Unifiedly Efficient Inference on All-Dimensional Targets for Large-Scale GLMs

Bo Fu[1], Dandan Jiang[1*]
[1] School of Mathematics, Xi'an Jiaotong University

December 9, 2025

Abstract

The scalability of Generalized Linear Models (GLMs) for large-scale, high-dimensional data often forces a trade-off between computational feasibility and statistical accuracy, particularly for inference on pre-specified parameters. While subsampling methods mitigate computational costs, existing estimators are typically constrained by a suboptimal $r^{-1/2}$ convergence rate, where r is the subsample size. This paper introduces a unified framework that systematically breaks this barrier, enabling efficient and precise inference regardless of the dimension of the target parameters. We propose three estimators tailored to different scenarios. For low-dimensional targets, we propose a de-variance subsampling (DVS) estimator that achieves a markedly improved rate of $\max\{r^{-1}, n^{-1/2}\}$, permitting valid inference even with very small subsamples. As r grows, a multi-step refinement of our estimator is proven to be asymptotically normal and semiparametric efficient when $r/\sqrt{n} \to \infty$, matching the full-sample estimator-confirmed by its Bahadur representation. Critically, for high-dimensional targets, we develop a novel decorrelated score function that facilitates simultaneous inference for a diverging number of pre-specified parameters. Numerical experiments demonstrate that our framework delivers a superior balance of computational efficiency and statistical accuracy across both low- and high-dimensional inferential tasks in large-scale GLM, thereby realizing the promise of unifiedly efficient inference for large-scale GLMs.

Keywords: High-dimensional inference, decorrelated score function, generalized linear models, subsampling

*Corresponging author: Dandan Jiang. Email address: jiangdd@mail.xjtu.edu.cn

1 Introduction

Driven by rapid advancement of technology, data volumes have surged, imposing substantial computational burden and theoretical limitations on conventional approaches. A canonical example arises in regression models involving a scalar response y upon a p-dimensional covariate vector x. The high dimensionality of the covariate x has motivated the development of high-dimensional statistics; for a comprehensive review, see Bühlmann & Van De Geer (2011), Hastie et al. (2015), Wainwright (2019). While a substantial body of literature has focused on estimation and variable selection, the literature on statistical inference has also developed rapidly in recent years. To tackle the high-dimensional inference problem, the debiased Lasso (Javanmard & Montanari 2014, van de Geer et al. 2014, Zhang & Zhang 2014) and the decorrelated score function (Ning & Liu 2017, Fang et al. 2020) emerge as two key statistical techniques that are both semiparametrically efficient, with numerous extensions to a variety of high-dimensional models (Cai et al. 2025, Cheng et al. 2022, Yan et al. 2023, Han et al. 2022, Cai et al. 2023).

In large-scale problems, where both dimensionality p and sample size n grow, computation becomes a major bottleneck. Subsampling provides an effective strategy by selecting a subset of size $r \ll n$ to reduce computational cost. Non-uniform subsampling further improves accuracy by prioritizing informative observations, with designs ranging from leverage-score sampling (Ma et al. 2014) to methods optimizing asymptotic variance (Wang et al. 2018, 2019, Zhang et al. 2021, Yu et al. 2022, Wang & Ma 2021, Fan et al. 2021, Ai et al. 2021). Although most subsampling estimators converge at rate $r^{-1/2}$, Su & Wang (2024) showed that a one-step correction can attain the rate $\max\{n^{-1/2}, r^{-1}\}$. However, existing designs mainly target low-dimensional covariates and may fail when $p > r$.

Motivated by the increasing prominence of large-scale data, recent work (Shan & Wang 2024, Shao et al. 2024) adopts decorrelated-score methods to obtain optimal subsampling

estimators in high-dimensional GLMs. Nevertheless, two limitations remain: (i) subsampling sacrifices accuracy relative to full-sample estimators, yet full-sample analysis incurs substantial computational cost; and (ii) existing procedures target only low-dimensional parameters and do not support high-dimensional simultaneous inference.

In this paper, we address the two aforementioned issues by developing statistically efficient procedures for inference on both low- and high-dimensional parameters in large scale GLMs. Leveraging the one-step idea of Su & Wang (2024), we generalize it to large scale GLMs and propose a de-variance subsampling (DVS) estimator for pre-specified low-dimensional parameters, attaining the convergence rate $\max\{n^{-1/2}, r^{-1}\}$ and enabling valid inference from a small subsample. In particular, when $r/\sqrt{n} \to \infty$, the estimator is asymptotically normal and semiparametrically efficient with the convergence rate $n^{-1/2}$. When full-data accuracy is essential, we further introduce a computationally efficient multi-step estimator that is also semiparametrically efficient. Moreover, in contrast to most prior works, which are restricted to inference on low-dimensional parameters, we generalize the decorrelated score function (Ning & Liu 2017) to high-dimensional parameters of interest by adopting the classic technique (Cai et al. 2011, Yan et al. 2023, Cai et al. 2025) to approximate the inverse Hessian matrix. Since simultaneous inference remains nontrivial in high dimensions, we employ the high-dimensional bootstrap (Chernozhukov et al. 2023) to obtain valid confidence regions. Our contributions can be summarized as follows:

1. **A Unifying Inference Framework for Low-Dimensional Targets.** Beyond merely improving upon existing subsampling estimators, we develop a two-stage inference framework that adapts to varying computational budgets and accuracy requirements. Theoretically, we break the long-standing $r^{-1/2}$ barrier through a novel DVS estimator, which achieves a strictly superior $\max\{n^{-1/2}, r^{-1}\}$ rate. This constitutes a fundamental advancement in subsampling theory. Practically, we complement

this with a multi-step estimator that transitions seamlessly to full-sample statistical efficiency when more computational resources are available. The two methods are unified under a non-asymptotic Bahadur representation framework, which not only certifies their asymptotic normality and semiparametric efficiency but also provides a coherent principle for selecting the appropriate estimator in practice. This systematic approach resolves the trade-off between computational cost and inferential accuracy that has limited prior subsampling methods.

2. **A Scalable Framework for High-Dimensional Simultaneous Inference.** For the challenging task of inference on a diverging number of parameters, we propose a novel subsampled decorrelated score method, which successfully integrates subsampling into high-dimensional simultaneous inference and overcomes the severe computational bottlenecks of full-data methods. By establishing a non-asymptotic Bahadur representation and employing a high-dimensional Gaussian approximation with bootstrap calibration, our method achieves statistical accuracy comparable to the full-sample debiased estimator while offering substantial computational savings. This provides a practical and theoretically sound solution for large-scale simultaneous inference problems previously considered intractable.

The rest of this paper is structured as follows. Section 2 presents a subsampling framework for large-scale GLMs constructed upon the decorrelated score function. Section 3 develops two estimators for the interested low-dimensional parameters in large-scale GLMs, the de-variance subsampling (DVS) and a multi-step estimator, tailored to different computational budgets for inferring low-dimensional parameters of interest in large-scale GLMs. Section 4 extends the decorrelated score approach to enable simultaneous inference on high-dimensional targeted parameters. Simulation studies are reported in Section 5, and a real data analysis is presented in Section 6. Section 7 delivers our concluding remarks and

some avenues for future exploration. Technical proofs and additional simulation results are relegated to the supplementary materials.

Notation: For a vector $\mathbf{v} = (v_1, \cdots, v_p)^\top \in \mathbb{R}^p$, denote $\|\mathbf{v}\| = (\sum_{i=1}^p |v_i|^2)^{1/2}$, $\|\mathbf{v}\|_1 = \sum_{i=1}^p |v_i|$, $\|\mathbf{v}\|_\infty = \max_{1\le i\le p} |v_i|$. For $\mathcal{S} \subseteq \{1, \cdots, p\}$, let $\mathbf{v}_\mathcal{S} = \{v_j, j \in \mathcal{S}\}$ and $\nabla_\mathcal{S} f(\mathbf{v})$ be the gradient of $f(\mathbf{v})$ with respect to $\mathbf{v}_\mathcal{S}$ for a given function f. For a matrix $\mathbf{A} = (a_{ij})_{1\le i\le k_1, 1\le j\le k_2} \in \mathbb{R}^{k_1\times k_2}$, denote $\|\mathbf{A}\|_\infty = \max_{1\le i\le k_1, 1\le j\le k_2} |a_{ij}|$, $\|\mathbf{A}\|_{1,1} = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} |a_{ij}|$, $\|\mathbf{A}\|_{L_1} = \max_{1\le i\le k_1} \sum_{j=1}^{k_2} |a_{ij}|$, and let $\|\mathbf{A}\|$ be the spectral norm of $\mathbf{A}$. We use $\otimes$ to depict the Kronecker product. $b'(t)$, $b''(t)$, $b'''(t)$ are the first, second, and third derivatives of $b(t)$, respectively. We define the sub-Gaussian norm and the sub-exponential norm of a random variable X as $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \le 2\}$ and $\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}\exp(|X|/t) \le 2\}$, respectively. For any two sequences $\{a_n\}$ and $\{b_n\}$, we use $a_n = O(b_n), a_n = o(b_n)$, and $a_n \asymp b_n$ to represent that there exists $c, C > 0$ such that $|a_n| \le C|b_n|$ for all n, $\lim_{n\to\infty} |a_n|/|b_n| = 0$, and $0 < c < |a_n/b_n| < C < \infty$, respectively. $a_n \lesssim b_n$ means $a_n = O(b_n)$. In addition, O_P and o_p have similar meanings as above except that the relationship now holds with high probability.

2 Preliminaries

2.1 Decorrelated score function for GLMs

In this section, we present a concise overview of GLMs and the decorrelated score function. Suppose that the observations $\{y_i, x_i\}_{i=1}^n$ are independent and identically distributed (i.i.d.) drawn from the population $\{y, x\}$ with some joint distribution P, where $y \in \mathbb{R}$ represents the response variable and $x \in \mathbb{R}^p$ is the covariate vector. Within the framework of GLMs with a canonical link, the conditional density of $y|x$ can be expressed as:

$$f(y|x, \beta_0, \sigma_0) \propto \exp\left\{\frac{yx^\top\beta_0 - b(x^\top\beta_0)}{c(\sigma_0)}\right\} = \exp\left\{\frac{y(\theta_0^\top z + \gamma_0^\top u) - b(\theta_0^\top z + \gamma_0^\top u)}{c(\sigma_0)}\right\},$$

where $\beta_0 = (\theta_0^\top, \gamma_0^\top)^\top$ is the unknown parameter vector, and $x^\top = (z^\top, u^\top)$. Specifically, the covariates are partitioned into two components: $z \in \mathbb{R}^d$ are the covariates corresponding to θ that is of interest and $d \leq p$, and $u \in \mathbb{R}^{p-d}$ are the covariates corresponding to nuisance parameter γ. Though d is assumed fixed in many previous works (Ning & Liu 2017, Shan & Wang 2024, Shao et al. 2024), we do not impose this condition in this work. The functions $b(\cdot)$ and $c(\cdot)$ are known functions, with σ_0 a known dispersion parameter σ, which is assumed to be fixed. In this formulation, we concentrate exclusively on θ. To streamline the notation, the intercept term is subsumed into γ_0 without loss of generality.

Consider the negative log-likelihood function $l(\beta_0) = l(\theta_0, \gamma_0) \propto -\log f(y|x, \beta_0, \sigma_0) = b(\theta_0^\top z + \gamma_0^\top u) - y(\theta_0^\top z + \gamma_0^\top u)$. The decorrelated score function in Ning & Liu (2017) takes the form $S(\beta, \mathbf{w}_0) = \nabla_\theta l(\beta) - \mathbf{w}_0 \nabla_\gamma l(\beta)$, where

$$\begin{aligned}\mathbf{w}_0 &= \underset{\mathbf{w}\in\mathbb{R}^{d\times(p-d)}}{\arg\min}\ \mathbb{E}[\|\nabla_\theta l(\beta_0) - \mathbf{w}\nabla_\gamma l(\beta_0)\|]^2 = \underset{\mathbf{w}\in\mathbb{R}^{d\times(p-d)}}{\arg\min}\ \mathbb{E}[b''(\beta_0^\top x)\|z - \mathbf{w}u\|^2] \\ &= \mathbb{E}[b''(\beta_0^\top x) z u^\top] \left(\mathbb{E}[b''(\beta_0^\top x) u u^\top]\right)^{-1}.\end{aligned}$$

The decorrelated score function offers two merits. First, γ does not influence the decorrelated score function around θ_0 since $\mathbb{E}[\nabla_\gamma S(\beta_0, \mathbf{w}_0)] = \mathbf{0}_{d\times(p-d)}$, ensuring the convergence rate of θ is not affected by the high-dimensional nuisance parameters γ. Second, it is orthogonal to the nuisance score function, as expressed by $\mathbb{E}[S(\beta_0, \mathbf{w}_0)^\top \nabla_\gamma l(\beta_0)] = \mathbf{0}_{d\times(p-d)}$. For the i.i.d. observations $\{y_i, x_i\}_{i=1}^n$, the full-data-based decorrelated score function is

$$S(\theta, \gamma_0, \mathbf{w}_0) = \frac{1}{n}\sum_{i=1}^n S_i(\theta, \gamma_0, \mathbf{w}_0) = \frac{1}{n}\sum_{i=1}^n \left(b'(\theta^\top z_i + \gamma_0^\top u_i) - y_i\right)(z_i - \mathbf{w}_0 u_i).$$

2.2 Subsampling-based decorrelated score function

Now we introduce the subsampling decorrelated score function implemented via Poisson subsampling. Denote $\mathcal{K}$ as the index set of a subsample attained via Poisson subsampling

defined later, the corresponding subsampling-based decorrelated score function is

$$S^*(\theta, \gamma_0, \mathbf{w}_0) = \frac{1}{r}\sum_{i\in\mathcal{K}} S_i(\theta, \gamma_0, \mathbf{w}_0) = \frac{1}{r}\sum_{i\in\mathcal{K}} \left(b'(\theta^\top z_i + \gamma_0^\top u_i) - y_i\right)(z_i - \mathbf{w}_0 u_i)\,.$$

Compared to full-data-based approaches, subsampling offers a practical solution to alleviate the computational burden. Here we adopt Poisson subsampling (Yu et al. 2022, Yao et al. 2023, Shan & Wang 2024), a technique that has garnered significant interest due to its computational efficiency relative to sampling with replacement. Specifically, let $\mathcal{P}$ be the index set of subsampled data points, then uniform Poisson subsampling involves generating n i.i.d. Bernoulli random variables $\delta_i, i = 1, \cdots, n$ to determine whether the i-th data point is selected, i.e., $\mathbb{P}(\delta_i = 1) = r/n$, where r is the expected subsample size. The general procedure for uniform Poisson subsampling is presented in Algorithm 1. A key merit

Algorithm 1: General uniform Poisson subsampling algorithm

Initialization $\mathcal{P} = \emptyset$, $p = r/n$, where r is the expected subsample size;
for $i = 1, \cdots, n$ **do**
    Generate a Bernoulli variable $\delta_i \sim \text{Bernoulli}(p)$;
    If $\delta_i = 1$ **do:** Update $\mathcal{P} = \mathcal{P} \cup \{(y_i, x_i)\}$;
end
Output subsample $\mathcal{P}$.

of Poisson subsampling is that the inclusion decision for each data (y_i, x_i) hinges only on δ_i, with the δ_i's, for $i = 1, \cdots, n$, being mutually independent. Moreover, δ_i can be generated either sequentially or in blocks, allowing for efficient parallel processing. We assume $r \leq n$ throughout this paper, a condition that naturally applies in the setting of large-scale datasets. When $r = n$, the subsample degenerates to the original dataset.

Since both the full-data-based decorrelated score function $S(\theta, \gamma_0, \mathbf{w}_0)$ and the subsampled version $S^*(\theta, \gamma_0, \mathbf{w}_0)$ depend on the unknown parameters γ_0 and $\mathbf{w}_0$, we draw another

subsample $\mathcal{K}_{\mathrm{p}}$ with an expected subsample size of r_{p} through Algorithm 1 to obtain their initial estimators similar to Shan & Wang (2024). Specifically, the two Lasso type estimators $\hat{\beta}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}$ can be calculated by employing `R` packages `glmnet` and `pqr`:

$$\hat{\beta}_{\mathrm{p}} = (\hat{\theta}_{\mathrm{p}}^{\top}, \hat{\gamma}_{\mathrm{p}}^{\top})^{\top} = \underset{\beta \in \mathbb{R}^{p}}{\arg\min} \left\{ \frac{1}{r_{\mathrm{p}}} \sum_{i \in \mathcal{K}_{\mathrm{p}}} \left(b(\beta^{\top} x_i^{*\mathrm{p}}) - y_i^{*\mathrm{p}} \beta^{\top} x_i^{*\mathrm{p}} \right) + \lambda \|\beta\|_1 \right\}, \tag{2.1}$$

$$\hat{\mathbf{w}}_{\mathrm{p}} = \underset{\mathbf{w} \in \mathbb{R}^{d \times (p-d)}}{\arg\min} \left\{ \frac{1}{r_{\mathrm{p}}} \sum_{i \in \mathcal{K}_{\mathrm{p}}} b''(\hat{\beta}_{\mathrm{p}}^{\top} x_i^{*\mathrm{p}}) \|z_i^{*\mathrm{p}} - \mathbf{w} u_i^{*\mathrm{p}}\|^2 + \tau \sum_{j=1}^{p-d} \|\mathbf{w}_j\|_1 \right\}, \tag{2.2}$$

where λ, τ are two regularized parameters and $\mathbf{w}_j$ is the j-th column of $\mathbf{w}$. Then the full-data-based estimator $\hat{\theta}_{full} \in \mathbb{R}^d$ is the solution to the equation below:

$$S(\theta, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}) = \frac{1}{n} \sum_{i=1}^{n} S_i(\theta, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}) = \frac{1}{n} \sum_{i=1}^{n} \left(b'(\theta^{\top} z_i + \hat{\gamma}_{\mathrm{p}}^{\top} u_i) - y_i \right) \left(z_i - \hat{\mathbf{w}}_{\mathrm{p}} u_i \right) = \mathbf{0}, \tag{2.3}$$

and the subsampling-based estimator $\hat{\theta}_{uni} \in \mathbb{R}^d$ is the solution to the equation below:

$$\begin{aligned} \mathbf{0} = S^{*}(\theta, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}) &= \frac{1}{r} \sum_{i \in \mathcal{K}} S_i(\theta, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}) = \frac{1}{r} \sum_{i \in \mathcal{K}} \left(b'(\theta^{\top} z_i + \hat{\gamma}_{\mathrm{p}}^{\top} u_i) - y_i \right) \left(z_i - \hat{\mathbf{w}}_{\mathrm{p}} u_i \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{r/n} \left(b'(\hat{\theta}_{\mathrm{p}}^{\top} z_i + \hat{\gamma}_{\mathrm{p}}^{\top} u_i) - y_i \right) \left(z_i - \mathbf{w}_0 u_i \right), \end{aligned} \tag{2.4}$$

where $\mathcal{K} = \{i | \delta_i = 1, i \in \{1, \cdots, n\}\}$ as the index set of subsample attained by performing Algorithm 1 with expected subsample size r.

However, the subsampled estimator $\hat{\theta}_{uni}$ defined in (2.4) only achieves $r^{-1/2}$ rate (Shan & Wang 2024), implying a loss of accuracy compared to the full-data estimator. To perform efficient inference without compromising statistical accuracy, we explore methods to perform fast inference via subsampling techniques in the following sections.

3 Inference for low-dimensional parameters of interest

In this section, we introduce two estimators for efficient inference of low-dimensional parameters of interest in large scale generalized linear models. Both estimators leverage suitably chosen subsample sizes to retain full-data statistical accuracy, while substantially reducing

computational cost. The first, referred to as the DVS estimator, emphasizes computational speed and flexibility, making it well-suited for scenarios where small pilot and main subsample sizes (r_{p} and r) are used. The second, a multi-step estimator, is more computationally economical when full-sample accuracy is desired.

3.1 Inference via the DVS estimator

As an intuitive statement, Ning & Liu (2017) considered a one-step estimator based on an initial Lasso estimator $\hat{\theta}_{\mathrm{p}}$, taking the following form:

$$\hat{\theta}_{\mathrm{p}} - \left(\nabla_\theta S(\hat{\theta}_{\mathrm{p}}, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}})\right)^{-1} S(\hat{\theta}_{\mathrm{p}}, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}), \tag{3.1}$$

except that $\hat{\theta}_{\mathrm{p}}$ in their work is the Lasso estimator based on the full sample. In contrast, we consider the one-step estimation based on $\hat{\theta}_{uni}$ attained via (2.4), which is asymptotically normal (Shan & Wang 2024). The estimator is formally defined as:

Definition 3.1 (de-variance subsampling (*DVS*) estimator)**.** *Continuing the notations in Section 2, define the DVS estimator as*

$$\tilde{\theta} = \hat{\theta}_{uni} - \left(\nabla_\theta S^*(\hat{\theta}_{uni}, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}})\right)^{-1} S(\hat{\theta}_{uni}, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}), \tag{3.2}$$

where $\nabla_\theta S^*(\hat{\theta}_{uni}, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}) = (1/r)\sum_{i\in\mathcal{K}} b''(\hat{\theta}_{uni}^\top z_i + \hat{\gamma}_{\mathrm{p}}^\top u_i)\left(z_i - \hat{\mathbf{w}}_{\mathrm{p}} u_i\right) z_i^\top$, *and* $S(\hat{\theta}_{uni}, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}) = (1/n)\sum_{i=1}^{n}\left(b'(\hat{\theta}_{uni}^\top z_i + \hat{\gamma}_{\mathrm{p}}^\top u_i) - y_i\right)\left(z_i - \hat{\mathbf{w}}_{\mathrm{p}} u_i\right)$.

Note that the additional cost of (3.2) is $O(n(p-d)d)$, which remains small since even computing the L-optimal subsampling probabilities in Shao et al. (2024), Shan & Wang (2024) already requires $O(n(p-d)d)$. Moreover, $\Phi := \mathbb{E}[b''(\beta_0^\top x)(z - \mathbf{w}_0 u)z^\top] = \mathbb{E}[b''(\beta_0^\top x)(z - \mathbf{w}_0 u)(z - \mathbf{w}_0 u)^\top]$ coincides with the submatrix of $(\mathbb{E}[b''(\beta_0^\top x)xx^\top])^{-1}$ corresponding to the coordinates of z, and is therefore symmetric. Thus, the DVS estimator resembles standard debiased estimators for low-dimensional targets, where the main task is approximating

the inverse Hessian, differing mainly in that it builds on $\hat{\theta}_{uni}$ rather than $\hat{\theta}_{\mathrm{p}}$ and incorporates decorrelated score functions with subsampling for computational gains. As shown in Ning & Liu (2017), the full-data decorrelated-score estimator is semiparametrically efficient, in line with debiased estimators in van de Geer et al. (2014), Zhang & Zhang (2014), Javanmard & Montanari (2014). Building on this insight, we use the asymptotic distribution of $\hat{\theta}_{uni}$ to perform inference with reduced pilot and main subsample sizes (r_{p} and r), yielding substantial computational savings while achieving semiparametric efficiency when $r/\sqrt{n} \to \infty$.

To derive the statistical properties of the proposed DVS estimator, we impose the following regularity conditions, which are commonly assumed in high-dimensional GLMs.

(A.1) For some constants $C_1, C_2 > 0$, $\lambda_{\min}(\mathbb{E}[b''(\beta_0^\top x)xx^\top]) \geq C_1$, $\lambda_{\max}(\mathbb{E}[b''(\beta_0^\top x)xx^\top]) \leq C_2$, where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ are the smallest eigenvalues of a given matrix.

(A.2) β_0 is sparse with support $\mathcal{S}_{\beta_0}$ and $|\mathcal{S}_{\beta_0}| = s_1$. The matrix $\mathbf{w}_0$ is sparse with support $\mathcal{S}_{\mathbf{w}_0} = \{j : \mathbf{w}_{0,j} \neq \mathbf{0}, j \in [p-d]\}$ and $|\mathcal{S}_{\mathbf{w}_0}| = s_2$.

(A.3) $\|x\|_\infty \leq C$, $\|y - b'(\beta_0^\top x)\|_{\psi_1} \leq C$, and $\max_{1\leq k\leq d} |(\mathbf{w}_0)_k u| \leq C$ for some constant C, where $(\mathbf{w}_0)_k$ is the k-th row of $\mathbf{w}_0$.

(A.4) There exists some constants $R_1 \leq R_2$ such that $\beta_0^\top x \in [R_1, R_2]$. b is third times differentiable and for any $t \in [R_1 - \varepsilon, R_2 + \varepsilon]$ with some constant $\varepsilon > 0$ and a sequence t_1 satisfying $|t_1 - t| = o(1)$, it holds that $0 < b''(t) \leq C$, $|b''(t_1) - b''(t)| \leq C|t_1 - t|b''(t)$, $|b'''(t)| \leq C$, and $|b'''(t_1) - b'''(t)| \leq C|t_1 - t|$ for some constant C.

(A.5) $\Phi = \mathbb{E}[b''(\beta_0^\top x)(z - \mathbf{w}_0 u)z^\top]$ is finite and $\lambda_{\min}(\Phi) > c$ for some positive constants c.

(A.6) $r^{-1}\log p = o(1)$, $r_{\mathrm{p}}^{-1/2}(s_1 \vee s_2)\log p = o(1)$, $r^{1/2}r_{\mathrm{p}}^{-1}(s_1 \vee s_2)\log p = o(1)$, $(s_1 \vee s_2)\sqrt{\log p/r} = o(1)$. The regularized parameters λ and τ are $\lambda \asymp \tau \asymp \sqrt{\log p/r_{\mathrm{p}}}$.

Assumptions (A.1)-(A.3) and Assumption (A.6) are also assumed by Shao et al. (2024), in which Assumptions (A.1)-(A.3) are common regular conditions for high-dimensional

GLMs (Ning & Liu 2017). Assumption (A.1) imposes an eigenvalue constraint on the design matrix, necessitating that the eigenvalues of $\mathbb{E}[xx^\top]$ are strictly bounded between 0 and ∞, a requirement satisfied when $0 < C_1 < b''(\beta_0^\top x) < C_2$ for some constants C_1 and C_2, and this is encompassed by Assumption (A.4). Assumption (A.2) imposes sparsity conditions on both β_0 and $\mathbf{w}_0$. For the sake of technical simplicity, Assumption (A.3) supposes that the design matrix is bounded and the residuals exhibit sub-exponential tails, which are similarly assumed by Ning & Liu (2017), Shao et al. (2024). Assumption (A.4) imposes a regularity condition on $b(t)$, which holds for a broad class of GLMs. Assumption (A.5) is similar to the condition in (Shao et al. 2024, Theorem 1) and can be directly deduced from Assumption (A.1) together with the Cauchy interlacing theorem in fact. Finally, Assumption (A.6) imposes a minimal restriction on the subsampling size and is non-restrictive.

Remark 3.1. *The sparsity condition for* $\mathbf{w}_0$ *can be changed into* $\max_{1\le k\le d}\|(\mathbf{w}_0)_k\|_0 = s_2$ *if one applies the standard Lasso estimator or the Dantzig type estimator to obtain the estimator of* $\mathbf{w}_0$ *(Cheng et al. 2022, Ning & Liu 2017). From this sight of view, Ning & Liu (2017) states that this condition is identical to the degree of the node Z in the graph if x follows a Gaussian graphical model. This is similar to the assumptions in van de Geer et al. (2014) if we only want to debias the parameters of interest.*

The following theorem states the properties of the DVS estimator, indicating that it can achieve a faster convergence rate than previous subsampling-based estimators.

Theorem 3.1. *Under Assumptions (A.1)-(A.6), we have*

$$\tilde{\theta} - \theta_0 = O_P\Big(n^{-1/2} + r^{-1} + \sqrt{s_1 s_2} r_{\mathrm{p}}^{-1}\log p + r^{-1/2} r_{\mathrm{p}}^{-1} s_1^2 \log p)\Big). \tag{3.3}$$

Furthermore, if r_{p} *satisfies*

$$\min\{\sqrt{n}, r\}\max\left\{\sqrt{s_1 s_2} r_{\mathrm{p}}^{-1}\log p, r^{-1/2} r_{\mathrm{p}}^{-1} s_1^2 \log p, \sqrt{r^{-1} s_2 r_{\mathrm{p}}^{-1}\log p}\right\} = o(1), \tag{3.4}$$

the following holds:

$$\min\{\sqrt{n}, r\}(\tilde{\theta} - \theta_0) \xrightarrow{d} h(U), \tag{3.5}$$

where U is a $(d^2 + 2d)$-dimensional random vector with distribution $\mathbb{N}(\mathbf{0}_{d^2+2d}, V)$, V is partitioned as $[V_{i,j}]_{i,j=1,2,3}$ with

$$\begin{cases} V_{11} = V_{22} = c(\sigma_0)\mathbb{E}[b''(\beta_0^\top x)(z - \mathbf{w}_0 u)(z - \mathbf{w}_0 u)^\top] = c(\sigma_0)\Phi, \\ V_{12} = V_{21}^\top = \sqrt{r/n}V_{11},\ V_{13} = V_{23} = V_{31}^\top = V_{32}^\top = \mathbf{0}_{d\times d^2}, \\ V_{33} = (1 - \dfrac{r}{n})\mathbb{E}[(b''(\beta_0^\top x))^2 vec\{(z - \mathbf{w}_0 u)z^\top\}vec\{(z - \mathbf{w}_0 u)z^\top\}^\top], \end{cases}$$

and

$$h(U) = \frac{1}{2}m_2\Phi^{-1}\mathcal{T}(\theta_0, \gamma_0, \mathbf{w}_0)((\Phi^{-1}U_1) \otimes (\Phi^{-1}U_1)) - m_1\Phi^{-1}U_2 - m_2\Phi^{-1}U_3\Phi^{-1}U_1, \tag{3.6}$$

where $m_1 = \min\{\sqrt{n}, r\}/\sqrt{n}, m_2 = \min\{\sqrt{n}, r\}/r$, $\mathcal{T}(\theta_0, \gamma_0, \mathbf{w}_0)$ is a $d \times d^2$ matrix with its j-th row being $vec\{\mathbb{E}[b'''(\beta_0^\top x)(z_j - (\mathbf{w}_0)_j u)zz^\top]\}$, and $U_1 \in \mathbb{R}^d, U_2 \in \mathbb{R}^d, U_3 \in \mathbb{R}^{d^2}$ denote the first d, next d and last d^2 components of U, respectively.

From Theorem 3.1, we have $\tilde{\theta} - \theta_0 = O_P(n^{-1/2} + r^{-1} + \sqrt{s_1 s_2}r_{\mathrm{p}}^{-1}\log p)$ if $r^{-1/2}s_1^{3/2}s_2^{-1/2} = o(1)$, which demonstrates that $\tilde{\theta} - \theta_0$ is composed of three components. The first component is of order $n^{-1/2}$, which is the convergence rate of the full-data-based estimator. The second component is of order r^{-1}, which is faster than the convergence rates of the subsampled estimators in previous work that are typically $r^{-1/2}$. The third component is $O_P(\sqrt{s_1 s_2}r_{\mathrm{p}}^{-1}\log p)$, which is the error brought by the initial estimators $\hat{\beta}_{\mathrm{p}}$ and $\hat{\mathbf{w}}_{\mathrm{p}}$. However, even when we choose $r_{\mathrm{p}} = r$, the convergence rate of $\tilde{\theta}$ is $\max\{n^{-1/2}, \sqrt{s_1 s_2}r^{-1}\log p\}$, showing that the convergence rate is always faster than $r^{-1/2}$ when $\sqrt{s_1 s_2/r}\log p = o(1)$ and the error is inversely proportional to r when $r \ll \sqrt{n}$, rather than being proportional to $r^{-1/2}$.

For the asymptotic distribution of the DVS estimator $\tilde{\theta}$, we need a larger r_{p} to ensure the accuracy. To guide practical implementations for constructing confidence intervals, we categorize the relationship between r and n into three scenarios to determine how to select pilot subsample size r_{p}. This division facilitates flexible choices of the subsample sizes r and r_{p} to strike a balance between computational cost and the length of confidence interval. We first consider the case when $r/\sqrt{n} \to \infty$. The following theorem exhibits the non-asymptotic Bahadur representation and the asymptotic distribution of the DVS estimator $\tilde{\theta}$. It verifies the semiparametric efficiency (Van der Vaart 2000) of the DVS estimator under $r/\sqrt{n} \to \infty$.

Theorem 3.2 (Case of $r/\sqrt{n} \to \infty$)**.** *Under Assumptions (A.1)-(A.6), we have*

$$\begin{aligned}&\sqrt{n}\{(\tilde{\theta}-\theta_0)+\Phi^{-1}S(\theta_0,\gamma_0,\mathbf{w}_0)\}\\ =&O_P(n^{1/2}r^{-1}+n^{1/2}(\sqrt{s_1s_2}\vee s_2)\log p/r_{\mathrm{p}}+n^{1/2}r^{-1/2}r_{\mathrm{p}}^{-1}s_1^2\log p)+o_P(1).\end{aligned}$$

If $r/\sqrt{n} \to \infty$ and r_{p} satisfies

$$r_{\mathrm{p}}/(\max\{n^{1/2}(\sqrt{s_1s_2}\vee s_2)\log p, n^{1/2}r^{-1/2}s_1^2\log p\}\to\infty, \tag{3.7}$$

we have $\sqrt{n}(\tilde{\theta}-\theta_0)\xrightarrow{d}N(0,c(\sigma_0)\Phi^{-1})$, which is the asymptotic distribution of the solution to $S(\theta,\gamma_0,\mathbf{w}_0)=\mathbf{0}$.

While Theorem 3.2 covers the case $r/\sqrt{n} \to \infty$, the other two scenarios are shown in Corollaries 3.1-3.2 below, using the notation of Theorem 3.1.

Corollary 3.1 (Case of $r/\sqrt{n} \to 0$)**.** *Under Assumptions (A.1)-(A.6), if $r/\sqrt{n} \to 0$ and r_{p} satisfies $r_{\mathrm{p}}/(\max\{r\sqrt{s_1s_2}\log p, r^{1/2}s_1^2\log p, rs_2\log p\}\to\infty$, we have*

$$r(\tilde{\theta}-\theta_0)\xrightarrow{d}\frac{1}{2}\Phi^{-1}\mathcal{T}(\theta_0,\gamma_0,\mathbf{w}_0)((\Phi^{-1}U_1)\otimes(\Phi^{-1}U_1))-\Phi^{-1}U_3\Phi^{-1}U_1.$$

Corollary 3.2 (Case of $r/\sqrt{n} \to C \in (0, \infty)$)**.** *Under Assumptions (A.1)-(A.6), if* $r/\sqrt{n} \to C \in (0, \infty)$ *and* r_{p} *satisfies* (3.7)*, we have*

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} \frac{\sqrt{n}}{2r}\Phi^{-1}\mathcal{J}(\theta_0, \gamma_0, \mathbf{w}_0)((\Phi^{-1}U_1) \otimes (\Phi^{-1}U_1)) - \Phi^{-1}U_2 - \frac{\sqrt{n}}{r}\Phi^{-1}U_3\Phi^{-1}U_1.$$

Corollary 3.1 establishes that for $r/\sqrt{n} \to 0$, the DVS estimator $\tilde{\theta}$ achieves a convergence rate of r and presents the detailed asymptotic distribution. Additionally, Corollary 3.2 further indicates that when $r/\sqrt{n} \to C \in (0, \infty)$, the asymptotic distribution takes the form of a mixture between a Gaussian component (as in the case $r/\sqrt{n} \to \infty$) and an additional term linked to the dominant item under the case of $r/\sqrt{n} \to 0$. In conclusion, regardless of the selection of r relative to $\sqrt{n}$, (3.6) can be utilized to construct the confidence interval by choosing an appropriate value for r_{p}, as discussed previously. Consequently, the confidence interval for the proposed estimator is derived using the Monte Carlo approximation of the asymptotic distribution presented in Theorem 3.1, as outlined in Algorithm 2.

It is noteworthy that in the Monte Carlo step, r_m refers to the number of random samples generated to estimate the quantiles used for constructing the confidence interval. These samples are drawn from $\tilde{h}(\tilde{U})$, where $\tilde{U}$ follows a normal distribution with mean $\mathbf{0}_{d^2+2d}$ and the covariance matrix $\tilde{V}$ is defined in Algorithm 2.

3.2 Inference via the multi-step estimator

In the previous subsection, the asymptotic properties of the DVS estimator is provided, affording practitioners enhanced flexibility in parameter estimation and confidence interval construction across varying selections of r and r_{p} under large scale data contexts. Specifically, if the computational efficiency is required, one can choose a small r and the corresponding small r_{p} to attain the DVS estimator and achieve fast inference. However, if high accuracy is desired, Algorithm 2 may remain computationally intensive while attaining $\hat{\theta}_{uni}$, though reducing much time compared to the full-data-based estimator.

Algorithm 2: Confidence interval construction via the DVS estimator

1 **Pilot Subsampling:** Run Algorithm 1 with expected subsample size r_{p} to get a initial subsample set $\mathcal{K}_{\mathrm{p}}$, then estimate $\hat{\beta}_{\mathrm{p}}$ and $\hat{\mathbf{w}}_{\mathrm{p}}$ as in (2.1) and (2.2) based on $\mathcal{K}_{\mathrm{p}}$.

2 **Subsampling:** Run Algorithm 1 with expected subsample size r to attain the subsample set $\mathcal{K}$, then attain $\hat{\theta}_{uni}$ by solving (2.4) based on the subsample set $\mathcal{K}$.

3 **One-step Estimation:** Get $\tilde{\theta}$ according to (3.2).

4 **Approximation of Function and Model:** For V, Φ, and $\mathcal{T}(\theta_0, \gamma_0, \mathbf{w}_0)$ in Theorem 3.1, replace their population versions with corresponding sample estimates $\tilde{V}$, $\tilde{\Phi}$, and $\mathcal{T}(\tilde{\theta}, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}})$ based on $\mathcal{K}_{\mathrm{p}}$ and substitute the unknown true parameters θ_0, γ_0, and $\mathbf{w}_0$ with their estimations $\tilde{\theta}$, $\hat{\gamma}_{\mathrm{p}}$, and $\hat{\mathbf{w}}_{\mathrm{p}}$. Estimate $h(\cdot)$ by substituting sample estimates $\tilde{\Phi}$ and $\mathcal{T}(\tilde{\theta}, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}})$, denoted as $\tilde{h}(\cdot) = (\tilde{h}_1(\cdot), \cdots, \tilde{h}_d(\cdot))$.

5 **Monte Carlo Simulation:** Generate r_m random samples $\{\tilde{U}^{(1)}, \cdots, \tilde{U}^{(r_m)}\}$ from $\mathbb{N}(\mathbf{0}_{d^2+2d}, \tilde{V})$. Let $\tilde{h}_{L,j}$ and $\tilde{h}_{U,j}$ be the lower and upper $\alpha/2$ quantiles of $\{\tilde{h}_j(\tilde{U}^{(i)})\}_{i=1}^{r_m}$ for $j = 1, \cdots, d$, respectively.

6 **Confidence Interval Construction:** For $j = 1, \cdots, d$, construct the confidence interval of the j-th component of θ_0 with confidence level $1 - \alpha$ as $[\tilde{\theta} - \tilde{h}_{U,j}/\min\{\sqrt{n}, r\}, \tilde{\theta} - \tilde{h}_{L,j}/\min\{\sqrt{n}, r\}]$.

Inspired by Ning & Liu (2017), A naive idea is to construct a one-step estimator directly based on the initial estimator $\hat{\theta}_{\mathrm{p}}$ provided in (2.1), rather than the estimator obtained via $\hat{\theta}_{uni}$. Note that in the previous subsection, $\hat{\theta}_{uni}$ can also be viewed as a one-step estimation based on $\hat{\theta}_{\mathrm{p}}$ and the subsample $\mathcal{K}$, and thus the DVS estimator can be regarded as a two-step estimator. Therefore, we propose a multi-step estimator as follows.

Definition 3.2. *Denote by $\hat{\theta}_{(0)} = \hat{\theta}_{\mathrm{p}}$ the initial Lasso estimator (2.1) based on the pilot subsample $\mathcal{K}_{\mathrm{p}}$. For $\ell = 1, \cdots$, we define*

$$\hat{\theta}_{(\ell)} = \hat{\theta}_{(\ell-1)} - \hat{\Phi}_{\mathrm{p}}^{-1} \frac{1}{n} \sum_{i=1}^{n} \left(b'(\hat{\theta}_{(\ell-1)}^{\top} z_i + \hat{\gamma}_{\mathrm{p}}^{\top} u_i) - y_i \right) \left(z_i - \hat{\mathbf{w}}_{\mathrm{p}} u_i \right), \tag{3.8}$$

where $\hat{\Phi}_{\mathrm{p}} = \nabla_{\theta} S^*(\hat{\theta}_{\mathrm{p}}, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}) = r_{\mathrm{p}}^{-1} \sum_{i \in \mathcal{K}_{\mathrm{p}}} b''(\hat{\beta}_{\mathrm{p}}^{\top} x_i) \left(z_i - \hat{\mathbf{w}}_{\mathrm{p}} u_i \right) z_i^{\top}$ is an estimator of Φ based on the pilot subsample $\mathcal{K}_{\mathrm{p}}$.

Note that only $\hat{\theta}_{(\ell-1)}^{\top} z_i$ and $b'(\cdot)$ should be updated in each iteration. Therefore, compared to the one-step estimator, the additional computational cost of $\hat{\theta}_{(\ell)}$ is $O(nd(\ell-1))$, which is small. Compared with the DVS estimator, the one-step estimator $\hat{\theta}_{(1)}$ saves the time of calculating $\hat{\theta}_{uni}$. The Bahadur representation of $\hat{\theta}_{(\ell)}$ is established as follows.

Theorem 3.3. *Under Assumptions (A.1)-(A.5), for each $\ell = 1, \cdots$, we have*

$$\begin{aligned} &\sqrt{n}\{(\hat{\theta}_{(\ell)} - \theta_0) + \Phi^{-1} S(\theta_0, \gamma_0, \mathbf{w}_0)\} \\ =&O_P(n^{1/2}(s_1 \vee \sqrt{s_1 s_2}) r_{\mathrm{p}}^{-1} \log p + n^{1/2}(s_1^3 \vee s_1^2 s_2)(\log p / r_{\mathrm{p}})^{3/2} + s_1 (\log p / r_{\mathrm{p}})^{1/2} + s_2 \log p / r_{\mathrm{p}}^{1/2}). \end{aligned} \tag{3.9}$$

Specifically, if $s_1 \log p / n^{1/2} = O(1)$ and $s_1 (\log p / r_{\mathrm{p}})^{1/2} = o(1)$, the error in (3.9) can be written as $\sqrt{n}\{(\hat{\theta}_{(\ell)} - \theta_0) + \Phi^{-1} S(\theta_0, \gamma_0, \mathbf{w}_0)\} = \sqrt{n}(\eta + \iota_{(\ell)})$, where $\eta = O_P((s_1^2 \vee s_1 s_2)(\log p / r_{\mathrm{p}})^{3/2} + (s_1 s_2)^{1/2} \log p / r_{\mathrm{p}} + s_1 \log p / (n r_{\mathrm{p}}^{1/2}))$ does not depend on ℓ, and $\iota_{(1)} = O_P(s_1^2 \log p / r_{\mathrm{p}})$, $\iota_{(\ell)} = O_P((\eta + n^{-1/2})^2 + (\eta + n^{-1/2})(\sqrt{\log p / r_{\mathrm{p}}} + s_1^2 \log p / r_{\mathrm{p}})) + (\eta + n^{-1/2} + \sqrt{\log p / r_{\mathrm{p}}} + s_1^2 \log p / r_{\mathrm{p}}) \iota_{(\ell-1)} + \iota_{(\ell-1)}^2)$ for $\ell \geq 2$.

Furthermore, under Assumptions (A.1)-(A.5), if r_{p} *satisfies*

$$r_{\mathrm{p}}/(\max\{n^{1/2}(s_1 \vee \sqrt{s_1 s_2})\log p, n^{1/3}(s_1^2 \vee s_1^{\frac{4}{3}} s_2^{\frac{2}{3}})\log p, s_2^2 \log^2 p\} \to \infty, \tag{3.10}$$

we have $\sqrt{n}(\hat{\theta}_{(\ell)} - \theta_0) \xrightarrow{d} N(0, c(\sigma_0)\Phi^{-1})$.

From Theorem 3.3, it follows that under some conditions and if $\iota_{(1)} = o_P(1), \eta + n^{-1/2} + \sqrt{\log p/r_{\mathrm{p}}} + s_1^2 \log p/r_{\mathrm{p}} = o_P(1)$, the error of the Bahadur representation decreases as ℓ increases. The error of the Bahadur representation of $\hat{\theta}_{(\ell)}$ will finally be dominated by η and the terms in $\iota_{(\ell)}$ that does not depend on $\iota_{(\ell-1)}$. Therefore, we let the iteration proceed until the relative change in the estimation becomes sufficiently small, that is, when the ratio $\left(\text{error}_{(\ell)} - \text{error}_{(\ell-1)}\right)/\text{error}_{(\ell-1)}$ falls below a predefined threshold. Here, $\text{error}_{(\ell)} = \|\hat{\theta}_{(\ell)} - \hat{\theta}_{(\ell-1)}\|$ denotes the stepwise estimation discrepancy between successive iterates. Similar to Theorem 3.2, Theorem 3.3 also implies the semiparametric efficiency of the multi-step estimator $\hat{\theta}_{(\ell)}$ ($\ell \geq 1$). We summarize the procedure for obtaining the multi-step estimator and constructing confidence intervals in Algorithm 3.

We now turn to the choice of r_{p}. Equation (3.9) indicates that the remainder term in the Bahadur representation is asymptotically negligible, and valid confidence intervals may be constructed via Algorithm 3, provided that $r_{\mathrm{p}} \gg \max\{n^{1/2}(s_1 \vee (s_1 s_2)^{1/2})\log p, n^{1/3}(s_1^2 \vee s_1^{4/3} s_2^{2/3})\log p, s_2^2 \log^2 p\}$. In practice, we recommend choosing $r_{\mathrm{p}} = n/C$ for some constant C (e.g., $C = 5$) to ensure computational efficiency while preserving inferential validity.

4 High-dimensional simultaneous inference

This section focuses on a high-dimensional parameter vector $\theta \in \mathbb{R}^d$ of scientific interest, where the dimension d may grow with the sample size. This setting arises in many modern applications, such as genomics, where one jointly tests the effects of many genes, or econometrics, where the goal is to assess a large set of policy variables. In such cases, standard

Algorithm 3: Confidence interval construction via the multi-step estimator

Subsampling: Run Algorithm 1 with expected subsample size r_{p} to get a subsample $\mathcal{K}_{\mathrm{p}}$, and attain $\hat{\beta}_{\mathrm{p}} = (\hat{\theta}_{\mathrm{p}}^{\top}, \hat{\gamma}_{\mathrm{p}}^{\top})^{\top}$ and $\hat{\mathbf{w}}_{\mathrm{p}}$ via (2.1)-(2.2).

Multi-step Estimation: Set $\hat{\theta}_{(0)} = \hat{\theta}_{\mathrm{p}}$, $\mathrm{error}_{(0)} = 1$.

for $\ell = 1, \cdots,$ *maxiter* **do**

  Calculate the multi-step estimator $\hat{\theta}_{(\ell)}$ according to (3.8). Denote $\mathrm{error}_{(\ell)} = \|\hat{\theta}_{(\ell)} - \hat{\theta}_{(\ell-1)}\|$.

  Break if $\left(\mathrm{error}_{(\ell)} - \mathrm{error}_{(\ell-1)}\right)/\mathrm{error}_{(\ell-1)} < 10^{-3}$. Let $\hat{\theta}_{(end)} = \hat{\theta}_{(\ell)}$.

end

Confidence interval construction: Denote the (j,j)-th entry of $\hat{\Phi}_{\mathrm{p}}^{-1}$ as $\nu_{j,j}$. For $j = 1, \cdots, d$, construct the confidence interval of the j-th component of θ_0 with confidence level $1-\alpha$ as $[\hat{\theta}_{(end)} - \nu_{j,j}\rho_{1-\frac{\alpha}{2}}/\sqrt{n}, \hat{\theta}_{(end)} + \nu_{j,j}\rho_{1-\frac{\alpha}{2}}/\sqrt{n}]$, where $\rho_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$-th quantile of the standard normal distribution.

coordinate-wise confidence intervals fail to provide a coherent statistical assessment, as the overall confidence level for jointly covering all parameters deteriorates rapidly due to multiple testing-a "curse of dimensionality'' for inference. Beyond this statistical challenge, the computational cost can become intractable as both n and p grow. Therefore, we develop a scalable framework for simultaneous inference on high-dimensional target parameters by deriving a uniform Bahadur representation for our debiased estimator based on the decorrelated score function, and then applying the multiplier bootstrap for high-dimensional simultaneous inference. This approach enables the construction of simultaneous confidence regions that remain both theoretically valid and computationally efficient even as d diverges.

Motivated by the debiased Lasso (Javanmard & Montanari 2014, van de Geer et al. 2014, Zhang & Zhang 2014) and the decorrelated score function (Ning & Liu 2017, Cheng et al. 2022, Fang et al. 2020), the crucial step for attaining the unbiased parameter is to construct an approximation of the inverse Hessian matrix, which serves a similar role to the estimation of Φ^{-1} in (3.2). When the dimension of parameters of interest is diverge, that is, d is also large, $(\nabla_\theta S^*(\hat{\theta}_{uni}, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}))^{-1}$ in (3.2) or $(\nabla_\theta S^*(\hat{\theta}_{\mathrm{p}}, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}))^{-1}$ in (3.8) may not be a good approximation of Φ^{-1}. Therefore, we adopt the approach in Cai et al. (2011), Yan et al. (2023) to attain the estimator of Φ^{-1}, denoted as $\hat{\mathbf{G}}$, via the following convex program:

$$\min_{\mathbf{G}\in\mathbb{R}^{d\times d}} \|\mathbf{G}\|_{1,1}, \quad \text{s.t.} \quad \|\mathbf{G}\check{\Phi}_{\mathrm{p}} - \mathbf{I}\|_\infty \le \gamma_n, \tag{4.1}$$

where $\check{\Phi}_{\mathrm{p}} = (1/r_{\mathrm{p}}) \sum_{i\in\mathcal{K}_{\mathrm{p}}} b''(\hat{\beta}_{\mathrm{p}}^\top x_i) \left(z_i - \hat{\mathbf{w}}_{\mathrm{p}} u_i\right) \left(z_i - \hat{\mathbf{w}}_{\mathrm{p}} u_i\right)^\top$ is a symmetric covariance-type estimator and is still a sample version of Φ based on the subsample $\mathcal{K}_{\mathrm{p}}$ and utilizing these initial estimators $\hat{\beta}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}$, and γ_n is a predetermined tuning parameter. While Φ is symmetric as illustrated in Section 3.1, $\hat{\mathbf{G}} = (\hat{g}_{i,j})_{1\le i,j\le d}$ is not symmetric typically. To address this problem, one can define $\tilde{\mathbf{G}} = (\tilde{g}_{i,j})_{1\le i,j\le d}$, where $\tilde{g}_{i,j} = \hat{g}_{i,j} I(|\hat{g}_{i,j}| \le |\hat{g}_{j,i}|) + \hat{g}_{j,i} I(|\hat{g}_{j,i}| < |\hat{g}_{i,j}|)$. Therefore, we assume that $\hat{\mathbf{G}}$ is symmetric without loss of

generality in our work. Then we propose the debiased estimator $\check{\theta}$ as follows:

$$\check{\theta} = \hat{\theta}_{\mathrm{p}} - \hat{\mathbf{G}} S(\hat{\theta}_{\mathrm{p}}, \hat{\gamma}_{\mathrm{p}}, \hat{\mathbf{w}}_{\mathrm{p}}) = \hat{\theta}_{\mathrm{p}} - \hat{\mathbf{G}} \frac{1}{n} \sum_{i=1}^{n} \left(b'(\hat{\beta}_{\mathrm{p}}^{\top} x_i) - y_i \right) \left(z_i - \hat{\mathbf{w}}_{\mathrm{p}} u_i \right). \tag{4.2}$$

To further develop the asymptotic behavior of $\check{\theta}$, another assumption is needed.

(B.1) Assume that $\Phi^{-1} = (\mathbf{g}_1, \cdots, \mathbf{g}_d)^{\top} = (g_{i,j})_{1 \le i,j \le d}$ is row-wisely sparse, i.e., $\max_{1 \le i \le d} \sum_{j=1}^{d} |g_{ij}|^q < C_g$ for some $q \in [0, 1)$ and positive constant C_g. The tuning parameter γ_n in (4.1) is set as $\gamma_n \asymp (s_1 \vee \sqrt{s_2})(\log p / r_{\mathrm{p}})^{1/2}$.

This assumption naturally holds for fixed d and requires Φ to be sparse in the sense of the ℓ_q norm of matrix row space if d diverges. Similar conditions can be seen in Cai et al. (2011), Yan et al. (2023), Ning & Liu (2017), Cai et al. (2025). It is the most common matrix sparsity assumption under the case of $q = 0$. More specifically, our sparsity assumptions on Φ^{-1} (Assumption (B.1)) and $\mathbf{w}_0$ (Assumption (A.2)) can be weaker than those sparsity functions on debiased lasso (van de Geer et al. 2014) as they can be deduced from the sparsity of $(\mathbb{E}[b''(\beta_0^{\top} x) x x^{\top}])^{-1}$, as illustrated in Remark 3.1. Based on Assumption (B.1), the estimation error $\|\hat{\mathbf{G}} - \Phi^{-1}\|_{L_1}$ can then be established. In the following theorem, we present the uniform Bahadur representation for $\check{\theta}$ under the case of $q = 0$ for convenience, and the whole analysis is exhibited in the supplementary materials.

Theorem 4.1. *Suppose Assumptions (A.1)-(A.5) and (B.1) hold under the case of $q = 0$, if n satisfies $(s_1 \vee s_2)\sqrt{\log p / n} = o(1)$, then*

$$\|\sqrt{n}\{(\check{\theta} - \theta_0) + \Phi^{-1} S(\theta_0, \gamma_0, \mathbf{w}_0)\}\|_{\infty} = O_P\left(n^{1/2} s_1 (s_1 \vee \sqrt{s_2})(\log p / r_{\mathrm{p}}) + s_2 \log p / r_{\mathrm{p}}^{1/2} \right).$$

Theorem 4.1 illustrates the uniform Bahadur representation for $\check{\theta}$, implying that $\sqrt{n}(\check{\theta} - \theta_0)$ can be expressed by a linear transform of $S(\theta_0, \gamma_0, \mathbf{w}_0)$ up to some ignorable terms. It also exhibits its semiparametric efficiency. If we choose $r_{\mathrm{p}} = n$, i.e., all samples are utilized, then $s_1(s_1 \vee \sqrt{s_2}) \log p / n^{1/2} = o(1)$ and $s_2^2 \log^2(p)/n = o(1)$ is required for the smallest

sample size n to make the residuals to be ignorable. Moreover, the subsampling technique is employed to alleviate computational cost if n is large.

Now we aim to construct simultaneous confidence intervals for θ. From Theorem 4.1, it is natural to derive that the asymptotic distribution of $\|\sqrt{n}(\check{\theta}-\theta)\|_\infty$ can be approximated by the maximum norm of a Gaussian random vector with mean $\mathbf{0}$ and variance $c(\sigma_0)\Phi^{-1}$. This is formally stated in the following proposition.

Proposition 4.1. *Suppose that Assumptions (A.1)-(A.5) and (B.1) hold under the case of* $q=0$, *if* $(s_1\vee s_2)\sqrt{\log p/n}=o(1)$, $n^{1/2}s_1(s_1\vee\sqrt{s_2})(\log p/r_{\mathrm{p}})=o(1)$, $s_2\log p/r_{\mathrm{p}}^{1/2}=o(1)$, $\log^7(dn)/n=o(1)$, *then there exists a* d*-dimensional Gaussian random vector* Q_n *with mean* $\mathbf{0}$ *and variance* $c(\sigma_0)\Phi^{-1}$ *such that*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\|\sqrt{n}(\check{\theta}-\theta)\|_\infty\le t\right)-\mathbb{P}(\|Q_n\|_\infty\le t)\right|=o(1).$$

However, when d is large, it is still unclear whether the maximum norm of Q_n can be reliably estimated by sampling from the d-dimensional Gaussian distribution with an estimated covariance matrix. In contrast, we adopt the multiplier bootstrap procedure. Let $e_i\sim\mathbb{N}(0,1), i=1,\cdots,n$ be independent of the data $\mathcal{D}=\{x_i,y_i\}_{i=1}^n$, and consider

$$\hat{Q}_n^e=\sqrt{n}\hat{\mathbf{G}}\frac{1}{n}\sum_{i=1}^n e_i\left(b'(\hat{\beta}_{\mathrm{p}}^\top x_i)-y_i\right)(z_i-\hat{\mathbf{w}}_{\mathrm{p}}u_i).$$

Denote the α-quantile of $\|\hat{Q}_n^e\|_\infty$ by $c_n(\alpha)=\inf_{t\in\mathbb{R}}\{t\in\mathbb{R}:\mathbb{P}_e(\|\hat{Q}_n^e\|_\infty\le t)\ge\alpha\}$, where $\mathbb{P}_e(\mathcal{A})$ represents the probability of the event $\mathcal{A}$ with respect to $e_1,\cdots,e_n$. The validity of the bootstrap approach is established by the following theorem.

Theorem 4.2. *Under the conditions of Proposition 4.1, if* $\log(p)\log\log(p)\{s_1^2\log p/r_{\mathrm{p}}+(s_1\vee s_2)(\log p/r_{\mathrm{p}})^{1/2}\}=o(1)$, *we have*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\|\hat{Q}_n^e\|_\infty\le t|\mathcal{D}\right)-\mathbb{P}\left(\|Q_n\|_\infty\le t\right)\right|=o(1).$$

Following Proposition 4.1 and Theorem 4.2, denote $\check{\theta} = (\check{\theta}_1, \cdots, \check{\theta}_d)$, the simultaneous α-level confidence interval for $\theta = (\theta_1, \cdots, \theta_d)^\top$ can be constructed as follows:

$$\left(\check{\theta}_i - n^{-1/2} c_n(1-\alpha), \check{\theta}_i + n^{-1/2} c_n(1-\alpha)\right), i = 1, \cdots, d.$$

Moreover, we propose a studentized statistic $\|\sqrt{n}\hat{g}_{j,j}^{-1/2}(\check{\theta} - \theta)_j\|_\infty$ for $j = 1, \cdots, d$ by leveraging the idea in Zhang & Cheng (2017), Cai et al. (2025), which can also be considered to attain confidence intervals with varying length, where $\hat{g}_{j,j}$ is the (j,j)-th element of $\hat{\mathbf{G}}$. Denote $S = \mathrm{diag}(\Phi^{-1})$ and $\hat{S} = \mathrm{diag}(\hat{\mathbf{G}})$. Let

$$\hat{Q}_n^{e,stu} = \sqrt{n}\hat{S}^{-1/2}\hat{\mathbf{G}}\frac{1}{n}\sum_{i=1}^{n} e_i \left(b'(\hat{\beta}_{\mathrm{p}}^\top x_i) - y_i\right)(z_i - \hat{\mathbf{w}}_{\mathrm{p}} u_i),$$

and denote the α-quantile of $\|\hat{Q}_n^{e,stu}\|_\infty$ by $c_n^{stu}(\alpha) = \inf_{t\in\mathbb{R}}\{t \in \mathbb{R} : \mathbb{P}_e(\|\hat{Q}_n^{e,stu}\|_\infty \le t) \ge \alpha\}$. The validity of the studentized bootstrap approach is established by the following theorem.

Theorem 4.3. *Under the conditions of Proposition 4.1, there exists a d-dimensional Gaussian random vector Q_n^{stu} with mean $\mathbf{0}$ and variance $c(\sigma_0)S^{-1/2}\Phi^{-1}S^{-1/2}$ such that*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\|\sqrt{n}\hat{S}^{-1/2}(\check{\theta} - \theta)\|_\infty \le t\right) - \mathbb{P}(\|Q_n^{stu}\|_\infty \le t)\right| = o(1).$$

Furthermore, under the conditions of Theorem 4.2,

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\|\hat{Q}_n^{e,stu}\|_\infty \le t\right) - \mathbb{P}(\|Q_n^{stu}\|_\infty \le t)\right| = o(1).$$

Following Theorem 4.3, the simultaneous α-level confidence interval with varying length for $\theta = (\theta_1, \cdots, \theta_d)^\top$ can be constructed as

$$(\check{\theta}_i - n^{-1/2}\hat{g}_{i,i}^{1/2} c_n^{stu}(1-\alpha), \check{\theta}_i + n^{-1/2}\hat{g}_{i,i}^{1/2} c_n^{stu}(1-\alpha)), \quad i = 1, \cdots, d.$$

The whole steps to attain $\check{\theta}$ and construct simultaneous confidence intervals with varying length are exhibited in Algorithm 4. The corresponding procedure for the

Algorithm 4: Simultaneous confidence intervals through Bootstrap

Subsampling and initial estimation: Run Algorithm 1 with expected subsample size r_{p} to get a initial subsample set $\mathcal{K}_{\mathrm{p}}$, and obtain $\hat{\beta}_{\mathrm{p}}$ and $\hat{\mathbf{w}}_{\mathrm{p}}$ via (2.1) and (2.2). Compute $\check{\Phi}_{\mathrm{p}} = \frac{1}{r_{\mathrm{p}}} \sum_{i \in \mathcal{K}_{\mathrm{p}}} b''(\hat{\beta}_{\mathrm{p}}^{\top} x_i) \left(z_i - \hat{\mathbf{w}}_{\mathrm{p}} u_i\right) \left(z_i - \hat{\mathbf{w}}_{\mathrm{p}} u_i\right)^{\top}$. Solve (4.1) to attain $\hat{\mathbf{G}}$, let $\hat{S} = \mathrm{diag}(\hat{\mathbf{G}}) = \mathrm{diag}(\hat{g}_{1,1}, \cdots, \hat{g}_{d,d})$, and attain $\check{\theta} = (\check{\theta}_1, \cdots, \check{\theta}_d)^{\top}$ via (4.2).

for $k = 1, \cdots, B$ **do**

  Generate standard normal random variables $e_{ki}, i = 1, \cdots, n$.

  Calculate $\hat{Q}_{n,k}^{e,stu} = \sqrt{n} \hat{S}^{-1/2} \hat{\mathbf{G}} \frac{1}{n} \sum_{i=1}^{n} e_{ki} \left(b'(\hat{\beta}_{\mathrm{p}}^{\top} x_i) - y_i \right) \left(z_i - \hat{\mathbf{w}}_{\mathrm{p}} u_i\right)$

end

Confidence interval construction: Find the α-quantile of $\{\|\hat{Q}_{n,k}^{e,stu}\|_{\infty}\}_{k=1}^{B}$, represented as $c_n^{stu}(1-\alpha)$. The simultaneous α-level confidence interval for θ_i is $\left(\check{\theta}_i - n^{-1/2} \hat{g}_{i,i}^{1/2} c_n^{stu}(1-\alpha), \check{\theta}_i + n^{-1/2} \hat{g}_{i,i}^{1/2} c_n^{stu}(1-\alpha)\right), i = 1, \cdots, d.$

non-studentized version is structurally similar and thus omitted for brevity. Note that conditions in Proposition 4.1 implies that r_{p} should be selected such that $r_{\mathrm{p}} \gg \max\{n^{1/2}s_1(s_1 \vee s_2^{1/2})\log p, s_2^2 \log^2 p\}$. In implementation, $r_{\mathrm{p}} = n/C$ for some constant C (e.g., $C = 5$) suffices similar to the multi-step estimator in Section 3.2.

5 Numerical experiments

In this section, we assess the finite-sample performance of our three estimators under both linear and logistic models. We repeated the simulations 500 times, all executed in `R` on a system equipped with an Intel Core i9-12900KF CPU and 64 GB of DDR5 RAM. Due to space limits, the logistic regression results are placed in the supplementary material.

We conduct numerical experiments as follows:

(i) For inference on individual coefficients, we compare our DVS estimator $\tilde{\theta}$ and the multi-step estimator $\hat{\theta}_{(end)}$ with several subsampled decorrelated-score estimators from Shan & Wang (2024), including the A- and L-optimal weighted and unweighted versions $\check{\theta}^{A}_{dw}$, $\check{\theta}^{L}_{dw}$, $\check{\theta}^{A}_{duw}$, $\check{\theta}^{L}_{duw}$, as well as the uniform subsampling version of subsampled decorrelated score estimator $\check{\theta}_{duni}$. For reference, we also report the computation time of the full-data estimator of Ning & Liu (2017).

(ii) For simultaneous inference, we compare our debiased estimator $\check{\theta}$ with nodewise-regression-based debiased estimators from van de Geer et al. (2014), restricting nodewise regression to the parameters of interest for computational fairness. Since their method is not directly applicable to high-dimensional simultaneous inference, we additionally adopt Zhang & Cheng (2017) for linear models that also uses a multiplier bootstrap. For logistic regression, we treat van de Geer et al. (2014) as a baseline, again combined with a multiplier bootstrap. Across all settings, we evaluate both non-studentized and studentized procedures, denoted by $\check{\theta}_{ns}$ and $\check{\theta}_{s}$ for our method, and $\hat{\theta}_{ns}$ and $\hat{\theta}_{s}$ for the

baseline.

We generate datasets of size n from the linear model $y_i = \beta_0^\top x_i + \varepsilon_i$, where $\beta_0 = (\sqrt{3}, \sqrt{3}, \sqrt{3})^\top \in \mathbb{R}^p$, x_i's are i.i.d. covariate vectors, and ε_i are i.i.d. random errors. The covariates x_i and errors ε_i are generated from the following four cases: (a) $x \sim N(\mathbf{0}, \Sigma)$, $\varepsilon \sim N(0,1)$; (b) $x \sim N(\mathbf{0}, \Sigma)$, $\varepsilon \sim 2t_5$; (c) $x \sim t_{10}(\mathbf{0}, \Sigma)$, $\varepsilon \sim N(0,1)$; (d) $x \sim t_{10}(\mathbf{0}, \Sigma)$, $\varepsilon \sim 2t_5$, where $\Sigma \in \mathbb{R}^{p\times p}$ is a Toeplitz matrix with its (j,k)-th component being $0.5^{|j-k|}$.

Since for the linear model, $c(\sigma_0)$ corresponds to the variance of the random error ε. For all the methods, we apply $\sum_{i=1}^n \|y_i - x_i^\top \hat{\beta}_{\mathrm{p}}\|_2^2/(n - \|\hat{\beta}_{\mathrm{p}}\|_0)$ as an estimator of $c(\sigma_0)$, where $\hat{\beta}_{\mathrm{p}}$ is a pilot Lasso estimator defined in (2.1). This estimator, similar to that in Reid et al. (2016), helps avoid noise underestimation when n is small.

5.1 Inference for single coefficient

We first let the whole data size $n = 10^5$ and the pilot subsample size $r_{\mathrm{p}} = 1000$. For the DVS estimator and the subsampling estimators in Shan & Wang (2024), we set subsample size $r = r_{\mathrm{p}}$. The dimension is set to $p = 500$, and the parameters of interest are the first d variables with $d = 5$ or $d = 10$. The confidence level is fixed as $\alpha = 0.05$. The comparisons of the MSE, running time, the average coverage probability (ACP) and the average length (AL) of the 95% confidence intervals are delineated in Table 1.

Take case (a) and $d = 5$ as an example, the full-data-based decorrelated score method takes about 150 seconds. Thus our methods serve as a computationally economical solution for inferring the target low-dimensional parameters if accuracy is desired. Compared with the subsampling estimators in Shan & Wang (2024), it can be seen from Table 1 that the empirical coverage probabilities of all methods are near 95% and they share similar computational costs. Among these methods, $\hat{\theta}_{(end)}$ achieves the smallest MSEs and the

Table 1: MSE, Time (seconds), ACP, and AL for Linear model with $n = 10^5$ and $p = 500$.

	d		$\hat{\theta}_{(end)}$	$\tilde{\theta}$	$\breve{\theta}^{A}_{duw}$	$\breve{\theta}^{L}_{duw}$	$\breve{\theta}^{A}_{dw}$	$\breve{\theta}^{L}_{dw}$	$\breve{\theta}_{duni}$
(a)	5	MSE	8.469×10^{-5}	1.400×10^{-4}	6.780×10^{-3}	6.781×10^{-3}	7.817×10^{-3}	7.806×10^{-3}	7.812×10^{-3}
		Time	1.667	1.646	1.522	1.544	1.526	1.562	1.386
		ACP	0.948	0.946	0.958	0.945	0.938	0.956	0.952
		AL	0.016	0.021	0.140	0.146	0.147	0.154	0.156
	10	MSE	1.743×10^{-4}	3.525×10^{-4}	0.014	0.017	0.015	0.015	0.015
		Time	2.667	2.810	2.868	2.567	2.340	2.571	2.033
		ACP	0.931	0.952	0.956	0.938	0.950	0.952	0.951
		AL	0.016	0.024	0.150	0.154	0.154	0.158	0.158
(b)	5	MSE	1.523×10^{-4}	2.380×10^{-4}	0.011	0.012	0.013	0.013	0.012
		Time	1.849	1.776	1.632	1.536	1.623	1.638	1.431
		ACP	0.940	0.950	0.944	0.952	0.939	0.952	0.960
		AL	0.020	0.027	0.180	0.190	0.190	0.199	0.202
	10	MSE	3.280×10^{-4}	6.871×10^{-4}	0.023	0.024	0.027	0.027	0.027
		Time	2.743	3.089	2.798	2.773	2.812	2.714	2.387
		ACP	0.934	0.937	0.958	0.954	0.940	0.946	0.944
		AL	0.021	0.031	0.192	0.198	0.199	0.206	0.206
(c)	5	MSE	7.942×10^{-5}	1.317×10^{-4}	4.722×10^{-3}	4.719×10^{-3}	5.094×10^{-3}	6.035×10^{-3}	6.166×10^{-3}
		Time	1.719	1.698	1.534	1.498	1.516	1.528	1.334
		ACP	0.929	0.945	0.941	0.952	0.952	0.951	0.960
		AL	0.014	0.020	0.117	0.123	0.127	0.133	0.140
	10	MSE	1.539×10^{-4}	3.473×10^{-4}	0.011	0.011	0.011	0.012	0.012
		Time	2.803	3.174	2.481	2.624	2.537	2.594	2.315
		ACP	0.930	0.957	0.940	0.944	0.955	0.946	0.954
		AL	0.014	0.024	0.126	0.129	0.134	0.138	0.142
(d)	5	MSE	1.082×10^{-4}	2.140×10^{-4}	7.871×10^{-3}	8.936×10^{-3}	9.345×10^{-3}	9.821×10^{-3}	0.010
		Time	1.742	1.712	1.526	1.542	1.552	1.575	1.374
		ACP	0.948	0.942	0.937	0.948	0.951	0.945	0.942
		AL	0.018	0.026	0.151	0.159	0.164	0.172	0.179
	10	MSE	2.442×10^{-4}	5.810×10^{-4}	0.016	0.019	0.020	0.020	0.021
		Time	2.932	3.210	2.952	2.686	2.806	2.576	2.635
		ACP	0.936	0.954	0.954	0.946	0.950	0.956	0.954
		AL	0.021	0.031	0.192	0.198	0.199	0.206	0.206

narrowest average lengths, and the DVS estimator is slightly worse than the multi-step estimator, but much better than other baseline methods. Overall, the average length of our method is several times or even nearly ten times shorter than other baseline methods.

Given the special properties of our DVS estimator $\tilde{\theta}$ and multi-step estimator $\hat{\theta}_{(end)}$, we provide a more intuitive comparison. Fixing $n = 10^5$, $p = 500$, $d = 5$, $r_{\mathrm{p}} = 2000$, we examine how the ACPs and ALs vary with the subsample size r for the DVS estimator and the subsampling methods under case (a). For reference, we also include the multi-step estimator with $r_{\mathrm{p}} = 2000$ and the full-data-based method. The results are depicted in Figure 1, while similar trends in other cases are omitted for brevity.

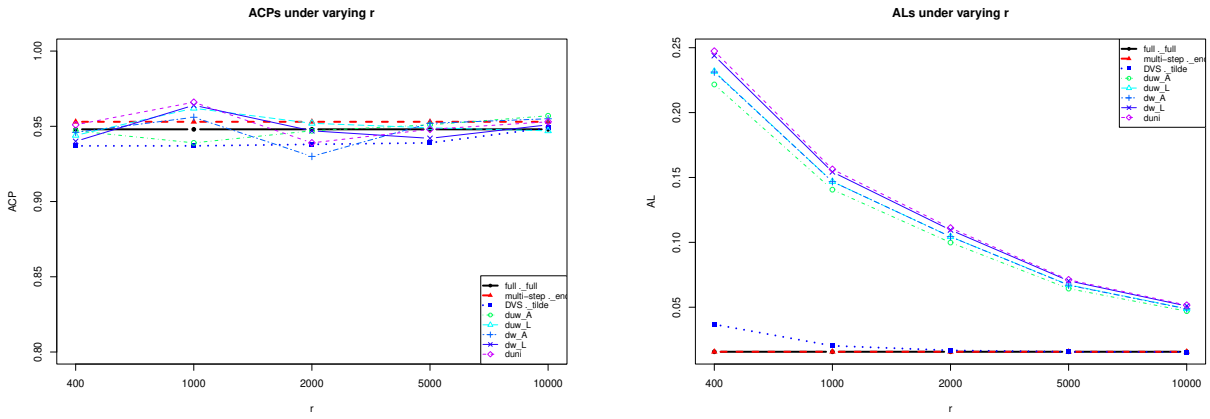

Figure 1: Comparison of ACPs and ALs under varying subsample sizes (r) for the linear model (Case a) with $n = 10^5$, $p = 500$, $d = 5$, and $r_{\mathrm{p}} = 2 \times 10^3$.

Figure 1 demonstrates that both the multi-step and DVS estimators consistently outperform the competing methods. The multi-step estimator attains average interval lengths nearly identical to the full-data method, while the DVS estimator exhibits a rapid decrease in interval length as the subsample size r increases, achieving comparable performance when $r = 2000$. These results further confirm the remarkable efficiency of our proposed approach. The experimental results for logistic regression are consistent with those for linear regression; further details are provided in the supplementary materials.

5.2 Simultaneous inference

Let the sample size n, dimension p, and the pilot subsample size be 5×10^5, 500 and 5000 in our method, respectively. The parameters of interest are the first d variables with $d = 50$. The confidence level is fixed as $\alpha = 0.05$. The comparisons of the MSE, running time, ACP and AL of the 95% simultaneous confidence intervals for four cases are shown in Table 2.

As shown in Table 2, the estimators proposed in Section 4 provide a computationally efficient approach for simultaneous inference of the high-dimensional target parameters. Specifically, the empirical coverage probabilities of all methods remain close to 95%, and their average interval lengths and MSEs are similar, whereas our approach requires only about 10% of the computational time of the baseline method, since merely 10% of the data are utilized to obtain the initial estimator. Therefore, our estimators serve as a an effective and practical alternative to existing methods, as they offer comparable performance with substantially lower computational cost. The results for logistic regression also closely mirror those for linear regression in the supplementary materials.

6 Real data analysis

In this Section, we apply our proposed methods to a superconductivity dataset (Hamidieh 2018) that is available from https://archive.ics.uci.edu/ml/datasets/Superconductivty+Data#. The dataset contains 21263 observations and 81 features, with the critical temperature of superconductors serving as the response variable.

First, we focus on the weighted mean of thermal conductivity (z_1) and the range of atomic radius (z_2) since they are important features (see Table 5 of Hamidieh (2018)). As a benchmark, we employing the decorrelated score approach (Ning & Liu 2017), the estimates of z_1 and z_2 are 0.509 and 0.241, respectively, with a computational time of 1.514 seconds.

Table 2: MSE, Time (seconds), ACP, and AL for simultaneous inference in Linear model with $n = 5 \times 10^4$, $p = 500$, $r_{\mathrm{p}} = 5000$, and $d = 50$.

	$\check{\theta}_{ns}$	$\check{\theta}_{s}$	$\hat{\theta}_{ns}$	$\hat{\theta}_{s}$	$\check{\theta}_{ns}$	$\check{\theta}_{s}$	$\hat{\theta}_{ns}$	$\hat{\theta}_{s}$
	(a)				(b)			
MSE	1.690×10^{-3}	1.689×10^{-3}	1.649×10^{-3}	1.649×10^{-3}	0.011	0.011	0.011	0.011
Time	128.89	130.69	1255.97	1261.22	134.94	138.41	1431.81	1439.47
ACP	0.940	0.945	0.945	0.950	0.932	0.936	0.932	0.932
AL	0.037	0.037	0.037	0.037	0.097	0.097	0.096	0.096
	(c)				(d)			
MSE	1.514×10^{-3}	1.487×10^{-3}	1.311×10^{-3}	1.311×10^{-3}	0.010	0.010	0.009	0.009
Time	131.73	142.49	1444.75	1452.32	138.97	136.71	1438.51	1443.10
ACP	0.936	0.924	0.940	0.932	0.960	0.956	0.952	0.952
AL	0.036	0.035	0.033	0.033	0.092	0.092	0.086	0.086

The corresponding 95% confidence intervals are $[0.455, 0.564]$ and $[0.194, 0.288]$, respectively, suggesting that superconductors with higher weighted mean of thermal conductivity and range of atomic radius tend to exhibit higher critical temperatures. Now we compare the estimators in Section 5.1 as follows. The multi-step estimator $\widehat{\theta}_{(end)}$ is applied with $r_{\mathrm{p}} = 5000$, while $r = r_{\mathrm{p}}$ is chosen as $\{500, 1000, 2000, 5000\}$ to leverage the DVS estimator $\tilde{\theta}$, and other baseline works $\check{\theta}^{A}_{dw}$, $\check{\theta}^{L}_{dw}$, $\check{\theta}^{A}_{duw}$, $\check{\theta}^{L}_{duw}$, and $\check{\theta}_{duni}$. Bias relative to the full-sample decorrelated estimates, standard deviation, and computational time (over 200 replications) are reported in Table 3, with 95% confidence intervals from one replication.

Table 3 shows that the proposed DVS estimator generally achieves the smallest SD, particularly for subsample sizes 500 or 1000, indicating more stable estimates than Shan & Wang (2024) under limited subsampling. Under the same r_{p}, all methods incur similar computational costs and are much faster than the full-sample decorrelated estimators. However, the DVS estimator demonstrates superior inferential performance. Notably, when $r_{\mathrm{p}} = 1000$,

Table 3: Bias, standard error (SD), Time (seconds), and confidence interval (CI) for z_1 and z_2 in the superconductivity dataset.

r_{p}		$\hat{\theta}_{(end)}$	$\tilde{\theta}$	$\breve{\theta}^{A}_{duw}$	$\breve{\theta}^{L}_{duw}$	$\breve{\theta}^{A}_{dw}$	$\breve{\theta}^{L}_{dw}$	$\breve{\theta}_{duni}$
	Time	–	0.356	0.340	0.335	0.366	0.367	0.352
	Bias1	–	-0.076	-0.126	-0.130	-0.154	-0.085	-0.125
	SD1	–	**0.039**	0.050	0.066	0.072	0.062	0.061
500	CI1	–	[-0.133,0.414]	[-2.040, 1.312]	[0.784,1.537]	[0.140,0.700]	[0.214,0.574]	[0.381,0.668]
	Bias2	–	-0.021	-0.020	0.052	-0.019	0.091	-0.022
	SD2	–	**0.030**	0.058	0.041	0.051	0.068	0.059
	CI2	–	[0.200,0.798]	[-1.776,3.767]	[-8.820,5.144]	[-0.023,0.524]	[-0.361,0.434]	[0.094,0.584]
	Time	–	0.378	0.376	0.368	0.372	0.357	0.364
	Bias1	–	-0.006	-0.116	-0.023	-0.076	0.068	-0.030
	SD1	–	**0.037**	0.046	0.052	0.059	0.052	0.059
1000	CI1	–	[0.117,0.838]	[-0.568,0.448]	[-0.352,1.757]	[-0.433,0.821]	[0.363,1.098]	[0.223,1.296]
	Bias2	–	0.025	-0.019	0.048	0.019	-0.060	0.021
	SD2	–	**0.037**	0.061	0.053	0.040	0.052	0.042
	CI2	–	[0.250,0.790]	[-0.282,1.113]	[0.008,0.809]	[-0.310,0.256]	[-0.249,0.451]	[-0.146,0.226]
	Time	–	0.410	0.408	0.424	0.424	0.436	0.418
	Bias1	–	-0.073	-0.070	-0.072	-0.036	-0.016	-0.056
	SD1	–	0.036	0.039	0.043	**0.031**	0.044	0.061
2000	CI1	–	[0.181,0.448]	[0.494,0.666]	[0.436,0.632]	[0.166,0.547]	[-0.156,0.748]	[-1.601,6.612]
	Bias2	–	0.017	0.082	0.068	-0.076	0.057	0.013
	SD2	–	0.023	0.033	0.037	**0.019**	0.038	0.045
	CI2	–	[0.129,0.355]	[0.060,0.315]	[0.104,0.335]	[0.059,0.406]	[0.057,0.348]	[-0.652,0.650]
	Time	0.586	0.557	0.796	0.869	0.876	0.892	0.905
	Bias1	-0.003	-0.014	-0.074	-0.043	0.020	-0.002	0.022
	SD1	**0.008**	0.017	0.013	0.013	0.018	0.014	0.042
5000	CI1	[0.407,0.526]	[0.414,0.568]	[0.392,0.722]	[0.095,0.452]	[0.540,0.732]	[0.363,1.098]	[0.172,1.011]
	Bias2	-0.003	4.773×10^{-4}	0.019	-0.008	0.030	0.053	0.007
	SD2	0.009	0.017	**0.008**	0.012	0.012	0.013	0.033
	CI2	[0.195,0.301]	[0.124,0.247]	[0.124,0.368]	[0.156,0.521]	[0.033,0.422]	[-34.62,33.00]	[-0.129,1.113]

Table 4: 95% simultaneous confidence intervals for 20 important features in the superconductivity dataset.

Methods							
	1	2	3	4	5	6	7
	[-0.112, -0.109]	[0.074, 0.077]	[0.236, 0.239]	[-0.342, -0.339]	[0.262, 0.264]	[-68.769, -68.766]	[-0.574, -0.571]
	8	9	10	11	12	13	14
$\breve{\theta}_{ns}$	[11.444, 11.446]	[-0.001, 0.001]	[-0.608, -0.605]	[-0.125, -0.122]	[0.547, 0.550]	[-4.742, -4.740]	[-0.352, -0.349]
	15	16	17	18	19	20	
	[0.003, 0.006]	[7.681, 7.684]	[-0.233, -0.230]	[-0.445, -0.442]	[-0.071, -0.068]	[-0.001, 0.002]	
	1	2	3	4	5	6	7
	[-0.089, -0.081]	[0.108, 0.116]	[0.142, 0.150]	[-0.177, -0.169]	[0.162, 0.170]	[-60.209, -60.201]	[-0.525, -0.518]
	8	9	10	11	12	13	14
$\breve{\theta}_{s}$	[20.530, 20.538]	[-0.004, 0.004]	[-0.611, -0.603]	[-0.125, -0.117]	[0.410, 0.418]	[-1.906, -1.898]	[-0.282, -0.274]
	15	16	17	18	19	20	
	[0.002, 0.005]	[3.111, 3.118]	[-0.232, -0.225]	[-0.075, -0.067]	[-0.196, -0.188]	[-0.001, 0.001]	
	1	2	3	4	5	6	7
	[-5.062, 4.421]	[-4.198, 5.285]	[-4.964, 4.519]	[-4.453, 5.030]	[-3.697, 5.786]	[19.132, 28.615]	[-4.380, 5.103]
	8	9	10	11	12	13	14
$\hat{\theta}_{ns,\text{zhang}}$	[-68.544, -59.061]	[-58.346, -48.864]	[-4.277, 5.206]	[-5.251, 4.232]	[-4.830, 4.653]	[56.744, 66.227]	[-6.755, 2.728]
	15	16	17	18	19	20	
	[-4.722, 4.761]	[34.088, 43.571]	[-5.395, 4.088]	[-3.677, 5.806]	[-3.579, 5.904]	[-4.753, 4.730]	
	1	2	3	4	5	6	7
	[-0.335, -0.306]	[0.507, 0.580]	[-0.251, -0.195]	[0.248, 0.328]	[0.996, 1.094]	[17.981, 29.767]	[0.295, 0.427]
	8	9	10	11	12	13	14
$\hat{\theta}_{s,\text{zhang}}$	[-68.239, -59.366]	[-59.304, -47.906]	[0.361, 0.567]	[-0.531, -0.488]	[-0.157, -0.019]	[57.198, 65.772]	[-2.144, -1.882]
	15	16	17	18	19	20	
	[0.018, 0.021]	[34.640, 43.018]	[-0.692, -0.616]	[0.970, 1.158]	[1.075, 1.249]	[-0.013, -0.011]	

the DVS estimator is the only method correctly identifying both z_1 and z_2 as significantly positive at the 0.05 level, consistent with the full-sample benchmark.

We further conduct simultaneous inference on the superconductivity dataset. As noted by Table 5 of Hamidieh (2018), 20 features (ordered by importance) are important to the critical temperature of the superconductors. We first fit a Lasso model on the full dataset and obtain the estimates: $(-0.084, 0.006, 0.227, -0.312, 0.253, -68.804, -0.568, 5.459, -0.284, -0.639, -0.157, 0.513, -4.201, -0.357, 0.005, 2.079, -0.216, -0.461, -0.134, 0.001)$. Then we conduct simultaneous inference on the 20 features. Denote the non-studentized and studentized versions of our estimators as $\check{\theta}_{ns}$ and $\check{\theta}_{s}$, respectively, and their counterparts in the baseline approach of Zhang & Cheng (2017) as $\hat{\theta}_{ns,\text{zhang}}$ and $\hat{\theta}_{s,\text{zhang}}$. The average running times across 100 replications are 4.784, 4.704, 18.438, and 18.464 seconds for $\check{\theta}_{ns}$, $\check{\theta}_{s}$, $\hat{\theta}_{ns,\text{zhang}}$, and $\hat{\theta}_{s,\text{zhang}}$, respectively, indicating that our methods are substantially faster. The inference results are presented in Table 4. Table 4 shows that the confidence intervals constructed by our proposed estimators closely align with the Lasso estimates, whereas those obtained by the methods of Zhang & Cheng (2017) deviate considerably. It also aligns with our previously inferential results, where z_1 corresponds to the 12th and z_2 corresponds to the 3rd feature. This discrepancy may stem from the strong multicollinearity among the covariates in the dataset-for instance, the dataset contains the geometric mean of density as an important feature and the mean of density in the remaining features that are highly correlated-which may undermine the inference accuracy of the method in Zhang & Cheng (2017). In contrast, our approach alleviates this issue to some extent through the use of decorrelated score functions, which mitigates the influence of nuisance parameters. Since $\check{\theta}_{s}$ yields feature-specific confidence intervals with adaptive lengths, we focus on its results. Except for the 9th and 20th features, all other 18 features are identified as significant. Among them, the 1st, 4th, 6th, 7th, 10th, 11th, 13th, 14th,

17th, 18th, and 19th features are significantly negative, while the 2nd, 3rd, 5th, 8th, 12th, 15th, 16th features are significantly positive. Hence, superconductors with small values of the negative features and large values of the positive features are more likely to exhibit high critical temperatures.

7 Conclusion

In this paper, we developed a unified and computationally efficient inference framework for both low- and high-dimensional parameters in large-scale GLMs. We propose two estimators—a DVS estimator and a multi-step estimator—that balance statistical efficiency and computational feasibility. The DVS estimator enables fast inference with a small subsample, while both estimators achieve statistical performance comparable to full-data approaches when $r/\sqrt{n} \to \infty$. For high-dimensional simultaneous inference, where conventional decorrelated score methods are inapplicable, we establish uniform Bahadur representations and asymptotic normality under general conditions. This enables valid simultaneous confidence intervals for large parameter sets and massive datasets, overcoming key statistical and computational challenges. Future work will focus on improving inference under measurement constraints and extending to prediction under covariate shift.

Disclosure statement

The authors declare that they have no conflicts of interest.

Acknowledgement

The authors are supported by Key technologies for coordination and interoperation of power distribution service resource under Grant No. 2021YFB2401300, and NSFC under Grant

No. 12571311 and 12326606.

Supplementary material

The supplementary material includes the additional simulation results for the Logistic regression model and the technical proofs. The code is made available at https://github.com/BoFuxjtu/BoFuxjtu_DVS.git.

References

Ai, M., Yu, J., Zhang, H. & Wang, H. (2021), 'Optimal subsampling algorithms for big data regressions', *Statistica Sinica* **31**(2), 749–772.

Bühlmann, P. & Van De Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Cai, L., Guo, X., Lian, H. & Zhu, L. (2025), 'Statistical inference for high-dimensional convoluted rank regression', *Journal of the American Statistical Association* pp. 1–25.

Cai, T., Liu, W. & Luo, X. (2011), 'A constrained ℓ 1 minimization approach to sparse precision matrix estimation', *Journal of the American Statistical Association* **106**(494), 594–607.

Cai, T. T., Guo, Z. & Ma, R. (2023), 'Statistical inference for high-dimensional generalized linear models with binary outcomes', *Journal of the American statistical association* **118**(542), 1319–1332.

Cheng, C., Feng, X., Huang, J. & Liu, X. (2022), 'Regularized projection score estimation of treatment effects in high-dimensional quantile regression', *Statistica Sinica* **32**(1), 23–41.

Chernozhukov, V., Chetverikov, D., Kato, K. & Koike, Y. (2023), 'High-dimensional data bootstrap', *Annual Review of Statistics and Its Application* **10**(1), 427–449.

Fan, Y., Liu, Y. & Zhu, L. (2021), 'Optimal subsampling for linear quantile regression models', *Canadian Journal of Statistics* **49**(4), 1039–1057.

Fang, E. X., Ning, Y. & Li, R. (2020), 'Test of significance for high-dimensional longitudinal data', *Annals of statistics* **48**(5), 2622.

Hamidieh, K. (2018), 'A data-driven statistical model for predicting the critical temperature of a superconductor', *Computational Materials Science* **154**, 346–354.

Han, D., Huang, J., Lin, Y. & Shen, G. (2022), 'Robust post-selection inference of high-dimensional mean regression with heavy-tailed asymmetric or heteroskedastic errors', *Journal of Econometrics* **230**(2), 416–431.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press.

Javanmard, A. & Montanari, A. (2014), 'Confidence intervals and hypothesis testing for high-dimensional regression', *The Journal of Machine Learning Research* **15**(1), 2869–2909.

Ma, P., Mahoney, M. & Yu, B. (2014), A statistical perspective on algorithmic leveraging, *in* 'International conference on machine learning', PMLR, pp. 91–99.

Ning, Y. & Liu, H. (2017), 'A general theory of hypothesis tests and confidence regions for sparse high dimensional models', *The Annals of Statistics* **45**(1), 158–195.

Reid, S., Tibshirani, R. & Friedman, J. (2016), 'A study of error variance estimation in lasso regression', *Statistica Sinica* pp. 35–67.

Shan, J. & Wang, L. (2024), 'Optimal poisson subsampling decorrelated score for high-dimensional generalized linear models', *Journal of Applied Statistics* pp. 1–25.

Shao, Y., Wang, L. & Lian, H. (2024), 'Optimal decorrelated score subsampling for high-dimensional generalized linear models under measurement constraints', *Journal of Computational and Graphical Statistics* pp. 1–10.

Su, M. & Wang, R. (2024), 'Subsampled one-step estimation for fast statistical inference', *arXiv preprint arXiv:2407.13446* .

van de Geer, S., Bühlmann, P., Ritov, Y. & Dezeure, R. (2014), 'On asymptotically optimal confidence regions and tests for high-dimensional models', *The Annals of Statistics* pp. 1166–1202.

Van der Vaart, A. W. (2000), *Asymptotic statistics*, Vol. 3, Cambridge university press.

Wainwright, M. J. (2019), *High-dimensional statistics: A non-asymptotic viewpoint*, Vol. 48, Cambridge university press.

Wang, H. & Ma, Y. (2021), 'Optimal subsampling for quantile regression in big data', *Biometrika* **108**(1), 99–112.

Wang, H., Yang, M. & Stufken, J. (2019), 'Information-based optimal subdata selection for big data linear regression', *Journal of the American Statistical Association* **114**(525), 393–405.

Wang, H., Zhu, R. & Ma, P. (2018), 'Optimal subsampling for large sample logistic regression', *Journal of the American Statistical Association.* **113**(522), 829–844.

Yan, Y., Wang, X. & Zhang, R. (2023), 'Confidence intervals and hypothesis testing for high-dimensional quantile regression: Convolution smoothing and debiasing', *Journal of Machine Learning Research* **24**(245), 1–49.

Yao, Y., Zou, J. & Wang, H. (2023), 'Optimal poisson subsampling for softmax regression', *Journal of Systems Science and Complexity* **36**(4), 1609–1625.

Yu, J., Wang, H., Ai, M. & Zhang, H. (2022), 'Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data', *Journal of the American Statistical Association* **117**(537), 265–276.

Zhang, C.-H. & Zhang, S. S. (2014), 'Confidence intervals for low dimensional parameters in high dimensional linear models', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **76**(1), 217–242.

Zhang, T., Ning, Y. & Ruppert, D. (2021), 'Optimal sampling for generalized linear models under measurement constraints', *Journal of Computational and Graphical Statistics* **30**(1), 106–114.

Zhang, X. & Cheng, G. (2017), 'Simultaneous inference for high-dimensional linear models', *Journal of the American Statistical Association* **112**(518), 757–768.