf{Y}_k, \mathbf{X}_k)$  is the log-likelihood function of the  $k$ th sub-dataset  $(\mathbf{Y}_k, \mathbf{X}_k)$ ,  $k \in \{1, \dots, K\}$ . If there existed a “god-made” computer with unlimited computational capacity to store and process the full data, the centralized MLE could be applied directly to obtain  $\hat{\gamma}_{mle} = \arg \max_{\gamma \in \mathbf{H}\beta} \mathcal{L}_N(\beta; \mathbf{Y}, \mathbf{X}) = \arg \max_{\gamma \in \mathbf{H}\beta} \sum_{k=1}^K \mathcal{L}_{n_k}(\beta; \mathbf{Y}_k, \mathbf{X}_k)$ , where  $\mathcal{L}_N(\beta; \mathbf{Y}, \mathbf{X})$  is the log-likelihood function of the full data  $(\mathbf{Y}, \mathbf{X})$ . Arguably,  $\hat{\gamma}_{mle}$  is the gold standard for inference. There are many ways to combine estimates  $\hat{\gamma}_k$  obtained from sub-datasets. This paper considers using the confidence distribution due to its generalizability under unified objective functions and its ease in establishing statistical inferences. For each sub-dataset  $(\mathbf{Y}_k, \mathbf{X}_k)$ , we first apply **Theorem 1** to construct the asymptotic confidence density  $\hat{h}_{n_k}(\gamma_0)$ ,  $k \in \{1, \dots, K\}$ . Then, in the same spirit as [21], we may combine the  $K$  confidence densities to derive a combined estimator of  $\gamma_0$ , denoted by  $\hat{\gamma}_{dac}$ , where *dac* refers to divide-and-combine, given as follows:

$$\begin{aligned} \hat{\gamma}_{dac} &= \arg \max_{\gamma} \log \prod_{k=1}^K \hat{h}_{n_k}(\gamma) \\ &= \arg \min_{\gamma} \sum_{k=1}^K \frac{1}{2\phi_k} (\gamma - \hat{\gamma}_{\lambda_k, k})^T \left[ \mathbf{H} \left\{ \mathbf{X}_k^T \mathbf{P}_{n_k}(\hat{\beta}_{\lambda_k, k}) \mathbf{X}_k \right\}^{-1} \mathbf{H}^T \right]^{-1} (\gamma - \hat{\gamma}_{\lambda_k, k}), \end{aligned} \quad (5)$$

where  $\hat{\gamma}_{\lambda_k, k} = \mathbf{H}\hat{\beta}_{\lambda_k, k}^c$  and  $\hat{\beta}_{\lambda_k, k}^c$  is the estimate given in (3) with respect to the  $k$ th sub-dataset  $(\mathbf{Y}_k, \mathbf{X}_k)$ . The key advantage of the approach in (5) is to derive an inference procedure for the combined estimator  $\hat{\gamma}_{dac}$ , as stated in **Theorem 2** under diverging  $p = O(n_{\inf}^\delta)$ ,  $\delta \in [0, 1)$ .

**Theorem 2.** Assume  $K = O(N^{1/2-\xi})$ ,  $\xi \in (0, 1/2]$ . Under conditions (C1)–(C4), if  $E \left[ \dot{\mathbf{S}}_{n_k}^{-1}(\beta_0) \mathbf{S}_{n_k}(\beta_0) \right] = \mathbf{0}$  and  $\mathbf{J}_{\gamma, k}(\beta_{0, k})$ , respectively, for all  $k$ , then the MODAC estimator  $\hat{\gamma}_{dac}$  obtained from (5) is consistent and asymptotically normally distributed, namely,  $N^{1/2}(\hat{\gamma}_{dac} - \gamma_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{J}_{dac, \gamma}(\beta_0))$  as  $n_{\inf} \rightarrow \infty$ , with  $\mathbf{J}_{dac, \gamma}(\beta_0) = \mathbf{J}_{\gamma}(\beta_0)$ , where the latter is the centralized Fisher information matrix of the full data. That is, the MODAC  $\hat{\gamma}_{dac}$  is asymptotically as efficient as the centralized MLE  $\hat{\gamma}_{mle}$ .

The key result of **Theorem 2** is that the combined estimator  $\hat{\gamma}_{dac}$  and the gold standard MLE  $\hat{\gamma}_{mle}$  are asymptotically equally efficient. Although it may be tempting to allocate CPUs to speed up computation, the order of  $K$  in **Theorem 2** guides us to choose a proper number of CPUs to ensure that each CPU has enough samples. It is worth noting that the dispersion parameter  $\phi_k$  is not required to be homogeneous across sub-datasets as it does not affect the estimation; and the divide-and-combine estimator  $\hat{\gamma}_{dac}$  does not require additional conditions than those required by



the regularized estimator in each sub-dataset. This is because constructing confidence densities makes the individual asymptotic normal distributions readily available, and the asymptotic distribution of the combined estimator follows. The practical implication of [Theorem 2](#) is that as long as the sample size of each sub-dataset is not too small, the proposed  $\hat{\gamma}_{dac}$  will have little loss of estimation efficiency, while enjoying fast computing in the analysis of big data. The proof of [Theorem 2](#) is given in [Appendix A](#).

For the ease of exposition, without loss of generality, we may take  $\gamma_{dac} = \beta_{dac}$ , i.e.,  $q = p$ . In this way, we can stick on the notation of  $\beta$  in the rest of this paper. By simple algebra, the solution to the divide-and-combine estimator  $\hat{\beta}_{dac}$  [\(5\)](#) can be expressed explicitly as a form of weighted average of  $\hat{\beta}_{\lambda_k, k}^c$ ,  $k = 1, \dots, K$ , as follows:

$$\hat{\beta}_{dac} = \{\sum_{k=1}^K n_k \hat{\Sigma}_{n_k}^{-1}(\hat{\beta}_{\lambda_k, k})\}^{-1} \{\sum_{k=1}^K n_k \hat{\Sigma}_{n_k}^{-1}(\hat{\beta}_{\lambda_k, k}) \hat{\beta}_{\lambda_k, k}^c\}, \quad (6)$$

where  $\hat{\Sigma}_{n_k}^{-1}(\hat{\beta}_{\lambda_k, k}) = (n_k \hat{\phi}_k)^{-1} \mathbf{X}_k^T \mathbf{P}_{n_k}(\hat{\beta}_{\lambda_k, k}) \mathbf{X}_k$ . Note that the inverse matrix  $\hat{\Sigma}_{n_k}^{-1}(\hat{\beta}_{\lambda_k, k})$  in [\(6\)](#) is readily available from the confidence distribution of each sub-dataset. The only matrix inversion required is for the sum of the Fisher information matrices. It follows that the variance-covariance matrix of  $\hat{\beta}_{dac}$  is estimated by  $\hat{\Sigma}_{dac} = \{\sum_{k=1}^K n_k \hat{\Sigma}_{n_k}^{-1}(\hat{\beta}_{\lambda_k, k})\}^{-1}$ , from which confidence regions for any sub-vector of  $\beta$  can be obtained by using standard multivariate analysis methods [\[15\]](#).

**Remark 3.** Note that when  $\lambda_k = 0$  for all  $k$ , our proposed estimator  $\hat{\beta}_{dac}$  in [\(6\)](#) reduces to the classical meta-estimator  $\hat{\beta}_{meta} = \{\sum_{k=1}^K n_k \hat{\Sigma}_{n_k}^{-1}(\hat{\beta}_k)\}^{-1} \{\sum_{k=1}^K n_k \hat{\Sigma}_{n_k}^{-1}(\hat{\beta}_k) \hat{\beta}_k\}$ . Lin and Xi [\[19\]](#) found a similar result as a special case of the aggregated estimating equation estimator. However, the aggregated estimating equation estimator requires a strong assumption of  $K = O(n_{inf}^r)$  ( $r < 1/3$ ), and it does not consider regularized estimation for variable selection. In addition, regardless of  $\hat{\beta}_{meta}$  and  $\hat{\beta}_{dac}$  in [\(6\)](#) taking the same form, they are derived from different criteria with different purposes. Specifically,  $\hat{\beta}_{meta}$  aims to improve statistical power via weighted average, while  $\hat{\beta}_{dac}$  is obtained by minimizing the combined confidence densities for the interest of statistical inference theory. The flexibility of the confidence density approach allows incorporating additional features in the combination; for example, the homogeneity may be relaxed by imposing a mixture of normals in [\(5\)](#), which is not feasible in the meta-estimator.

**Remark 4.** A majority voting approach [\[4\]](#) to combine sparse estimates from  $K$  sub-datasets takes the form  $\hat{\beta}_{mv} = \mathbf{A} \{\sum_{k=1}^K n_k \mathbf{A}^T \hat{\Sigma}_{n_k}(\hat{\beta}_k) \mathbf{A}\}^{-1} \{\sum_{k=1}^K n_k \mathbf{A}^T \hat{\Sigma}_{n_k}(\hat{\beta}_k) \mathbf{A} \hat{\beta}_{k, \hat{\mathcal{A}}^{(v)}}\}$ , where  $\hat{\mathcal{A}}^{(v)} = \{j : \sum_{k=1}^K I(\hat{\beta}_{k,j} \neq 0) > w\}$  is a set of selected signals in terms of a prespecified voting threshold  $w \in [0, K]$ ,  $\hat{\beta}_{k, \hat{\mathcal{A}}^{(v)}}$  denotes a corresponding sub-vector of the lasso estimate  $\hat{\beta}_k$ , and  $\mathbf{A}$  is a  $p \times |\hat{\mathcal{A}}^{(v)}|$  subsetting matrix corresponding to set  $\hat{\mathcal{A}}^{(v)}$ . The majority voting estimator  $\hat{\beta}_{mv}$  has been shown to have the oracle property, which, however, is not applicable to statistical inference.

**Remark 5.** The role of tuning in individual datasets is not to induce sparsity in the final aggregated estimate, but to produce intermediate sparse estimates that give rise to a robust approximation of the covariance in the individual confidence distributions. Since the bias-correction procedure offsets the effect of sparsity tuning, the choice of tuning parameter becomes of little relevance to the means of the derived confidence distributions. The purpose of our integrative inference distinguishes from those estimation methods given in [\[17,33\]](#) that aim to produce aggregated sparse estimates, in which a common tuning parameter has to be chosen across all  $K$  sub-datasets. As a result, their estimation methods require one more round of synchronization, whereas in ours, tuning can be done in parallel (from [Corollary 1](#)).

The overall computational complexity of centralized MLE based on Fisher's scoring is of order  $O(Np^{2+\epsilon})$ ,  $\epsilon \in (0, 1)$  [\[32\]](#), which is dominated by the cost of matrix inversion. The complexity of divided procedures in MODAC involves coordinate descent (of order  $O(2np)$  when  $\lambda$  is given [\[12\]](#)) and evaluating Fisher information matrix (of order  $O(np^2)$ ), for each sub-dataset. The aggregation step involves summation of order  $O(Kp^2)$  and matrix inversion of order  $O(p^{2+\epsilon})$ . Therefore, the complexity under the ideal parallel situation is of order  $O(2np + np^2 + Kp^2 + p^{2+\epsilon})$ . Even in the worst scenario when parallel procedures are run sequentially, the upper bound of overall complexity of MODAC is  $O(Knp^2)$ , which remains comparable to that of the centralized MLE. Similarly, the complexity of the distributed meta-estimator is at order  $O(np^{2+\epsilon} + Kp^2 + p^{2+\epsilon})$  with an upper bound  $O(Knp^{2+\epsilon})$ . The value  $\epsilon$  is purely dependent on the choice of a matrix inversion algorithm, and it ranges over  $(0.3, 0.4)$  for some efficient algorithms.

#### 4. Simulation studies

In this section, we demonstrate the numerical performance of our method under linear, logistic and Poisson regressions through simulation experiments. Specifically, we compare across three divide-and-combine methods, including the meta-analysis method by inverse variance weighted averaging described in [Remark 3](#), the majority voting method described in [Remark 4](#), and our method. Note that when  $K = 1$ , under no data partition, meta-analysis is equivalent to the centralized MLE, the majority voting method is equivalent to the centralized lasso regression [\[31\]](#), and our method is equivalent to centralized lasso with post-selection inference from [Theorem 1](#).

All methods are compared thoroughly on the performance of variable selection, statistical inference and computation time. The evaluation metrics for variable selection include the sensitivity and specificity of correctly identifying non-zero coefficients. The evaluation metrics for statistical inference include mean squared error, absolute bias, coverage

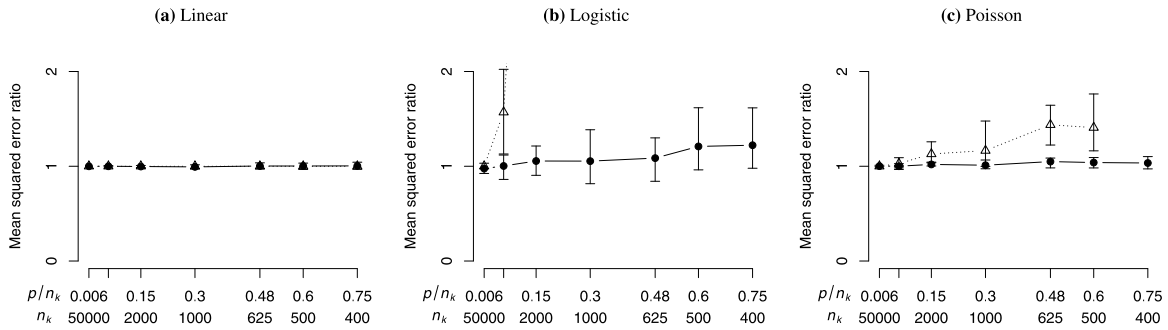
**Table 1**

Simulation results, summarized from 500 replications, under the setting of  $N = 50,000$  and  $p = 300$  for linear, logistic and Poisson models. Methods with different  $K$  are compared. CMLE denotes the centralized MLE method; META denotes the meta-analysis method; MV denotes the majority voting method; and MODAC denotes the proposed method of divide-and-combine.

| Linear model  |                     |                      |                       |                   |   |  |                      |                       |                        |
|---|---------------------|----------------------|-----------------------|-------------------|---|--|----------------------|-----------------------|------------------------|
|   | CMLE<br>( $K = 1$ ) | META<br>( $K = 25$ ) | META<br>( $K = 100$ ) | MV<br>( $K = 1$ ) | MV<br>( $K = 25$ )<br>( $\omega = 12$ ) | MV<br>( $K = 100$ )<br>( $\omega = 50$ ) | MODAC<br>( $K = 1$ ) | MODAC<br>( $K = 25$ ) | MODAC<br>( $K = 100$ ) |
| Sensitivity   | 1.00                | 1.00                 | 1.00                  | 1.00              | 1.00                                    | 1.00                                     | 1.00                 | 1.00                  | 1.00                   |
| Specificity   | 0.95                | 0.95                 | 0.95                  | 0.91              | 1.00                                    | 1.00                                     | 0.95                 | 0.95                  | 0.95                   |
| MSE of $\hat{\beta}_{\mathcal{A}_0}$ ( $\times 100$ )   | 0.01                | 0.01                 | 0.01                  | 0.03              | 0.01                                    | 0.01                                     | 0.01                 | 0.01                  | 0.01                   |
| MSE of $\hat{\beta}_{\mathcal{A}_0^c}$ ( $\times 100$ ) | 0.01                | 0.01                 | 0.01                  | 0.00              | 0.00                                    | 0.00                                     | 0.01                 | 0.01                  | 0.01                   |
| Absolute bias of $\hat{\beta}_{\mathcal{A}_0}$          | 0.01                | 0.01                 | 0.01                  | 0.01              | 0.01                                    | 0.01                                     | 0.01                 | 0.01                  | 0.01                   |
| Absolute bias of $\hat{\beta}_{\mathcal{A}_0^c}$        | 0.01                | 0.01                 | 0.01                  | 0.00              | 0.00                                    | 0.00                                     | 0.01                 | 0.01                  | 0.01                   |
| Cov. prob. of $\beta_{\mathcal{A}_0}$                   | 0.95                | 0.95                 | 0.95                  | –                 | –                                       | –  | 0.95                 | 0.95                  | 0.95                   |
| Cov. prob. of $\beta_{\mathcal{A}_0^c}$                 | 0.95                | 0.95                 | 0.95                  | –                 | –                                       | –  | 0.95                 | 0.95                  | 0.95                   |
| Asymp. st. err. of $\hat{\beta}_{\mathcal{A}_0}$        | 0.01                | 0.01                 | 0.01                  | –                 | –                                       | –  | 0.01                 | 0.01                  | 0.01                   |
| Asymp. st. err. of $\hat{\beta}_{\mathcal{A}_0^c}$      | 0.01                | 0.01                 | 0.01                  | –                 | –                                       | –  | 0.01                 | 0.01                  | 0.01                   |
| Computation time  | 34.85               | 0.62                 | 0.20                  | 31.50             | 2.16                                    | 2.08                                     | 36.61                | 2.28                  | 2.14                   |
| Logistic model  |                     |                      |                       |                   |   |  |                      |                       |                        |
|   | CMLE<br>( $K = 1$ ) | META<br>( $K = 25$ ) | META<br>( $K = 100$ ) | MV<br>( $K = 1$ ) | MV<br>( $K = 25$ )<br>( $\omega = 7$ )  | MV<br>( $K = 100$ )<br>( $\omega = 20$ ) | MODAC<br>( $K = 1$ ) | MODAC<br>( $K = 25$ ) | MODAC<br>( $K = 100$ ) |
| Sensitivity   | 1.00                | 1.00                 | 0.00                  | 1.00              | 1.00                                    | 1.00                                     | 1.00                 | 1.00                  | 1.00                   |
| Specificity   | 0.95                | 1.00                 | 1.00                  | 0.89              | 1.00                                    | 1.00                                     | 0.95                 | 0.95                  | 0.96                   |
| MSE of $\hat{\beta}_{\mathcal{A}_0}$ ( $\times 100$ )   | 0.08                | 0.57                 | 189.38                | 0.23              | 0.20                                    | 0.29                                     | 0.08                 | 0.09                  | 0.10                   |
| MSE of $\hat{\beta}_{\mathcal{A}_0^c}$ ( $\times 100$ ) | 0.08                | 0.05                 | 4.15                  | 0.00              | 0.00                                    | 0.00                                     | 0.08                 | 0.08                  | 0.07                   |
| Absolute bias of $\hat{\beta}_{\mathcal{A}_0}$          | 0.02                | 0.07                 | 1.36                  | 0.04              | 0.04                                    | 0.05                                     | 0.02                 | 0.02                  | 0.02                   |
| Absolute bias of $\hat{\beta}_{\mathcal{A}_0^c}$        | 0.02                | 0.02                 | 0.16                  | 0.00              | 0.00                                    | 0.00                                     | 0.02                 | 0.02                  | 0.02                   |
| Cov. prob. of $\beta_{\mathcal{A}_0}$                   | 0.95                | 0.36                 | 1.00                  | –                 | –                                       | –  | 0.95                 | 0.94                  | 0.92                   |
| Cov. prob. of $\beta_{\mathcal{A}_0^c}$                 | 0.95                | 1.00                 | 1.00                  | –                 | –                                       | –  | 0.95                 | 0.95                  | 0.96                   |
| Asymp. st. err. of $\hat{\beta}_{\mathcal{A}_0}$        | 0.03                | 0.03                 | 1895.12               | –                 | –                                       | –  | 0.03                 | 0.03                  | 0.03                   |
| Asymp. st. err. of $\hat{\beta}_{\mathcal{A}_0^c}$      | 0.03                | 0.03                 | 1893.23               | –                 | –                                       | –  | 0.03                 | 0.03                  | 0.03                   |
| Computation time  | 66.01               | 1.63                 | 1.40                  | 260.48            | 15.78                                   | 10.42                                    | 266.09               | 15.92                 | 10.53                  |
| Poisson model   |                     |                      |                       |                   |   |  |                      |                       |                        |
|   | CMLE<br>( $K = 1$ ) | META<br>( $K = 25$ ) | META<br>( $K = 100$ ) | MV<br>( $K = 1$ ) | MV<br>( $K = 25$ )<br>( $\omega = 7$ )  | MV<br>( $K = 100$ )<br>( $\omega = 26$ ) | MODAC<br>( $K = 1$ ) | MODAC<br>( $K = 25$ ) | MODAC<br>( $K = 100$ ) |
| Sensitivity   | 1.00                | 1.00                 | 1.00                  | 1.00              | 1.00                                    | 1.00                                     | 1.00                 | 1.00                  | 1.00                   |
| Specificity   | 0.95                | 0.94                 | 0.91                  | 0.91              | 1.00                                    | 1.00                                     | 0.95                 | 0.95                  | 0.95                   |
| MSE of $\hat{\beta}_{\mathcal{A}_0}$ ( $\times 100$ )   | 0.70                | 0.80                 | 0.90                  | 1.70              | 0.80                                    | 0.50                                     | 0.70                 | 0.70                  | 0.70                   |
| MSE of $\hat{\beta}_{\mathcal{A}_0^c}$ ( $\times 100$ ) | 0.70                | 0.70                 | 0.80                  | 0.00              | 0.10                                    | 0.00                                     | 0.70                 | 0.70                  | 0.70                   |
| Absolute bias of $\hat{\beta}_{\mathcal{A}_0}$          | 0.01                | 0.01                 | 0.01                  | 0.01              | 0.01                                    | 0.00                                     | 0.01                 | 0.01                  | 0.01                   |
| Absolute bias of $\hat{\beta}_{\mathcal{A}_0^c}$        | 0.01                | 0.01                 | 0.01                  | 0.00              | 0.00                                    | 0.00                                     | 0.01                 | 0.01                  | 0.01                   |
| Cov. prob. of $\beta_{\mathcal{A}_0}$                   | 0.95                | 0.93                 | 0.90                  | –                 | –                                       | –  | 0.95                 | 0.95                  | 0.95                   |
| Cov. prob. of $\beta_{\mathcal{A}_0^c}$                 | 0.95                | 0.94                 | 0.91                  | –                 | –                                       | –  | 0.95                 | 0.95                  | 0.95                   |
| Asymp. st. err. of $\hat{\beta}_{\mathcal{A}_0}$        | 0.01                | 0.01                 | 0.01                  | –                 | –                                       | –  | 0.01                 | 0.01                  | 0.01                   |
| Asymp. st. err. of $\hat{\beta}_{\mathcal{A}_0^c}$      | 0.01                | 0.01                 | 0.01                  | –                 | –                                       | –  | 0.01                 | 0.01                  | 0.01                   |
| Computation time  | 42.26               | 1.46                 | 0.40                  | 132.06            | 26.57                                   | 25.00                                    | 136.85               | 26.67                 | 25.08                  |

probability and asymptotic standard error of coefficients in the signal set  $\mathcal{A}_0$  and the non-signal set  $\mathcal{A}_0^c$ , respectively. Coverage probabilities and standard errors are not reported for the majority voting method since it does not provide inference. We use results from the centralized MLE,  $\hat{\beta}_{mle}$ , as our gold standard in all comparisons. In order to ensure the best variable selection results of the majority voting method, we carefully select  $\omega$  in  $\hat{\beta}_{mv}$  such that the sum of sensitivity and specificity is maximized. The computation time of all methods includes the time taken to read data from disks to memory and the time taken by numerical calculations. Under the divide-and-combine setting with  $K > 1$ , computation time is reported as the sum of the maximum time used among parallelized jobs and the time used to combine results. All shrinkage estimates are obtained by applying the R package `glmnet` with tuning parameter  $\lambda_k$  selected to yield the smallest average 10-fold cross-validated error. All simulation experiments are conducted by R software on a standard Linux cluster with 16 GB of random-access memory per CPU.

Table 1 presents the simulation results from a moderate size dataset with  $N = 50,000$  and  $p = 300$  so that methods without data partition can be repeated in multiple rounds of simulations within a reasonable amount of time. Clearly, this



**Fig. 1.** The y-axis measures the ratio of mean squared error over that of the gold standard CMLE, for regression coefficients in set  $\mathcal{A}_0$ . Median and interquartile ranges of the ratios of META (triangles) and MODAC (solid dots) are shown as the ratio  $p/n_k$  increases. We fix  $N$  at 50,000 and  $p$  at 300. META fails to converge for logistic and Poisson regressions when  $p/n_k$  is large and the results are unavailable.

is a typical regression data setting with  $N \gg p$ . We consider linear, logistic and Poisson models, with responses generated from the mean model  $E(y_i) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}_0)$ ,  $i \in \{1, \dots, N\}$ , with covariates  $\{\mathbf{x}_i\}_{i=1}^N$  generated from the multivariate normal distribution with mean zero and variance one, and with a compound symmetric covariance structure with correlation  $\rho = 0.8$ , a simulation setting similar to that considered by [13]. We report scenarios when the full dataset is randomly divided into  $K = 25$  and  $K = 100$  subsets of equal sizes, each with sample size  $n_k = 2000$  and  $n_k = 500$ , respectively. Results for  $K = 1$  are also reported. We randomly select  $s_0 = 10$  coefficients from  $\boldsymbol{\beta}_0$  to be set at non-zero. The non-zero coefficients are set to 0.3 for linear models, 0.3 for logistic models, and 0.1 for Poisson models. CMLE, META, MV and MODAC denote the centralized MLE, meta-analysis, majority voting and our method of divide-and-combine (MODAC), respectively. Results are based on 500 replications.

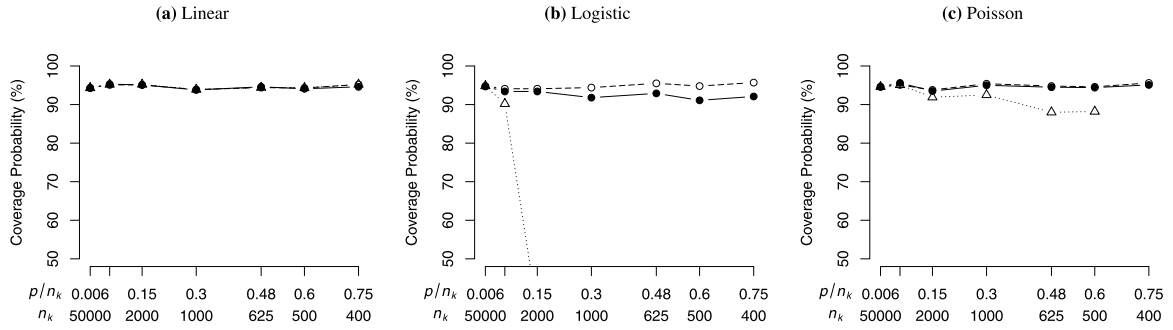
The results of the Gaussian linear model in Table 1 reassuringly show that all methods perform as well as the gold standard. META and MODAC exhibit identical performances as that of CMLE regardless of the choices of  $K$ . This is because under the linear model, CMLE can be directly parallelized, so META and MODAC solutions are exact and identical to CMLE. Among all methods, MV has the highest sensitivity and specificity when  $\omega = 12$  for  $K = 25$  and  $\omega = 50$  for  $K = 100$ . This shows the improvement of selection consistency by divide-and-combine. Under the same model settings, Figs. 1 and 2 display additional simulation results at varying choices of  $K$  with  $N$  fixed at 50,000, summarized over 100 replications. Fig. 1 shows the ratio comparison of mean squared error of META and MODAC, respectively, to that of CMLE, for  $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}$  as  $K$  increases. Fig. 2 compares the coverage probabilities of  $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}$  between CMLE, META and MODAC. Since META and MODAC are identical to CMLE, their mean-squared errors and coverage probabilities are almost identical, as shown in Figs. 1(a) and 2(a).

The existence of exact solutions for divide-and-combine methods under the linear model no longer holds in other generalized linear models, where iterative numerical procedures are needed to search for the estimates. For example in the logistic model, the  $p/n_k$  ratio is responsible for numerical stability, as shown in Figs. 1(b) and 2(b). When  $p/n_k$  approaches one, the mean squared errors and coverage probabilities of META quickly deviate from those of CMLE, whereas MODAC remains stable. Although  $p$  is much smaller than  $N$ , data partitioning may sometimes result in  $p$  closer to  $n_k$  for some sub-datasets. Regularization is shown in our simulation to be an appealing strategy to reduce the dimension of the optimization to achieve more stable numerical performance. The regularization helps stabilize the Newton-Raphson iterative updating algorithm, in which the Hessian matrix may be otherwise poorly estimated in case of  $p/n_k$  being close to one. The numerical results of META appear to be unstable within each sub-dataset in both cases  $K = 25$  and  $K = 100$ . Such poor numerical performance results from the estimated probabilities  $\hat{\mu}_i$  approaching the boundaries in  $[0, 1]$ , causing the variance estimates  $\hat{\mu}_i(1 - \hat{\mu}_i)$  too close to 0. In short, META gives biased parameter estimates and overestimated standard errors of these parameter estimates, and is very sensitive to the choice of  $K$ . On the other hand, through the regularized estimation of  $\mu_i$ , the proposed MODAC exhibits stable performance similar to that of CMLE. The bias of MV for  $\mathcal{A}_0^c$  is higher than that of CMLE as expected due to the  $\ell_1$  penalty.

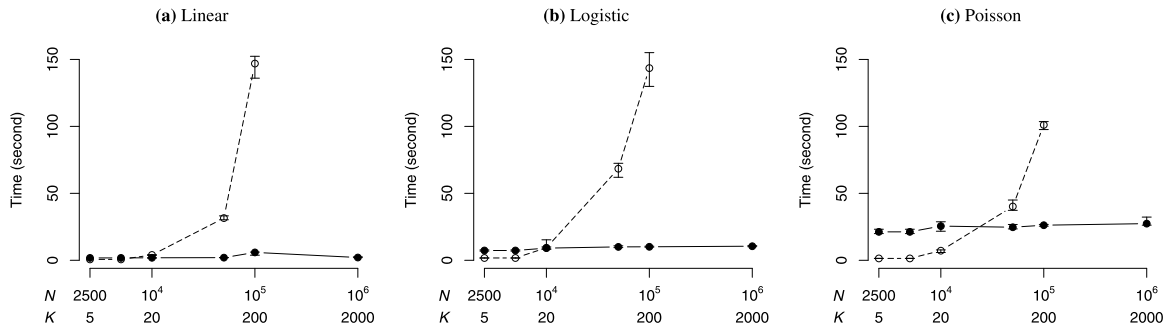
In regard to the Poisson model, Table 1 shows that similar to our findings in the linear and logistic models, MODAC again gives the most stable results among all divide-and-combine methods. On the other hand, META gives improper coverage probabilities in comparison to the nominal 95% level, as well as poorer selection accuracy than CMLE. In Fig. 1(c), we see that the mean squared errors of MODAC are stable against the change of  $p/n_k$ . In contrast, the mean squared errors of META quickly deviate from the mean squared error of CMLE for the Poisson model as  $p/n_k$  increases. In Fig. 2(c), as similar to the logistic model, the 95% confidence interval coverage probabilities of MODAC remains close to the nominal level, whereas the coverage probabilities of META deviate from 95% when  $p/n_k$  gets close to one. MV gives the best variable selection with  $\omega$  carefully chosen.

The key message from Table 1 and Figs. 1–2 is that the invocation of regularization greatly helps to achieve consistent and stable mean and variance estimation in the application of divide-and-combine methods. We see that MODAC produces the most comparable results to those of the gold standard, and is virtually unaffected by the partition size  $K$ . In contrast,





**Fig. 2.** Coverage probabilities of regression coefficients in set  $\mathcal{A}_0$  for the gold standard CMLE (open dots), META (triangles) and MODAC (solid dots) as the ratio of  $p$  and  $n_k$  increases. The total sample size  $N$  and number of covariates  $p$  are fixed at 50,000 and 300, respectively, for all cases. META fails to converge for logistic regression when  $p/n_k \geq 0.3$  and the results are unavailable.



**Fig. 3.** Median computation time and interquartile ranges for CMLE (open dots) and MODAC (solid dots) as  $N$  increases. The sample size of each sub-dataset  $n_k$  in our method is fixed at 500 by increasing  $K$ . CMLE fails when  $N = 10^6$  due to memory limitation and the results are unavailable.

the performances of META and MV vary over the partition size  $K$ . Another noticeable advantage of MODAC is the saving of computation time in comparison to gold standard CMLE due to MODAC's scalability, as shown in Fig. 3 with increase in  $N$  and  $n_k = 500$  in MODAC, based on 100 replications. We see that the computational burden increases sharply for CMLE as  $N$  increases, whereas the computation time for MODAC remains almost the same in all three types of models, which clearly demonstrates its scalability. Computation time for CMLE when  $N = 10^6$  is not reported because the computation exceeds the maximum memory limit allowed on the Linux cluster. In summary, MODAC achieves significant computation time reduction without sacrificing statistical accuracy. Despite the fact that META is the fastest as it does not involve a tuning parameter selection step, its results are clearly unstable in both the logistic and Poisson models. We present additional simulations in the Supplementary Material to show (i) sensitivity of MV regarding choices of  $\omega$ , (ii) sensitivity of MODAC under different levels of correlation in design matrices, and (iii) comparison with another faster version of CMLE given by R package `speedglm`.

## 5. Real data application

We illustrate our method using a publicly available dataset from the National Highway and National Automotive Sampling System Crashworthiness Data System between the years of 2009 and 2015. Details on the access of the data are provided in the Supplementary Material. This national database contains detailed information of about 5000 crashes each year sampled across the United States. The response variable of interest is a binary outcome of injury severity, where 1 corresponds to a crash leading to moderate or severer injury, and 0 for minor or no injury. Most of the predictors included in this study are categorical, and are transformed into dummy variables before regression. Our logistic regression analysis includes 37,535 drivers with 48 predictors after the transformation. The full data are randomly partitioned into  $K = 50$  sub-datasets, each with sample size of about 750. The logistic regression estimation and inference results are provided in Table 2, which shows the estimated coefficients, standard errors and  $p$ -values of 48 potential risk factors obtained by CMLE, META and MODAC. Recall that CMLE is the centralized MLE method, which reads in all  $K$  data batches and fit one logistic regression. CMLE gives the exact solution of maximum likelihood estimate and thus serves as our gold standard for comparisons. In terms of time, MODAC requires 0.66 s, one half of that by CMLE, which is 1.17 s. MODAC yields the exact same inference result as that of CMLE. Although META is the fastest and finishes in 0.03 s, its inference results deviate from those of CMLE and MODAC. For example, as inferred by both CMLE and MODAC, African American is less likely to have moderate to severe injury in a crash than White, and accidents are more likely to result in minor injuries on Wednesday than Sunday; in contrast, META is unable to capture these factors at the same confidence level.

**Table 2**

Estimation and inference results of association study between potential risk factors and binary injury outcome. Logistic model is fitted using the centralized maximum likelihood estimation method (CMLE), the meta-analysis method (META), and our proposed method of divide-and-combine (MODAC). Run time is presented in square brackets.

| Predictors  | CMLE (1.17s) |          |         | META (0.03s) |          |         | MODAC (0.62s) |          |         |
|---|--------------|----------|---------|--------------|----------|---------|---------------|----------|---------|
|   | Estimate     | St. Err. | p-value | Estimate     | St. Err. | p-value | Estimate      | St. Err. | p-value |
| Age   | 0.08         | 0.01     | 0.00    | 0.08         | 0.01     | 0.00    | 0.08          | 0.01     | 0.00    |
| If any other passenger                            | -0.17        | 0.03     | 0.00    | -0.16        | 0.03     | 0.00    | -0.17         | 0.03     | 0.00    |
| If passenger below 14                             | -0.31        | 0.06     | 0.00    | -0.27        | 0.06     | 0.00    | -0.26         | 0.05     | 0.00    |
| If driver female                                  | -0.08        | 0.03     | 0.01    | -0.08        | 0.03     | 0.02    | -0.08         | 0.03     | 0.01    |
| Driver weight                                     | 0.10         | 0.01     | 0.00    | 0.09         | 0.01     | 0.00    | 0.10          | 0.01     | 0.00    |
| Driver height                                     | -0.09        | 0.02     | 0.00    | -0.08        | 0.02     | 0.00    | -0.09         | 0.02     | 0.00    |
| If restraint used                                 | -1.07        | 0.03     | 0.00    | -1.00        | 0.03     | 0.00    | -1.05         | 0.03     | 0.00    |
| Number of lanes                                   | 0.03         | 0.01     | 0.03    | 0.03         | 0.01     | 0.04    | 0.03          | 0.01     | 0.03    |
| Speed limit                                       | 0.01         | 0.01     | 0.65    | 0.00         | 0.01     | 0.81    | 0.00          | 0.01     | 0.72    |
| Vehicle age                                       | 0.01         | 0.01     | 0.43    | 0.01         | 0.01     | 0.40    | 0.01          | 0.01     | 0.40    |
| Vehicle curb weight                               | -0.02        | 0.02     | 0.30    | -0.01        | 0.02     | 0.48    | -0.02         | 0.02     | 0.34    |
| If truck  | -0.05        | 0.04     | 0.19    | -0.05        | 0.04     | 0.21    | -0.05         | 0.04     | 0.18    |
| If vehicle in previous accident                   | -0.11        | 0.03     | 0.00    | -0.10        | 0.03     | 0.00    | -0.10         | 0.03     | 0.00    |
| If four wheel drive                               | 0.01         | 0.04     | 0.69    | 0.02         | 0.04     | 0.67    | 0.01          | 0.04     | 0.70    |
| If drinking involved                              | 0.00         | 0.04     | 0.90    | 0.01         | 0.05     | 0.78    | 0.00          | 0.04     | 0.90    |
| If drug involved                                  | 0.03         | 0.04     | 0.51    | 0.03         | 0.05     | 0.49    | 0.03          | 0.04     | 0.54    |
| If Hispanic                                       | 0.12         | 0.04     | 0.00    | 0.11         | 0.04     | 0.00    | 0.11          | 0.04     | 0.00    |
| If roadway condition bad                          | 0.00         | 0.05     | 0.98    | 0.03         | 0.05     | 0.60    | 0.00          | 0.05     | 0.98    |
| If inclement weather                              | -0.02        | 0.06     | 0.77    | -0.03        | 0.06     | 0.58    | -0.02         | 0.06     | 0.77    |
| Driver race - White (baseline)                    |              |          |         |              |          |         |               |          |         |
| Driver race - Black                               | -0.07        | 0.03     | 0.03    | -0.06        | 0.03     | 0.07    | -0.07         | 0.03     | 0.03    |
| Driver race - Asian                               | -0.08        | 0.07     | 0.23    | -0.01        | 0.07     | 0.83    | -0.08         | 0.07     | 0.23    |
| Region - West (baseline)                          |              |          |         |              |          |         |               |          |         |
| Region - Mid-Atlantic                             | -0.16        | 0.04     | 0.00    | -0.15        | 0.04     | 0.00    | -0.15         | 0.04     | 0.00    |
| Region - Northeast                                | -0.07        | 0.06     | 0.22    | -0.04        | 0.06     | 0.52    | -0.07         | 0.06     | 0.23    |
| Region - Northwest                                | 0.27         | 0.05     | 0.00    | 0.26         | 0.05     | 0.00    | 0.28          | 0.05     | 0.00    |
| Region - South                                    | -0.29        | 0.05     | 0.00    | -0.26        | 0.05     | 0.00    | -0.27         | 0.04     | 0.00    |
| Region - Southeast                                | -0.29        | 0.06     | 0.00    | -0.25        | 0.06     | 0.00    | -0.26         | 0.06     | 0.00    |
| Region - Southwest                                | -0.13        | 0.04     | 0.00    | -0.12        | 0.04     | 0.00    | -0.12         | 0.04     | 0.00    |
| Light condition - daylight (baseline)             |              |          |         |              |          |         |               |          |         |
| Light condition - dark                            | 0.05         | 0.05     | 0.24    | 0.07         | 0.05     | 0.16    | 0.05          | 0.04     | 0.26    |
| Light condition - dawn/dusk                       | -0.02        | 0.06     | 0.76    | 0.03         | 0.07     | 0.71    | -0.02         | 0.06     | 0.76    |
| Light condition - dark/lighted                    | -0.03        | 0.03     | 0.33    | -0.02        | 0.03     | 0.45    | -0.03         | 0.03     | 0.33    |
| Season - Summer (baseline)                        |              |          |         |              |          |         |               |          |         |
| Season - Spring                                   | 0.12         | 0.03     | 0.00    | 0.11         | 0.04     | 0.00    | 0.11          | 0.03     | 0.00    |
| Season - Fall                                     | 0.01         | 0.04     | 0.83    | 0.00         | 0.04     | 0.93    | 0.00          | 0.03     | 0.88    |
| Season - Winter                                   | 0.03         | 0.04     | 0.34    | 0.03         | 0.04     | 0.46    | 0.03          | 0.04     | 0.37    |
| Trafficway flow - divided with barrier (baseline) |              |          |         |              |          |         |               |          |         |
| Trafficway flow - divide without barrier          | 0.02         | 0.04     | 0.64    | 0.01         | 0.04     | 0.73    | 0.01          | 0.04     | 0.71    |
| Trafficway flow - not divided                     | -0.02        | 0.04     | 0.63    | -0.03        | 0.04     | 0.49    | -0.02         | 0.04     | 0.53    |
| Trafficway flow - one way                         | -0.19        | 0.06     | 0.00    | -0.16        | 0.06     | 0.01    | -0.17         | 0.06     | 0.00    |
| Day of Week - Sun (baseline)                      |              |          |         |              |          |         |               |          |         |
| Day of week - Mon                                 | -0.21        | 0.05     | 0.00    | -0.19        | 0.05     | 0.00    | -0.21         | 0.04     | 0.00    |
| Day of week - Tue                                 | -0.22        | 0.05     | 0.00    | -0.21        | 0.05     | 0.00    | -0.21         | 0.04     | 0.00    |
| Day of week - Wed                                 | -0.09        | 0.04     | 0.03    | -0.09        | 0.05     | 0.06    | -0.09         | 0.04     | 0.03    |
| Day of week - Thu                                 | -0.17        | 0.04     | 0.00    | -0.17        | 0.05     | 0.00    | -0.17         | 0.04     | 0.00    |
| Day of week - Fri                                 | -0.15        | 0.04     | 0.00    | -0.15        | 0.04     | 0.00    | -0.15         | 0.04     | 0.00    |
| Day of week - Sat                                 | -0.19        | 0.04     | 0.00    | -0.18        | 0.04     | 0.00    | -0.18         | 0.04     | 0.00    |
| Year - 2009 (baseline)                            |              |          |         |              |          |         |               |          |         |
| Year - 2010                                       | -0.06        | 0.04     | 0.15    | -0.05        | 0.04     | 0.26    | -0.05         | 0.04     | 0.18    |
| Year - 2011                                       | 0.01         | 0.04     | 0.78    | 0.01         | 0.04     | 0.83    | 0.01          | 0.04     | 0.81    |
| Year - 2012                                       | 0.11         | 0.04     | 0.01    | 0.11         | 0.04     | 0.01    | 0.10          | 0.04     | 0.01    |
| Year - 2013                                       | 0.08         | 0.04     | 0.08    | 0.07         | 0.04     | 0.09    | 0.07          | 0.04     | 0.09    |
| Year - 2014                                       | 0.04         | 0.05     | 0.32    | 0.06         | 0.05     | 0.22    | 0.04          | 0.04     | 0.34    |
| Year - 2015                                       | 0.14         | 0.05     | 0.00    | 0.15         | 0.05     | 0.00    | 0.14          | 0.05     | 0.00    |

## 6. Discussion

In this paper, we proposed a scalable regression method in the context of GLM with reliable statistical inference through the seminal work of confidence distribution. Although the divide-and-combine idea has been widely adopted in practice to solve computational challenges arising from the analysis of big data, statistical inference has been little investigated

in such setting. We found in this paper that regularized estimation is appealing in the context of the GLM, especially in the logistic regression because regularization can effectively increase the numerical stability of regression analysis where there are many noisy features. In fact, such divide-and-combine inference may adopt other regularized estimators with regular limiting distributions, but we recommend sparse estimators for better numerical stability in estimating the bias terms and approximating the Fisher information matrices.

In practice, heterogeneity in covariate distributions may arise from various forms of distributed data storage over time and/or space. Some careful analyses are required to understand the nature of heterogeneity, which guide us to choose suitable methods in the integrative inference. Our method is essentially applicable to the targeted regression parameters that are the same across the sub-datasets, while both untargeted regression parameters and parameters of the second moments are allowed to differ across sub-datasets. When such targeted parameters are not clearly defined *a priori*, we may run an additional subgrouping analysis to identify the unknown subpopulations [see examples considered in 30,35], and then apply the proposed method to perform a group-based inference. Additionally, extension to allow unbalanced covariates' distributions and/or missing covariates across sub-datasets is an important direction to account for potential imbalances of data divisions, yet proper inference will require additional conditions similar to those proposed in [18,34] to handle these complications.

Our method can be readily built in into some of the most popular open source parallel computing platforms, such as MapReduce [5] and Spark [38], to handle massive datasets where sample sizes are in the order of millions. Examples include estimating conversion rates using the Criteo online advertising data that have more than 2 million observations [6] and predicting patient disease status based on 9 million patients' electronic health records [33]. Although divide-and-combine is not needed for small datasets, our simulation results show that it is still preferable to impose regularization for large  $p$  using the bias-correction technique. For reproducibility, R code is provided in the Supplementary Material.

## Acknowledgments

The authors thank Editor-in-Chief, Associate Editor and two anonymous reviewers for their constructive comments. Zhou's research was partially supported by Fund of National Natural Science (Nos. 11901470, 11931014 and 11829101) and by Fundamental Research Funds for the Central Universities (Nos. JBK190904 and JBK1806002). Song's research was supported by a National Institutes of Health grant R01 ES024732 and a National Science Foundation grant DMS1811734.

## Appendix A. Proofs

**Proof of Theorem 1.** We present here the key steps in the proof of Theorem 1 and relegate the complete proof to the Supplementary Material (Section S1.1). We explicitly write subscript  $k$  in the proof because the results will be used in Corollary 1. Denote some positive constants by  $C_l, l \in \{1, \dots, 4\}$ . For any fixed integer  $q$ , consider the bias-corrected estimator  $\hat{\gamma}_{\lambda_k, k} = \mathbf{H} \hat{\beta}_{\lambda_k, k}^c$  with  $\hat{\beta}_{\lambda_k, k}^c = \hat{\beta}_{\lambda_k, k} + \{-\hat{\mathbf{S}}_{n_k}(\hat{\beta}_{\lambda_k, k})\}^{-1} \mathbf{S}_{n_k}(\hat{\beta}_{\lambda_k, k})$ , where the lasso estimator  $\hat{\beta}_{\lambda_k, k}$  satisfies the KKT condition  $\mathbf{S}_{n_k}(\hat{\beta}_{\lambda_k, k}) - \lambda_k \hat{\mathbf{e}}_k = 0$ . Let  $\mathbf{A}_{n_k}(\beta) = \frac{1}{n_k} \sum_{i=1}^{n_k} v(\mu_{k,i}) \mathbf{x}_{k,i} \mathbf{x}_{k,i}^T$  and  $\mathbf{B}_{n_k} = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_{k,i} \mathbf{x}_{k,i}^T$ , where  $\mu_{k,i} = g^{-1}(\mathbf{x}_{k,i}^T \beta)$ . Under condition (C2), it is easy to show that for any  $\mu_{k,i} \in \Omega_\delta$ ,

$$c_k b_k^2 \mathbf{I}_{p \times p} \leq c_k \mathbf{B}_{n_k} \leq \mathbf{A}_{n_k}(\beta) \leq C_k \mathbf{B}_{n_k} \leq C_k B_k^2 \mathbf{I}_{p \times p}, \quad (7)$$

where  $\mathbf{A} \leq \mathbf{B}$  indicates  $\mathbf{B} - \mathbf{A}$  is positive semi-definite. So  $\hat{\mathbf{S}}_{n_k}(\beta_{0,k}) = \phi_k^{-1} \mathbf{A}_{n_k}(\beta_0)$  is invertible. With  $PL(\hat{\beta}_\lambda; \mathbf{Y}, \mathbf{X}) \geq PL(\beta_0; \mathbf{Y}, \mathbf{X})$ , we have

$$\begin{aligned} \lambda_k \|\beta_{0,k}\|_1 &\geq \frac{1}{n_k} \left\{ \mathcal{L}_{n_k}(\beta_{0,k}; \mathbf{Y}_k, \mathbf{X}_k) - \mathcal{L}_{n_k}(\hat{\beta}_{\lambda_k, k}; \mathbf{Y}_k, \mathbf{X}_k) \right\} + \lambda_k \|\hat{\beta}_{\lambda_k, k}\|_1 \\ &= -\mathbf{S}_{n_k}(\beta_{0,k})^T (\hat{\beta}_{\lambda_k, k} - \beta_{0,k}) + \frac{1}{2\hat{\phi}_k} (\hat{\beta}_{\lambda_k, k} - \beta_{0,k})^T \mathbf{A}_{n_k}(\tilde{\beta}_k) (\hat{\beta}_{\lambda_k, k} - \beta_{0,k}) + \lambda_k \|\hat{\beta}_{\lambda_k, k}\|_1, \end{aligned}$$

where  $\tilde{\beta}_k$  is a certain value between  $\beta_{0,k}$  and  $\hat{\beta}_{\lambda_k, k}$ . It follows that

$$\|\mathbf{P}_{n_k}^{1/2}(\tilde{\beta}_k) \mathbf{X}_k (\hat{\beta}_{\lambda_k, k} - \beta_{0,k})\|_2^2 / (n_k \hat{\phi}_k) + 2\lambda_k \|\hat{\beta}_{\lambda_k, k}\|_1 \leq 2\mathbf{S}_{n_k}(\beta_{0,k})^T (\hat{\beta}_{\lambda_k, k} - \beta_{0,k}) + 2\lambda_k \|\beta_{0,k}\|_1.$$

According to Corollary 6.2 in [3] and conditions (C1)–(C3), we show that

$$\|\mathbf{X}_k (\hat{\beta}_{\lambda_k, k} - \beta_{0,k})\|_2^2 / n_k + \lambda_k \|\hat{\beta}_{\lambda_k, k} - \beta_{0,k}\|_1 \leq C_1 \lambda_k^2 s_{0,k}, \quad (8)$$

where  $s_{0,k} = \sum_{j=1}^p I(\beta_{0,k,j} \neq 0)$ .

To show the consistency and asymptotic normality of  $\hat{\gamma}_{\lambda_k, k}$ , we begin with the first-order Taylor expansion on the KKT condition. Under conditions (C1)–(C3), we obtain

$$\hat{\gamma}_{\lambda_k, k} - \gamma_{0,k} = \hat{\phi}_k \mathbf{H} \mathbf{A}_{n_k}^{-1}(\beta_{0,k}) \mathbf{S}_{n_k}(\beta_{0,k}) - \mathbf{R}_{n_k}(\tilde{\beta}_k, \beta_{0,k}; \mathbf{H}) + \mathbf{B}_{n_k}(\hat{\beta}_{\lambda_k, k}, \beta_{0,k}; \mathbf{H}), \quad (9)$$

where  $\mathbf{R}_{n_k}(\tilde{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_{0,k}; \mathbf{H}) = \mathbf{H}\mathbf{A}_{n_k}^{-1}(\boldsymbol{\beta}_{0,k})\frac{1}{n_k}\mathbf{X}_k^T \left\{ \mathbf{P}_{n_k}(\tilde{\boldsymbol{\beta}}_k) - \mathbf{P}_{n_k}(\boldsymbol{\beta}_{0,k}) \right\} \mathbf{X}_k(\hat{\boldsymbol{\beta}}_{\lambda_k,k} - \boldsymbol{\beta}_{0,k})$  and

$$\begin{aligned} \mathbf{B}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}, \boldsymbol{\beta}_{0,k}; \mathbf{H}) &= \mathbf{H}\hat{\boldsymbol{\phi}}_k \left\{ \mathbf{A}_{n_k}^{-1}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}) - \mathbf{A}_{n_k}^{-1}(\boldsymbol{\beta}_{0,k}) \right\} \left\{ \mathbf{S}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}) - \mathbf{S}_{n_k}(\boldsymbol{\beta}_{0,k}) \right\} \\ &\quad + \mathbf{H}\hat{\boldsymbol{\phi}}_k \left\{ \mathbf{A}_{n_k}^{-1}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}) - \mathbf{A}_{n_k}^{-1}(\boldsymbol{\beta}_{0,k}) \right\} \mathbf{S}_{n_k}(\boldsymbol{\beta}_{0,k}) \stackrel{\text{def}}{=} \mathbf{I}_{1,k} + \mathbf{I}_{2,k}. \end{aligned}$$

Note that from condition (C2) and (7),

$$\|\mathbf{R}_{n_k}(\tilde{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_{0,k}; \mathbf{H})\|_2^2 \leq c_2 c_k^{-2} b_k^{-4} p \left\{ n_k^{-1} \|\mathbf{X}_k(\hat{\boldsymbol{\beta}}_{\lambda_k,k} - \boldsymbol{\beta}_{0,k})\|_2^2 \right\}^2 \bar{\sigma}^2(\mathbf{H})q. \quad (10)$$

Similarly, we have

$$\begin{aligned} \|\mathbf{I}_{1,k}\|_2^2 &\leq c_3 p \left\{ n_k^{-1} \|\mathbf{X}_k(\hat{\boldsymbol{\beta}}_{\lambda_k,k} - \boldsymbol{\beta}_{0,k})\|_2^2 \right\}^2 (\lambda_k^2 s_{0,k}^2 + \lambda_k s_{0,k} + 1) \bar{\sigma}^2(\mathbf{H})q, \\ \|\mathbf{I}_{2,k}\|_2^2 &\leq c_4 n_k^{-1} \lambda_k^2 s_{0,k}^2 \bar{\sigma}^2(\mathbf{H})q. \end{aligned} \quad (11)$$

Furthermore, by the multivariate Lindeberg–Levy central limit theorem [24] and Slutsky's theorem, the first term in (9) satisfies  $\hat{\boldsymbol{\phi}}_k \mathbf{H} \mathbf{A}_{n_k}^{-1}(\boldsymbol{\beta}_{0,k}) \mathbf{S}_{n_k}(\boldsymbol{\beta}_{0,k}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{J}_{\gamma}(\boldsymbol{\beta}_{0,k}))$  asymptotically as  $n_k \rightarrow \infty$ . Also, under condition (C3) that  $\lambda_k = O\{(\log p/n_k)^{1/2}\}$  and  $s_{0,k} = o\left\{n_k^{\frac{1-\delta}{2}}(\log p)^{-1/2}\right\}$ , inequalities (8), (10) and (11) guarantee  $\|\mathbf{R}_{n_k}(\tilde{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_{0,k}; \mathbf{H})\|_2 = o_p(n_k^{-1/2})$  and  $\|\mathbf{B}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}, \boldsymbol{\beta}_{0,k}; \mathbf{H})\|_2 = o_p(n_k^{-1/2})$ . Thus, Theorem 1 follows.  $\square$

**Proof of Corollary 1.** See the Supplementary Material.  $\square$

**Proof of Theorem 2.** Denote  $\mathbf{r}_N(\boldsymbol{\gamma}) = \frac{1}{N} \sum_{k=1}^K \partial \log \hat{h}_{n_k}(\boldsymbol{\gamma}) / \partial \boldsymbol{\gamma}$  and  $\mathbf{r}(\boldsymbol{\gamma}) = \lim_{n_{\inf} \rightarrow \infty} \mathbf{r}_N(\boldsymbol{\gamma})$ . It is easy to show  $\mathbf{r}(\hat{\boldsymbol{\gamma}}_{dac}) \rightarrow \mathbf{0}$ . On the other hand,

$$\begin{aligned} \mathbf{r}_N(\boldsymbol{\gamma}_0) &= -\frac{1}{N} \sum_{k=1}^K n_k \left\{ \mathbf{H} \dot{\mathbf{S}}_{n_k}^{-1}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}) \mathbf{H}^T \right\}^{-1} \left\{ \boldsymbol{\gamma}_0 - \mathbf{H} \hat{\boldsymbol{\beta}}_{\lambda_k,k} + \mathbf{H} \dot{\mathbf{S}}_{n_k}^{-1}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}) \mathbf{S}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}) \right\} \\ &= \frac{1}{N} \sum_{k=1}^K n_k \left\{ \mathbf{H} \dot{\mathbf{S}}_{n_k}^{-1}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}) \mathbf{H}^T \right\}^{-1} \left\{ \mathbf{H} \dot{\mathbf{S}}_{n_k}^{-1}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}) \mathbf{S}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}) + \mathbf{H}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{\lambda_k,k}) \right\} \\ &= \frac{1}{N} \sum_{k=1}^K n_k \left\{ \mathbf{H} \dot{\mathbf{S}}_{n_k}^{-1}(\boldsymbol{\beta}_0) \mathbf{H}^T \right\}^{-1} \mathbf{H} \dot{\mathbf{S}}_{n_k}^{-1}(\boldsymbol{\beta}_0) \mathbf{S}_{n_k}(\boldsymbol{\beta}_0) + o_p \left\{ N^{-1} \left( \sum_{k=1}^K n_k^{1/2} \right) \right\}, \end{aligned} \quad (12)$$

where the second equality holds under conditions (C1)–(C4). Then, by the law of large numbers,  $\mathbf{r}(\boldsymbol{\gamma}_0) = \sum_{k=1}^K \frac{n_k}{N} E \left[ \left\{ \mathbf{H} \dot{\mathbf{S}}_{n_k}^{-1}(\boldsymbol{\beta}_0) \mathbf{H}^T \right\}^{-1} \mathbf{H} \dot{\mathbf{S}}_{n_k}^{-1}(\boldsymbol{\beta}_0) \mathbf{S}_{n_k}(\boldsymbol{\beta}_0) \right] = \mathbf{0}$ , where the first equation follows from the condition that  $K = o(N)$  and the second equation follows from condition that  $E \left\{ \mathbf{H} \dot{\mathbf{S}}_{n_k}^{-1}(\boldsymbol{\beta}_0) \mathbf{S}_{n_k}(\boldsymbol{\beta}_0) \right\} = \mathbf{0}$ . Furthermore, we have  $\dot{\mathbf{r}}(\boldsymbol{\gamma}_0) = \mathbf{J}_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{\beta}_0)$ , which is a negative-definite matrix given conditions (C1) and (C2). By combining this with  $\mathbf{r}(\boldsymbol{\gamma}_0) = \mathbf{0}$  and  $\mathbf{r}(\hat{\boldsymbol{\gamma}}_{dac}) \rightarrow \mathbf{0}$ , the consistency of  $\hat{\boldsymbol{\gamma}}_{dac}$  follows.

By simple algebra, we obtain

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_{dac} &= \left[ \sum_{k=1}^K n_k \left\{ -\mathbf{H} \dot{\mathbf{S}}_{n_k}^{-1}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}) \mathbf{H}^T \right\}^{-1} \right]^{-1} \left[ \sum_{k=1}^K n_k \left\{ -\mathbf{H} \dot{\mathbf{S}}_{n_k}^{-1}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}) \mathbf{H}^T \right\}^{-1} \hat{\boldsymbol{\gamma}}_{\lambda_k,k} \right] \\ &= \left[ \frac{1}{N} \sum_{k=1}^K n_k \left\{ \mathbf{H} \dot{\mathbf{S}}_{n_k}^{-1}(\boldsymbol{\beta}_0) \mathbf{H}^T \right\}^{-1} \right]^{-1} \left[ \frac{1}{N} \sum_{k=1}^K n_k \left\{ \mathbf{H} \dot{\mathbf{S}}_{n_k}^{-1}(\boldsymbol{\beta}_0) \mathbf{H}^T \right\}^{-1} \hat{\boldsymbol{\gamma}}_{\lambda_k,k} \right] + O_p(N^{-1}K) + o_p(N^{-1/2}), \end{aligned}$$

and  $\text{var}(\hat{\boldsymbol{\gamma}}_{dac}) = N^{-1} \mathbf{J}_{dac, \boldsymbol{\gamma}}(\boldsymbol{\beta}_0)$ . Applying the condition that  $K = O(N^{1/2-\xi})$  with  $\xi \in (0, 1/2]$  and the central limit theorem, we establish the asymptotic normal distribution of  $\hat{\boldsymbol{\gamma}}_{dac}$ .

Finally, it suffices to show that  $\hat{\boldsymbol{\gamma}}_{mle}$  has the same asymptotic distribution as  $\hat{\boldsymbol{\gamma}}_{dac}$ . By the definition of  $\hat{\boldsymbol{\gamma}}_{mle}$  in Theorem 2, we have  $\hat{\boldsymbol{\gamma}}_{mle} - \boldsymbol{\gamma}_0 = -\mathbf{H} \dot{\mathbf{S}}_N^{-1}(\boldsymbol{\beta}_0) \mathbf{S}_N(\boldsymbol{\beta}_0) + o_p(N^{-1/2})$ . The asymptotically equivalent efficiency claimed in Theorem 2 follows by the central limit theorem.  $\square$

## Appendix B. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.jmva.2019.104567>.

## References

- [1] A.V. Aho, J.E. Hopcroft, Design & Analysis of Computer Algorithms, Pearson Education India, 1974.
- [2] H. Battey, J. Fan, H. Liu, J. Lu, Z. Zhu, Distributed testing and estimation under sparse high dimensional models, *Ann. Statist.* 46 (3) (2018) 1352–1382.
- [3] P. Bühlmann, S. van de Geer, Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer Science & Business Media, 2011.
- [4] X. Chen, M. Xie, A split-and-conquer approach for analysis of extraordinarily large data, *Statist. Sinica* 24 (4) (2014) 1655–1684.
- [5] J. Dean, S. Ghemawat, MapReduce: simplified data processing on large clusters, *Commun. ACM* 51 (1) (2008) 107–113.
- [6] E. Diemert, J. Meynet, P. Galland, D. Lefortier, Attribution modeling increases efficiency of bidding in display advertising, in: Proceedings of the AdKDD and TargetAd Workshop, KDD, 14 August 2017, ACM, Halifax, NS, Canada, 2017, page 2.
- [7] D.L. Donoho, J.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, *Biometrika* 81 (3) (1994) 425–455.
- [8] B. Efron, Bayes and likelihood calculations from confidence intervals, *Biometrika* 80 (1) (1993) 3–26.
- [9] J. Fan, F. Han, H. Liu, Challenges of big data analysis, *Natl. Sci. Rev.* 1 (2) (2014) 293–314.
- [10] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.* 96 (456) (2001) 1348–1360.
- [11] R.A. Fisher, Statistical Methods and Scientific Inference, Hafner Publishing Co., Oxford, England, 1956.
- [12] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, *J. Stat. Softw.* 33 (1) (2010) 1.
- [13] S. van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, On asymptotically optimal confidence regions and tests for high-dimensional models, *Ann. Statist.* 42 (3) (2014) 1166–1202.
- [14] L.V. Hedges, I. Olkin, Statistical Methods for Meta-Analysis, Academic Press, 2014.
- [15] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, London, 2002.
- [16] A. Kleiner, A. Talwalkar, P. Sarkar, M.I. Jordan, A scalable bootstrap for massive data, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 76 (4) (2014) 795–816.
- [17] J.D. Lee, Q. Liu, Y. Sun, J.E. Taylor, Communication-efficient sparse regression, *J. Mach. Learn. Res.* 18 (5) (2017) 1–30.
- [18] F. Li, K.L. Morgan, A.M. Zaslavsky, Balancing covariates via propensity score weighting, *J. Amer. Statist. Assoc.* 113 (521) (2018) 390–400.
- [19] N. Lin, R. Xi, Aggregated estimating equation estimation, *Stat. Interface* 4 (1) (2011) 73–83.
- [20] D. Lin, D. Zeng, On the relative efficiency of using summary statistics versus individual-level data in meta-analysis, *Biometrika* 97 (2) (2010) 321–332.
- [21] D. Liu, R.Y. Liu, M. Xie, Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness, *J. Amer. Statist. Assoc.* 110 (509) (2015) 326–340.
- [22] L.W. Mackey, M.I. Jordan, A. Talwalkar, Divide-and-conquer matrix factorization, in: Advances in Neural Information Processing Systems, 2011, pp. 1134–1142.
- [23] P. McCullagh, J.A. Nelder, Generalized Linear Models, Chapman & Hall, 1989.
- [24] R.J. Serfling, Approximation Theorems of Mathematical Statistics, volume 162, John Wiley & Sons, 2009.
- [25] J. Shao, X. Deng, Estimation in high-dimensional linear models with deterministic design matrices, *Ann. Statist.* 40 (2) (2012) 812–831.
- [26] K. Singh, M. Xie, W.E. Strawderman, Combining information from independent sources through confidence distributions, *Ann. Statist.* 33 (1) (2005) 159–183.
- [27] P.X.-K. Song, Correlated Data Analysis: Modeling, Analytics, and Applications, Springer, New York, 2007.
- [28] D. Stangl, D.A. Berry, Meta-Analysis in Medicine and Health Policy, CRC Press, 2000.
- [29] A.J. Sutton, J. Higgins, Recent developments in meta-analysis, *Stat. Med.* 27 (5) (2008) 625–650.
- [30] L. Tang, P.X. Song, Fused lasso approach in regression coefficients clustering: learning parameter heterogeneity in data integration, *J. Mach. Learn. Res.* 17 (1) (2016) 3915–3937.
- [31] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B Stat. Methodol.* (1996) 267–288.
- [32] P. Toulis, E.M. Airoldi, Scalable estimation strategies based on stochastic approximations: classical results and new insights, *Stat. Comput.* 25 (4) (2015) 781–795.
- [33] Y. Wang, N. Palmer, Q. Di, J. Schwartz, I. Kohane, T. Cai, A fast divide-and-conquer sparse cox regression, *Biostatistics* (2019) <http://dx.doi.org/10.1093/biostatistics/kxz036>.
- [34] F. Wang, P.X.-K. Song, L. Wang, Merging multiple longitudinal studies with study-specific missing covariates: A joint estimating function approach, *Biometrics* 71 (4) (2015) 929–940.
- [35] F. Wang, L. Wang, P.X.-K. Song, Fused lasso with the adaptation of parameter ordering in combining multiple studies with repeated measurements, *Biometrics* 72 (4) (2016) 1184–1193.
- [36] M. Xie, K. Singh, Confidence distribution, the frequentist distribution estimator of a parameter: a review, *Internat. Statist. Rev.* 81 (1) (2013) 3–39.
- [37] M. Xie, K. Singh, W.E. Strawderman, Confidence distributions and a unifying framework for meta-analysis, *J. Amer. Statist. Assoc.* 106 (493) (2011) 320–333.
- [38] M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker, I. Stoica, Spark: Cluster computing with working sets, *HotCloud* 10 (10–10) (2010) 95.
- [39] Y. Zhang, J. Duchi, M. Wainwright, Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates, *J. Mach. Learn. Res.* 16 (2015) 3299–3340.
- [40] C.-H. Zhang, S.S. Zhang, Confidence intervals for low dimensional parameters in high dimensional linear models, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 76 (1) (2014) 217–242.
- [41] H. Zou, The adaptive lasso and its oracle properties, *J. Amer. Statist. Assoc.* 101 (476) (2006) 1418–1429.
- [42] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67 (2) (2005) 301–320.