The Annals of Statistics
2017, Vol. 45, No. 1, 158–195
DOI: 10.1214/16-AOS1448

A GENERAL THEORY OF HYPOTHESIS TESTS AND CONFIDENCE REGIONS FOR SPARSE HIGH DIMENSIONAL MODELS[1]

BY YANG NING AND HAN LIU

Cornell University and Princeton University

We consider the problem of uncertainty assessment for low dimensional components in high dimensional models. Specifically, we propose a novel decorrelated score function to handle the impact of high dimensional nuisance parameters. We consider both hypothesis tests and confidence regions for generic penalized M-estimators. Unlike most existing inferential methods which are tailored for individual models, our method provides a general framework for high dimensional inference and is applicable to a wide variety of applications. In particular, we apply this general framework to study five illustrative examples: linear regression, logistic regression, Poisson regression, Gaussian graphical model and additive hazards model. For hypothesis testing, we develop general theorems to characterize the limiting distributions of the decorrelated score test statistic under both null hypothesis and local alternatives. These results provide asymptotic guarantees on the type I errors and local powers. For confidence region construction, we show that the decorrelated score function can be used to construct point estimators that are asymptotically normal and semiparametrically efficient. We further generalize this framework to handle the settings of misspecified models. Thorough numerical results are provided to back up the developed theory.

1. Introduction. Given n independent and identically distributed multivariate random variables $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$, assume that they are generated by a statistical model $\mathcal{P} = \{\mathbb{P}_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \Omega\}$, where $\boldsymbol{\beta}$ is a d-dimensional unknown vector of parameters with d much larger than the sample size n and Ω is the parameter space. In high dimensional settings, one general approach to estimate $\boldsymbol{\beta}$ is given by the penalized M-estimator

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \Omega} \ell(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}), \tag{1.1}$$

where $\ell(\boldsymbol{\beta})$ is a loss function (e.g., the negative log-likelihood) and $P_\lambda(\boldsymbol{\beta})$ is a penalty function with a tuning parameter λ. These penalty functions can be classified into two categories: convex penalties and nonconvex penalties. The most

Received September 2015; revised January 2016.

[1]Supported in part by NSF CAREER Award DMS-14-54377, NSF IIS1546482, NSF IIS1408910, NSF IIS1332109, NIH R01MH102339, NIH R01GM083084 and NIH R01HG06841.

MSC2010 subject classifications. Primary 62E20, 62F03; secondary 62F25.

Key words and phrases. High dimensional inference, nuisance parameter, sparsity, hypothesis test, confidence interval, score function, model misspecification.

popular convex penalty is the L_1 penalty, also known as the Lasso penalty [33], whose theoretical properties have been extensively studied in the literature. For instance, the statistical rate of the Lasso estimator is established by [5], and the variable selection consistency is studied by [24, 43]. The class of nonconvex penalties includes MCP [41], SCAD [8] and capped-L_1 penalty. Theoretical properties of these nonconvex estimators are investigated by [8, 37, 38, 41], among others.

Though significant progress has been made toward understanding the estimation theory of penalized M-estimators, it remains less explored on quantifying the uncertainty of the obtained results. To bridge this gap, this paper proposes a new device, named as decorrelated score function, to test statistical hypotheses and construct confidence regions for low dimensional components in high dimensional models. In particular, we partition the parameter $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma})$, where θ is a finite-dimensional parameter of interest and $\boldsymbol{\gamma}$ is a nuisance parameter. We aim to test the null hypothesis $H_0 : \theta^* = 0$, where θ^* is the true value of θ. The main challenge of this problem is the presence of high dimensional nuisance parameters, which invalidates the classical inferential theory. To handle this challenge, we apply a decorrelation operation on the high dimensional score function, so that the obtained decorrelated score function for θ becomes uncorrelated with the nuisance score functions. Unlike the classical score function, the decorrelated score can handle the impact of high dimensional nuisance parameters. With the decorrelated score function as a key ingredient, our framework is quite general. For example, it provides valid inference for penalized M-estimators with both convex and nonconvex penalties.

Theoretically, for hypothesis testing, we prove the limiting distributions of the decorrelated score test statistic under both the null hypothesis and local alternatives. These results characterize the asymptotic type I errors and local powers of the test. We further establish the uniform weak convergence of the test statistic, which implies honesty of the score test in terms of the type I errors and powers. For confidence region construction, we show that the decorrelated score function can be used to construct an estimator that is asymptotically normal and achieves the information lower bound, leading to an optimal confidence region. These theorems are established under a general framework. Thus, this paper provides a general theory for hypothesis tests and confidence regions in high dimensional models. We further illustrate the proposed methods by several commonly-used models including linear regressions, logistic regressions, Poisson regressions, Gaussian graphical models and additive hazards models.

To further broaden our framework, we provide valid inferential results under general misspecified models. In particular, we show that the proposed score test is robust to model misspecification, thus obtains valid inference on oracle parameters (i.e., least false parameters). This generalizes the classical misspecified model theory developed by [40]. Results on model misspecification is illustrated for linear regression.

We point out that the decorrelated score test can be viewed as a high dimensional extension of the Rao's score test in statistics [7] and the Lagrange multiplier test in econometrics [11]. In particular, the decorrelated score test is asymptotically equivalent to these two tests in low dimensional models. However, in high dimensions, the type I error can be controlled by the decorrelated score test rather than the two classical tests.

1.1. *Related works.* In the literature, there exist some recent works on uncertainty assessment for the regularized estimators in high dimensional models. In an early work, [15] showed that the limiting distribution of the Lasso estimator is nonnormal even in the low dimensional setting. Recently, in the high dimensional setting, several authors including [25, 26, 39] considered p-values based on the sample splitting technique or subsampling. Unlike these approaches, our approach avoids sample splitting and is potentially more powerful. For the L_1-regularized linear regression, [16, 22] considered the conditional inference given the event that some covariates have been selected, which is philosophically different from our unconditional inference. An instrumental variable approach together with a double selection procedure is proposed by [3]. From a different perspective, [13, 14, 34, 42] proposed a debiased method (named as LDPE) or desparsifying method to construct confidence intervals for linear or generalized linear models with the Lasso penalty. Unlike these works which are tailored for individual models, our decorrelated score method provides a general framework for high dimensional inference. In addition, our method can be used to infer the oracle parameter under model misspecification. With a class of nonconvex penalty functions, [8] established the oracle properties of the obtained estimator. However, such oracle results require minimal signal strength conditions which may not hold in many applications and the uncertainty of the estimation for small signals cannot be evaluated. In contrast, our method does not require variable selection consistency and provides valid inference for small signals. For hypothesis testing, [36] proposed a penalized score test. However, they focused on a null hypothesis depending on the tuning parameter and their test is biased for $H_0 : \theta^* = 0$. In addition, the validity of their test hinges on an irrepresentable type condition, which is not needed here.

1.2. *Organization of the paper.* In Section 2, we propose the decorrelated score function. In Section 3, we establish general results for hypothesis tests and confidence regions. In Section 4, we apply these general results to linear regression, logistic regression, Poisson regression, Gaussian graphical model, and additive hazards model. In Section 5, we consider misspecified models. Section 6 provides numerical results and Section 7 contains more discussions. We defer technical details to the supplementary materials [27].

1.3. *Notation.* The following notation is adopted throughout this paper. For $\mathbf{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, and $1 \le q \le \infty$, we define $\|\mathbf{v}\|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$, $\|\mathbf{v}\|_0 = |\operatorname{supp}(\mathbf{v})|$, where $\operatorname{supp}(\mathbf{v}) = \{j : v_j \neq 0\}$ and $|A|$ is the cardinality of a set A. Denote $\|\mathbf{v}\|_\infty = \max_{1\le i\le d} |v_i|$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$. For a matrix $\mathbf{M} = [M_{jk}]$, let $\|\mathbf{M}\|_{\max} = \max_{jk} |M_{jk}|$, $\|\mathbf{M}\|_1 = \sum_{jk} |M_{jk}|$, $\|\mathbf{M}\|_{\ell_\infty} = \max_j \sum_k |M_{jk}|$. If the matrix $\mathbf{M}$ is symmetric, then $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ are the minimal and maximal eigenvalues of $\mathbf{M}$. For $S \subseteq \{1, \ldots, d\}$, let $\mathbf{v}_S = \{v_j : j \in S\}$ and $\bar{S}$ be the complement of S. The gradient of a function $f(\boldsymbol{x})$ is denoted by $\nabla f(\boldsymbol{x})$. Let $\nabla_S f(\boldsymbol{x})$ denote the gradient of $f(\boldsymbol{x})$ with respect to $\boldsymbol{x}_S$. For two positive sequences a_n and b_n, we write $a_n \asymp b_n$ if $C \le a_n/b_n \le C'$ for some $C, C' > 0$. Similarly, we use $a_n \lesssim b_n$ to denote $a_n \le Cb_n$ for some constant $C > 0$. For a sequence of random variables X_n, we write $X_n \rightsquigarrow X$, if X_n converges weakly to X. Given $a, b \in \mathbb{R}$, let $a \vee b$ and $a \wedge b$ denote the maximum and minimum of a and b. A random variable X is sub-exponential if there exists some constant $K_1 > 0$ such that $\mathbb{P}(|X| > t) \le \exp(1 - t/K_1)$ for all $t \ge 0$. The sub-exponential norm of X is defined as $\|X\|_{\psi_1} = \sup_{p\ge 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}$. A random variable X is sub-Gaussian if there exists some constant $K_2 > 0$ such that $\mathbb{P}(|X| > t) \le \exp(1 - t^2/K_2^2)$ for all $t \ge 0$. The sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} = \sup_{p\ge 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}$. For simplicity, we use C, C', C'' to denote generic constants, whose values can change from line to line.

2. Score function for high dimensional models. We first introduce a general modeling framework and several examples. Then we review the classical Rao's score test for low dimensional models, and highlight the difficulty for directly applying it to models with high dimensional nuisance parameters. Finally, we propose a new device, named as decorrelated score function, to construct tests and confidence regions in high dimensions.

2.1. *A general statistical model framework and examples.* Let $\boldsymbol{U}$ denote a multivariate random variable following from a high dimensional statistical model $\mathcal{P} = \{\mathbb{P}_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \Omega\}$, where $\boldsymbol{\beta}$ is a d dimensional unknown parameter and Ω is the parameter space. To infer the true value of $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}^*$ (which is an interior point of Ω), we assume that there exist n independently identically distributed copies of $\boldsymbol{U}$, that is, $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$. In many statistical problems, the unknown parameter $\boldsymbol{\beta}$ can be partitioned as $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma}^T)^T$, where θ is a univariate parameter of interest and $\boldsymbol{\gamma}$ is a $(d-1)$ dimensional nuisance parameter. For simplicity and without loss of generality, we consider only univariate parameter θ. Extension to finite-dimensional parameter θ is straightforward and is deferred to the supplementary materials [27]. The statistical inferential problem can be formulated as testing the validity of the null hypothesis $H_0 : \theta^* = 0$ versus $H_1 : \theta^* \neq 0$ or constructing confidence intervals for θ^*. Instead of only relying on the likelihood based inference, our framework allows the inference based on a general

loss function. To ease presentation, we introduce the notation of the loss function $\ell(\theta, \boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\theta, \boldsymbol{\gamma})$, where $\ell_i(\theta, \boldsymbol{\gamma})$ is the loss function for the ith observation. For instance, $\ell(\theta, \boldsymbol{\gamma})$ could be the negative log-likelihood for generalized linear models. Given $\ell(\boldsymbol{\beta})$, we define $\mathbf{I} = \mathbb{E}_{\boldsymbol{\beta}}(\nabla^2 \ell(\boldsymbol{\beta}))$, and $I_{\theta|\boldsymbol{\gamma}} = I_{\theta\theta} - \mathbf{I}_{\theta\boldsymbol{\gamma}}\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\mathbf{I}_{\boldsymbol{\gamma}\theta}$, where $I_{\theta\theta}$, $\mathbf{I}_{\theta\boldsymbol{\gamma}}$, $\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ and $\mathbf{I}_{\boldsymbol{\gamma}\theta}$ are the corresponding partitions of $\mathbf{I}$. When $\ell(\theta, \boldsymbol{\gamma})$ is the negative log-likelihood, $\mathbf{I}$ and $I_{\theta|\boldsymbol{\gamma}}$ are called Fisher information and partial Fisher information, respectively. Similarly, denote $\mathbf{I}^* = \mathbb{E}_{\boldsymbol{\beta}^*}(\nabla^2 \ell(\boldsymbol{\beta}^*))$ and $\mathbf{w}^* = \mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{*-1}\mathbf{I}_{\boldsymbol{\gamma}\theta}^*$. Hereafter, we use $\mathbb{P}_{\boldsymbol{\beta}^*}(\cdot)$ and $\mathbb{E}_{\boldsymbol{\beta}^*}(\cdot)$ to denote the probability and expectation evaluated under the joint probability density of $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)$ indexed by the true parameter $\boldsymbol{\beta}^*$.

In this paper, we apply our general framework to study the high dimensional inferential problems for the following five models.

2.1.1. *Example* 1: *Linear regression models*. Consider the linear regression, $Y_i = \theta^* Z_i + \boldsymbol{\gamma}^{*T}\boldsymbol{X}_i + \varepsilon_i$, where $Z_i \in \mathbb{R}$, $\boldsymbol{X}_i \in \mathbb{R}^{d-1}$, and the error ε_i satisfies $\mathbb{E}(\varepsilon_i) = 0, \mathbb{E}(\varepsilon_i^2) = \sigma^2$ for $i = 1, \ldots, n$. Let $\boldsymbol{Q}_i = (Z_i, \boldsymbol{X}_i^T)^T$ denote the collection of all covariates for sample i and $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma}^T)^T$. The negative Gaussian quasi-log-likelihood (i.e., the least square loss) has the form $\ell(\boldsymbol{\beta}) = (2n)^{-1}\sum_{i=1}^{n}(Y_i - \boldsymbol{\beta}^T\boldsymbol{Q}_i)^2$. For the purpose of theoretical derivation, we assume that noise ε_i satisfies $\|\varepsilon_i\|_{\psi_2} \le C$ for some constant C, and $2\kappa \le \lambda_{\min}(\mathbb{E}(\boldsymbol{Q}_i^{\otimes 2})) \le \lambda_{\max}(\mathbb{E}(\boldsymbol{Q}_i^{\otimes 2})) \le 2/\kappa$ for some constant $\kappa > 0$. In addition, we assume $\boldsymbol{Q}_i$ is a sub-Gaussian vector.

2.1.2. *Example* 2: *Logistic regression models*. Assume that the binary outcome $Y_i \in \{0, 1\}$ given covariates $\boldsymbol{Q}_i = (Z_i, \boldsymbol{X}_i^T)^T \in \mathbb{R}^d$ follows from the logistic regression, whose negative log-likelihood is

$$\ell(\theta, \boldsymbol{\gamma}) = -\frac{1}{n}\sum_{i=1}^{n}\{Y_i(\theta Z_i + \boldsymbol{\gamma}^T\boldsymbol{X}_i) - \log[1 + \exp(\theta Z_i + \boldsymbol{\gamma}^T\boldsymbol{X}_i)]\}.$$

We derive the Fisher information matrix as $\mathbf{I}^* = \mathbb{E}_{\boldsymbol{\beta}^*}(\exp(\boldsymbol{\beta}^{*T}\boldsymbol{Q}_i)\boldsymbol{Q}_i^{\otimes 2}/(1 + \exp(\boldsymbol{\beta}^{*T}\boldsymbol{Q}_i))^2)$. We assume that $\lambda_{\min}(\mathbf{I}^*) \ge \kappa^2$ for some constant $\kappa > 0$, $\|\boldsymbol{Q}_i\|_\infty \le K$, and $|\mathbf{w}^{*T}\boldsymbol{X}_i| \le K$ for some constant $K > 0$. These are technical conditions imposed the design variables, which are weaker than those in [34]; see Section 4 for more detailed discussions.

2.1.3. *Example* 3: *Poisson regression models*. Assume that the outcome $Y_i \in \{0, 1, 2, \ldots\}$ given covariates $\boldsymbol{Q}_i = (Z_i, \boldsymbol{X}_i^T)^T \in \mathbb{R}^d$ follows from the Poisson regression, whose negative log-likelihood is

$$\ell(\theta, \boldsymbol{\gamma}) = -\frac{1}{n}\sum_{i=1}^{n}\{Y_i(\theta Z_i + \boldsymbol{\gamma}^T\boldsymbol{X}_i) - \exp(\theta Z_i + \boldsymbol{\gamma}^T\boldsymbol{X}_i)\}.$$

The Fisher information matrix is $\mathbf{I}^* = \mathbb{E}_{\boldsymbol{\beta}^*}(\exp(\boldsymbol{\beta}^{*T}\boldsymbol{Q}_i)\boldsymbol{Q}_i^{\otimes 2})$. Similarly, we assume that $\lambda_{\min}(\mathbf{I}^*) \geq \kappa^2$ for some constant $\kappa > 0$, $\|\boldsymbol{Q}_i\|_\infty \leq K$, $|\mathbf{w}^{*T}\boldsymbol{X}_i| \leq K$, and $|\boldsymbol{\beta}^{*T}\boldsymbol{Q}_i| \leq K$ for some constant $K > 0$.

2.1.4. *Example* 4: *Gaussian graphical models*. Consider the Gaussian graphical model $\boldsymbol{X} = (X_1, \ldots, X_d)^T \sim N(0, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{\Sigma}^*$ is the true covariance matrix. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n independent copies of $\boldsymbol{X}$. Since the conditional independence among $\boldsymbol{X}$ is characterized by the sparsity pattern of the precision matrix $\boldsymbol{\Theta}^* = \boldsymbol{\Sigma}^{*-1}$, inferring the component of the unknown precision matrix $\boldsymbol{\Theta}^*$ is often of interest. Denote the kth column of $\boldsymbol{\Theta}$ as $\boldsymbol{\beta}$. Without loss of generality, assume we are interested in $\theta = \Theta^*_{1k}$. Then we can similarly partition $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma}^T)^T$. We now consider the inference based on the following column-wise loss function proposed by [21]:

$$\ell(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \mathbf{e}_k^T\boldsymbol{\beta}, \tag{2.1}$$

where $\hat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T$ is the sample covariance matrix. Note that the objective function in (2.1) is a quadratic loss function instead of the log-likelihood for Gaussian graphical models. Assume that for some constant ν, it holds that $1/\nu \leq \lambda_{\min}(\boldsymbol{\Sigma}^*) \leq \lambda_{\max}(\boldsymbol{\Sigma}^*) \leq \nu$.

2.1.5. *Example* 5: *Additive hazards models*. The additive hazards model provides a convenient approach for the regression analysis of right censored survival data [17]. In the following, we first introduce some notation commonly used in the survival analysis. Let T be the time to event, R be the censoring time and $\boldsymbol{Q}(t) = (Z(t), \boldsymbol{X}^T(t))^T$ be the d dimensional time-dependent covariate vector at time t. Let $W = \min\{T, R\}$ and $\Delta = I(T \leq R)$ denote the observed survival time and censoring indicator. Assume that T and R are conditionally independent given all the covariates. The observed data consist of $(W_i, \Delta_i, \boldsymbol{Q}_i(\cdot))$, for $i = 1, \ldots, n$, which are n independent copies of $(W, \Delta, \boldsymbol{Q}(\cdot))$. Let $\lambda(t \mid \boldsymbol{Q}(t))$ be the conditional hazard function at time t given the covariates $\boldsymbol{Q}(t)$. The additive hazards model assumes that $\lambda(t \mid \boldsymbol{Q}(t)) = \lambda_0(t) + \boldsymbol{\beta}^{*T}\boldsymbol{Q}(t)$, where $\lambda_0(t)$ is an unknown baseline hazard function and $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is an unknown vector. We partition $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma}^T)^T$.

In the following, we introduce some counting process notation. Define $N_i(t) = I(W_i \leq t, \Delta_i = 1)$ to be the observed counting process, $Y_i(t) = I(W_i \geq t)$ to be the at-risk process. Let $\bar{\boldsymbol{Q}}(t) = \sum_{i=1}^n Y_i(t)\boldsymbol{Q}_i(t)/\sum_{i=1}^n Y_i(t)$ be the averaged covariates over the risk set. Denote

$$\begin{aligned} \boldsymbol{b} &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \{\boldsymbol{Q}_i(t) - \bar{\boldsymbol{Q}}(t)\}\, dN_i(t) \quad \text{and} \\ \mathbf{V} &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau Y_i(t)\{\boldsymbol{Q}_i(t) - \bar{\boldsymbol{Q}}(t)\}^{\otimes 2}\, dt, \end{aligned} \tag{2.2}$$

where τ is the end of study time. Under the additive hazards model, the inference on $\boldsymbol{\beta}^*$ is performed based on the following loss function:

$$\ell(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^T \mathbf{V}\boldsymbol{\beta} - \boldsymbol{b}^T\boldsymbol{\beta}. \tag{2.3}$$

Recall that for any vector $\mathbf{v}$, we write $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$, $\mathbf{v}^{\otimes 1} = \mathbf{v}$ and $\mathbf{v}^{\otimes 0} = 1$. Define $\mathbf{s}^{(k)}(t) = \mathbb{E}(Y_i(t)\boldsymbol{Q}_i(t)^{\otimes k})$ for $k = 0, 1, 2$,

$$\begin{aligned} \mathbf{V}^* &= \mathbb{E}\Big(\int_0^\tau Y_i(t)\{\boldsymbol{Q}_i(t) - \mathbf{e}(t)\}^{\otimes 2}\,dt\Big), \\ \mathbf{W}^* &= \mathbb{E}\Big(\int_0^\tau \{\boldsymbol{Q}_i(t) - \mathbf{e}(t)\}^{\otimes 2}\,dN_i(t)\Big), \end{aligned} \tag{2.4}$$

where $\mathbf{e}(t) = \mathbf{s}^{(1)}(t)/\mathbf{s}^{(0)}(t)$. We assume that

$$\int_0^\tau \lambda_0(t)\,dt < \infty, \qquad \mathbb{P}\big(Y_i(\tau) = 1\big) > 0, \qquad \big\|\boldsymbol{Q}_i(t)\big\|_\infty \le K \tag{2.5}$$

for some constant $K > 0$. The sample paths of $Q_{ij}(t)$ have uniformly bounded variation for $j = 1, \ldots, d$. The above assumptions are commonly used to study the estimation properties for high dimensional additive hazards model (e.g., condition 2 in [18]). In addition, we assume $|\mathbf{w}^{*T}\boldsymbol{X}_i(t)| \le K$ and $\lambda_{\min}(\mathbf{V}^*) \ge \kappa^2$ and $\lambda_{\min}(\mathbf{W}^*) \ge \kappa^2$ for some constants $K, \kappa > 0$. These assumptions are similar to those imposed for logistic and Poisson models.

2.2. *Challenges of score test in high dimensional models.* To illustrate the challenge of the classical score test, we assume $\ell(\boldsymbol{\beta})$ is the negative log-likelihood. When the dimension of the parameter vector is fixed and much smaller than the sample size, the classical Rao's score test for $H_0 : \theta^* = 0$ versus $H_1 : \theta^* \neq 0$ is based on the profile score function $\nabla_\theta \ell(0, \hat{\boldsymbol{\gamma}}(0))$, where $\hat{\boldsymbol{\gamma}}(\theta) = \operatorname{argmin}_{\boldsymbol{\gamma}} \ell(\theta, \boldsymbol{\gamma})$ is the constrained maximum likelihood estimator (MLE) of $\boldsymbol{\gamma}$ for fixed θ. Under the null hypothesis, it is well known that [7],

$$n^{1/2}\nabla_\theta \ell\big(0, \hat{\boldsymbol{\gamma}}(0)\big) \rightsquigarrow N\big(0, I^*_{\theta|\boldsymbol{\gamma}}\big). \tag{2.6}$$

The Rao's score test statistic is given by $S_c = n\{\nabla_\theta \ell(0, \hat{\boldsymbol{\gamma}}(0))\}^2 \hat{I}^{-1}_{\theta|\boldsymbol{\gamma}}$, where $\hat{I}_{\theta|\boldsymbol{\gamma}}$ is some consistent estimator of $I^*_{\theta|\boldsymbol{\gamma}}$. The score test is obtained by rejecting H_0, if and only if the value of S_c is large. In low dimensions, the score test is known to be asymptotically optimal against local alternatives.

In this paper, we are interested in inference for high dimensional models, in which d can be much larger than n. When the nuisance parameter $\boldsymbol{\gamma}$ is of high dimension, the constrained maximum likelihood estimator (MLE) $\hat{\boldsymbol{\gamma}}(\theta)$ is no longer consistent [28, 29]. Even though the corresponding maximum penalized likelihood estimator (MPLE) such as the Lasso estimator is consistent under certain conditions, it does not have a tractable limiting distribution even in the fixed dimensional

case [15]. To illustrate the infeasibility of the Rao's score test for high dimensional models, for any estimator $\tilde{\boldsymbol{\gamma}}$, consider the following Taylor expansion:

$$n^{1/2}\nabla_\theta \ell(0, \tilde{\boldsymbol{\gamma}}) = n^{1/2}\nabla_\theta \ell(0, \boldsymbol{\gamma}^*) + \Delta + \text{Rem}, \tag{2.7}$$

where Rem represents the remainder and $\Delta = n^{1/2}\nabla^2_{\theta\boldsymbol{\gamma}}\ell(0, \boldsymbol{\gamma}^*)(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)$. Although $n^{1/2}\nabla_\theta \ell(0, \boldsymbol{\gamma}^*)$ usually converges weakly to a normal distribution due to the central limit theorem, the asymptotic normality of $n^{1/2}\nabla_\theta \ell(0, \tilde{\boldsymbol{\gamma}})$ fails due to the nonignorable estimation bias and sparsity effect of $\tilde{\boldsymbol{\gamma}}$ in Δ and Rem. First, to ensure Rem is asymptotically ignorable, $\tilde{\boldsymbol{\gamma}}$ must have a fast convergence rate, which rules out the nonsparse MLE. Second, for those sparse estimators such as MPLE, following the arguments in [15], Δ may converge to some intractable limiting distribution, if it exists. Hence, the score function with $\boldsymbol{\gamma}$ estimated by either MLE or MPLE does not have a simple limiting distribution in the high dimensional setting.

2.3. *A decorrelated score method for high dimensional models.* As seen in the previous section, the standard score function with estimated nuisance parameters cannot be used for inference in high dimensional models. This motivates us to construct a new type of score function applicable in this more challenging regime. Assume that $\boldsymbol{\beta}$ can be estimated by the penalized M-estimator (1.1). In many applications, the penalty function $P_\lambda(\boldsymbol{\beta})$ in (1.1) is decomposable in the sense that $P_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^d p_\lambda(\beta_j)$. In our framework, we allow both convex and nonconvex penalties. For instance, $p_\lambda(\beta_j)$ can be taken as the L_1 penalty [33] $p_\lambda(\beta_j) = \lambda|\beta_j|$, the SCAD penalty [8],

$$p_\lambda(\beta_j) = \int_0^{|\beta_j|} \left\{ \lambda I(z \le \lambda) + \frac{(a\lambda - z)_+}{a-1} I(z > \lambda) \right\} dz,$$

for some $a > 2$ and the MCP penalty [41],

$$p_\lambda(\beta_j) = \lambda \int_0^{|\beta_j|} \left(1 - \frac{z}{\lambda b}\right)_+ dz,$$

for some $b > 0$. To infer the parameter θ, we propose a new type of score function given by

$$S(\theta, \boldsymbol{\gamma}) = \nabla_\theta \ell(\theta, \boldsymbol{\gamma}) - \mathbf{w}^T \nabla_{\boldsymbol{\gamma}} \ell(\theta, \boldsymbol{\gamma}) \qquad \text{with } \mathbf{w}^T = \mathbf{I}_{\theta\boldsymbol{\gamma}} \mathbf{I}^{-1}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}. \tag{2.8}$$

We call $S(\theta, \boldsymbol{\gamma})$ as the decorrelated score function for θ. This name comes from the fact that $S(\theta, \boldsymbol{\gamma})$ is uncorrelated with the nuisance score functions $\nabla_{\boldsymbol{\gamma}} \ell(\boldsymbol{\beta})$, that is, $\mathbb{E}_{\boldsymbol{\beta}}(S(\boldsymbol{\beta})^T \nabla_{\boldsymbol{\gamma}} \ell(\boldsymbol{\beta})) = 0$. We can show that the decorrelation operation is crucial to control the variability of higher order terms in the Taylor expansions, similar to I_2 in (2.7). Geometric insight of $S(\theta, \boldsymbol{\gamma})$ and its connection with the profile score function will be discussed in Section 2.5.

To construct a score test for $\theta^* = 0$ based on $S(\theta, \boldsymbol{\gamma})$, one needs to estimate the nuisance parameter $\boldsymbol{\gamma}$ and the unknown vector $\mathbf{w}$. The whole procedure is described in Algorithm 2.3. The output of this algorithm is the estimated decorrelated

Algorithm 1 Calculate the estimated decorrelated score function $\hat{S}(\theta, \hat{\boldsymbol{\gamma}})$

Require: Negative log-likelihood $\ell(\theta, \boldsymbol{\gamma})$, penalty function $P(\cdot)$ and tuning parameters λ and λ'.

(i): Calculate $\hat{\boldsymbol{\beta}}$ in (1.1) and partition $\hat{\boldsymbol{\beta}}$ as $\hat{\boldsymbol{\beta}} = (\hat{\theta}, \hat{\boldsymbol{\gamma}})$;

(ii): Estimate $\mathbf{w}$ by the Dantzig type estimator $\hat{\mathbf{w}}$,

$$\hat{\mathbf{w}} = \operatorname{argmin} \|\mathbf{w}\|_1 \qquad \text{s.t. } \big\|\nabla^2_{\theta\boldsymbol{\gamma}} \ell(\hat{\boldsymbol{\beta}}) - \mathbf{w}^T \nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \ell(\hat{\boldsymbol{\beta}})\big\|_\infty \le \lambda', \tag{2.9}$$

(iii): Calculate the estimated decorrelated score function

$$\hat{S}(\theta, \hat{\boldsymbol{\gamma}}) = \nabla_\theta \ell(\theta, \hat{\boldsymbol{\gamma}}) - \hat{\mathbf{w}}^T \nabla_{\boldsymbol{\gamma}} \ell(\theta, \hat{\boldsymbol{\gamma}}). \tag{2.10}$$

return $\hat{S}(\theta, \hat{\boldsymbol{\gamma}})$.

score function $\hat{S}(\theta, \hat{\boldsymbol{\gamma}})$, where $\boldsymbol{\gamma}$ is estimated by $\hat{\boldsymbol{\gamma}}$ in (1.1) and $\mathbf{w}$ is estimated by $\hat{\mathbf{w}}$ in (2.9). Hence, we can calculate the value of the estimated decorrelated score function $\hat{S}(\theta, \hat{\boldsymbol{\gamma}})$ at $\theta = 0$ to evaluate the validity of the null hypothesis. Note that the key step in the algorithm is to estimate $\mathbf{w}$, which essentially searches for the best sparse linear combination of the nuisance score functions $\nabla_{\boldsymbol{\gamma}} \ell(\theta, \boldsymbol{\gamma})$ to approximate the score function $\nabla_\theta \ell(\theta, \boldsymbol{\gamma})$, in a computationally efficient way. This can be also seen from the alternative formulation in (2.12). Since the nuisance score functions $\nabla_{\boldsymbol{\gamma}} \ell(\boldsymbol{\beta})$ and $\mathbf{w}$ have dimension $d-1$, we may need to impose some sparsity assumption on $\mathbf{w}$ to control the estimation error. The implication of this assumption and the comparison with the existing methods are further discussed in Section 4.1.

REMARK 1. Our method allows a wide range of penalty functions $P(\cdot)$ including nonconvex penalties, whereas most of the existing works mainly consider the Lasso penalty. For instance, the method in [34] is based on inverting Karush–Kuhn–Tucker (KKT) conditions of Lasso, which seems not directly applicable to nonconvex estimators obtained by a statistical optimization algorithm [37, 38].

REMARK 2. Based on our numerical experience, we find that a refitted Lasso estimator, which takes the support set of $\hat{\boldsymbol{\beta}}$ in (1.1) and on its support re-estimate $\boldsymbol{\beta}$ by the MLE, usually leads to better finite sample performance of the score test. The intuition is that this refitted estimator may have less bias, and is also less sensitive to the choice of tuning parameters in (1.1).

REMARK 3. In fact, our framework also allows a variety of procedures for estimating $\mathbf{w}$ in step (ii). Besides (2.9), some examples are given by the following penalized M-estimators $\tilde{\mathbf{w}}$ and $\bar{\mathbf{w}}$:

$$\tilde{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{2n} \sum_{i=1}^{n} \{\mathbf{w}^T \nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \ell_i(\hat{\boldsymbol{\beta}}) \mathbf{w} - 2\mathbf{w}^T \nabla_{\boldsymbol{\gamma}\theta} \ell_i(\hat{\boldsymbol{\beta}})\} + Q_{\lambda'}(\mathbf{w}), \tag{2.11}$$

$$\bar{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{2n} \sum_{i=1}^{n} \{\nabla_\theta \ell_i(\hat{\boldsymbol{\beta}}) - \mathbf{w}^T \nabla_\gamma \ell_i(\hat{\boldsymbol{\beta}})\}^2 + Q_{\lambda'}(\mathbf{w}), \tag{2.12}$$

where $Q(\cdot)$ is a general penalty function.

We note that a similar refitted estimator of $\mathbf{w}$ may also improve the finite sample performance of the test.

REMARK 4. For notational simplicity, we assume that $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma})$ is estimated under the full model and $\hat{\boldsymbol{\gamma}}$ is the corresponding component of $\hat{\boldsymbol{\beta}}$. It is well known that the classical Rao's score test requires the estimator under the null hypothesis. Indeed, we can use the similar strategy to estimate the nuisance parameters. Specifically, instead of plugging $\hat{\boldsymbol{\gamma}}$ into the decorrelated score function, we can estimate $\boldsymbol{\gamma}$ by $\hat{\boldsymbol{\gamma}}_0 = \operatorname{argmin}_{\boldsymbol{\gamma}} \{\ell(0, \boldsymbol{\gamma}) + P_\lambda(\boldsymbol{\gamma})\}$, and replace $\hat{\boldsymbol{\beta}}$ in (2.9) with $(0, \hat{\boldsymbol{\gamma}}_0)$. As shown in Theorem 3.1, under mild conditions, no matter which estimator ($\hat{\boldsymbol{\gamma}}$ or $\hat{\boldsymbol{\gamma}}_0$) is plugged in, the estimated decorrelated score function is asymptotically equivalent to $S(\theta^*, \boldsymbol{\gamma}^*)$.

In this section, to ease presentation, we confine our attention to the univariate parameter of interest and the likelihood framework. Similar to (2.8), we can define the decorrelated score function for a multi-dimensional parameter of interest; see the supplementary materials [27] for the extension. In addition, for the general loss function $\ell(\boldsymbol{\beta})$, we can still define the estimated decorrelated score function $\hat{S}(\theta, \hat{\boldsymbol{\gamma}})$ following the procedures in Algorithm 2.3. In Section 3, we show that the general theory is applicable even if $\ell(\boldsymbol{\beta})$ is a loss function other than the negative log-likelihood.

2.4. *One-step estimator and confidence regions.* Though the decorrelated score method is target to hypothesis testing, in what follows, we consider how to use the decorrelated score function to construct a valid confidence interval for the parameter of interest θ. This is based on the key observation that the estimated decorrelated score function $\hat{S}(\theta, \hat{\boldsymbol{\gamma}})$ in (2.10) can be regarded as an approximately unbiased estimating function for θ [10]. Thus, one general strategy to define an estimator through the estimating function $\hat{S}(\theta, \hat{\boldsymbol{\gamma}})$ is to solve this equation, that is, $\hat{S}(\theta, \hat{\boldsymbol{\gamma}}) = 0$. However, as commented in Chapter 5 of [35], this Z-estimation approach may have several drawbacks. For instance, the estimating function $\hat{S}(\theta, \hat{\boldsymbol{\gamma}})$ may have multiple roots, such that the estimator becomes ill-posed. To overcome these issues, similar to [4], we consider the one-step method, which solves a first-order approximation of $\hat{S}(\theta, \hat{\boldsymbol{\gamma}}) = 0$. Given the penalized M-estimator $\hat{\theta}$, a one-step estimator $\tilde{\theta}$ is the solution to $\hat{S}(\hat{\boldsymbol{\beta}}) + \hat{I}_{\theta|\boldsymbol{\gamma}}(\theta - \hat{\theta}) = 0$, defined as

$$\tilde{\theta} = \hat{\theta} - \hat{S}(\hat{\boldsymbol{\beta}})/\hat{I}_{\theta|\boldsymbol{\gamma}} \qquad \text{where } \hat{I}_{\theta|\boldsymbol{\gamma}} = \nabla^2_{\theta\theta} \ell(\hat{\boldsymbol{\beta}}) - \hat{\mathbf{w}}^T \nabla^2_{\boldsymbol{\gamma}\theta} \ell(\hat{\boldsymbol{\beta}}). \tag{2.13}$$

In Section 3, we show that, under mild conditions on $\ell(\boldsymbol{\beta})$ and the penalized M-estimator, the one-step estimator $\tilde{\theta}$ is asymptotically normal with mean θ^* and

is semiparametrically efficient. Based on the asymptotic normality of $\tilde{\theta}$, we can easily construct the optimal confidence interval for θ^*. A similar method can be used to construct the optimal confidence region for a multi-dimensional parameter of interest; see the supplementary materials [27].

2.5. *Geometric interpretation and further discussion.* Given the random variable $\boldsymbol{U}$, consider $S_U(\theta, \boldsymbol{\gamma}) = \nabla_\theta \log f(\boldsymbol{U}; \boldsymbol{\beta}) - \mathbf{w}^T \nabla_{\boldsymbol{\gamma}} \log f(\boldsymbol{U}; \boldsymbol{\beta})$, where $f(\boldsymbol{U}, \boldsymbol{\beta})$ is the probability density of $\boldsymbol{U}$ under the model $\mathbb{P}_{\boldsymbol{\beta}}$, and the decorrelated score function is $S(\theta, \boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^n S_{U_i}(\theta, \boldsymbol{\gamma})$. For simplicity, we focus on the geometric interpretation for $S_U(\theta, \boldsymbol{\gamma})$. For any $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma}^T)^T$, the linear space spanned by the score functions can be expressed by $T = \{a_{\boldsymbol{\beta}} \nabla_\theta \log f(\boldsymbol{U}; \boldsymbol{\beta}) + \mathbf{b}_{\boldsymbol{\beta}}^T \nabla_{\boldsymbol{\gamma}} \log f(\boldsymbol{U}; \boldsymbol{\beta})\}$, where $a_{\boldsymbol{\beta}}$ is a nonrandom scalar and $\mathbf{b}_{\boldsymbol{\beta}}$ is a nonrandom $(d-1)$ dimensional vector. As suggested by the notation, $a_{\boldsymbol{\beta}}$ and $\mathbf{b}_{\boldsymbol{\beta}}$ can depend on $\boldsymbol{\beta}$. It is shown by [31] that the space T is a Hilbert space with inner product given by $\langle g_1(\boldsymbol{U}; \boldsymbol{\beta}), g_2(\boldsymbol{U}; \boldsymbol{\beta}) \rangle = \mathbb{E}_{\boldsymbol{\beta}}(g_1(\boldsymbol{U}; \boldsymbol{\beta}) g_2(\boldsymbol{U}; \boldsymbol{\beta}))$, for any $g_1(\boldsymbol{U}; \boldsymbol{\beta}), g_2(\boldsymbol{U}; \boldsymbol{\beta}) \in T$. Consider the linear space spanned by the nuisance score functions $T_N = \{\mathbf{b}_{\boldsymbol{\beta}}^T \nabla_{\boldsymbol{\gamma}} \log f(\boldsymbol{U}; \boldsymbol{\beta})\}$, where $\mathbf{b}_{\boldsymbol{\beta}}$ is a nonrandom $(d-1)$ dimensional vector, and its orthogonal complement $U_N := T_N^{\perp} = \{g(\boldsymbol{U}; \boldsymbol{\beta}) \in T, \langle g(\boldsymbol{U}; \boldsymbol{\beta}), s(\boldsymbol{U}; \boldsymbol{\beta}) \rangle = 0, \forall s(\boldsymbol{U}; \boldsymbol{\beta}) \in T_N\}$. Since $\nabla_\theta \log f(\boldsymbol{U}; \boldsymbol{\beta}) \in T$ and U_N is a closed space, the projection of $\nabla_\theta \log f(\boldsymbol{U}; \boldsymbol{\beta})$ to U_N is well defined and identical to the decorrelated score function $S_U(\theta, \boldsymbol{\gamma})$. Note that, for the estimation purpose, we assume that the projection of $\nabla_\theta \log f(\boldsymbol{U}; \boldsymbol{\beta})$ to T_N is identical to the projection of $\nabla_\theta \log f(\boldsymbol{U}; \boldsymbol{\beta})$ to a low dimensional subspace $T_N(S)$ of T_N. Here, $T_N(S)$ is defined as $T_N(S) = \{\mathbf{c}_{\boldsymbol{\beta}}^T \nabla_{\boldsymbol{\gamma}_S} \log f(\boldsymbol{U}; \boldsymbol{\beta})\}$, where S is a subset of $\{1, \ldots, d-1\}$ with $|S| \ll n$, and $\mathbf{c}_{\boldsymbol{\beta}}$ is a nonrandom $|S|$ dimensional vector. In other words, $T_N(S)$ is a linear space spanned by the components of nuisance score functions corresponding to $\boldsymbol{\gamma}_S$. For the clarification purpose, we illustrate these geometric structures in Figure 1.

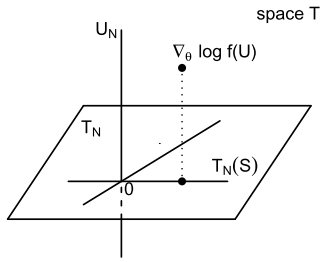

FIG. 1. *Geometric illustration. The entire space is T. Each point in the space represents a function. The plane is the nuisance space T_N and the perpendicular line is its orthogonal space U_N. Within the space T_N, there exists a low dimensional subspace $T_N(S)$, such that the projection of $\nabla_\theta \log f(\boldsymbol{U})$ to T_N is in $T_N(S)$. The 0 in T_N is the 0 function.*

Recall that in semiparametric models, the efficient score function is defined as the projection of $\nabla_\theta \log f(\boldsymbol{U}; \boldsymbol{\beta})$ to the orthogonal complement of the tangent set for nuisance parameters; see Section 25.4 in [35] for a rigorous definition. Note that our decorrelated score function has a similar projection interpretation. However, the main difference is as follows. In high dimensional settings, to ensure the projection is estimable, we introduce a more refined low dimensional structure $T_N(S)$ within the high dimensional Hilbert space T_N. This makes our decorrelated score function different from the efficient score function in semiparametric literature.

Finally, we comment on the connection between the decorrelated score and profile score function. Recall that the profile score function is given by $\nabla_\theta \ell(\theta, \hat{\boldsymbol{\gamma}}_\theta)$, where $\ell(\theta, \hat{\boldsymbol{\gamma}}_\theta)$ is the negative profile log-likelihood, and $\hat{\boldsymbol{\gamma}}_\theta$ is the constrained maximum likelihood estimator, that is, $\hat{\boldsymbol{\gamma}}_\theta = \operatorname{argmin}_{\boldsymbol{\gamma}} \ell(\theta, \boldsymbol{\gamma})$. It is easily seen that the decorrelated score function $S(\theta, \hat{\boldsymbol{\gamma}}_\theta)$ with $\boldsymbol{\gamma}$ estimated by $\hat{\boldsymbol{\gamma}}_\theta$ is identical to the profile score function, due to $\nabla_{\boldsymbol{\gamma}} \ell(\theta, \hat{\boldsymbol{\gamma}}_\theta) = 0$. Hence, in low dimensional problems, the profile score function implicitly performs decorrelation. From this perspective, we view our decorrelated score function as a natural high dimensional extension of the profile score function.

3. A general theory for tests and confidence regions. The goal of this section is to develop a general theory for the score test and the confidence regions. We first establish the pointwise weak convergence of score test under the null hypothesis. Then we establish the asymptotic normality of the one-step estimator. In Appendix A, we further strengthen the pointwise results to the uniform convergence of score test under the null and alternative hypotheses.

In this section, we focus on correctly specified models and allow $\ell(\boldsymbol{\beta})$ to be a general loss function.

3.1. *Weak convergence of score test under null hypothesis.* To study the properties of the decorrelated score test under the null hypothesis, we impose some technical conditions.

These conditions can be classified into four categories: (i) Consistency conditions for initial parameter estimation (Assumption 3.1); (ii) Concentration of the gradient and Hessian matrix (Assumption 3.2); (iii) Local smoothness on the loss function (Assumption 3.3); (iv) Central limit theorem for the score function (Assumption 3.4).

ASSUMPTION 3.1 (Consistency conditions for initial parameter estimation). Recall that $\mathbf{w}^* = \mathbf{I}^{*-1}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \mathbf{I}^*_{\boldsymbol{\gamma}\theta}$. For some sequences $\eta_1(n)$ and $\eta_2(n)$ converging to 0, as $n \to \infty$, it holds

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\beta}^*}\big(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \lesssim \eta_1(n)\big) = 1 \quad \text{and} \quad \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\beta}^*}\big(\|\hat{\mathbf{w}} - \mathbf{w}^*\|_1 \lesssim \eta_2(n)\big) = 1.$$

The estimation error bounds for $\boldsymbol{\beta}^*$ in terms of the L_1 norm have been thoroughly studied for a variety of estimators, including the Lasso estimator, nonconvex estimator and Dantzig selector; see [5, 6, 37, 38], among many others. The estimation error bound for $\mathbf{w}^*$ can be derived by the similar method. But the key difference is that $\hat{\mathbf{w}}$ may depend on $\hat{\boldsymbol{\beta}}$, which causes some extra technical difficulty in bounding $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_1$. The following lemma shows that Assumption 3.1 holds for all five models in Section 2.

LEMMA 3.1. *In linear models* (*Example* 1), *logistic models* (*Example* 2), *Poisson models* (*Example* 3), *Gaussian graphical models* (*Example* 4) *and additive hazards models* (*Example* 5), *we have*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s^*\sqrt{\log d/n}), \qquad \|\hat{\mathbf{w}} - \mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s'\sqrt{\log d/n}),$$

where $\hat{\boldsymbol{\beta}}$ *and* $\hat{\mathbf{w}}$ *are defined in Section* 4, $s^* = \|\boldsymbol{\gamma}^*\|_0$ *and* $s' = \|\mathbf{w}^*\|_0$.

PROOF. A detailed proof is shown in the supplementary materials [27]. □

ASSUMPTION 3.2 (Concentration of the gradient and Hessian). Let $\mathbf{v}^* = (1, -\mathbf{w}^{*T})^T$. We assume $\|\nabla \ell(\boldsymbol{\beta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log d/n})$ and

$$\|\mathbf{v}^{*T}\nabla^2 \ell(\boldsymbol{\beta}^*) - \mathbb{E}_{\boldsymbol{\beta}^*}(\mathbf{v}^{*T}\nabla^2 \ell(\boldsymbol{\beta}^*))\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log d/n}).$$

Assumption 3.2 holds for i.i.d. data provided there exists a constant K such that for all $1 \le j \le d$,

$$\max\{\|\nabla_j \ell_i(\boldsymbol{\beta}^*)\|_{\psi_1}, \|[\mathbf{v}^{*T}\nabla^2 \ell_i(\boldsymbol{\beta}^*)]_j\|_{\psi_1}\} \le K.$$

Thus, this assumption essentially imposes the sub-exponential conditions for some random variables related to the gradient and Hessian matrix. In low dimensional settings, it is often sufficient to prove asymptotic normality of estimators under some finite moment assumptions on the gradient and Hessian matrix. In contrast, to apply sharper exponential inequalities (i.e., Bernstein inequality) to bound the rate of convergence in high dimensions as in Assumption 3.2, we may need stronger sub-exponential type conditions. The following lemma shows that Assumption 3.2 holds for all five models considered in Section 2.

LEMMA 3.2. *Assumption* 3.2 *holds for linear models* (*Example* 1), *logistic models* (*Example* 2), *Poisson models* (*Example* 3), *Gaussian graphical models* (*Example* 4) *and additive hazards models* (*Example* 5).

PROOF. A detailed proof is shown in the supplementary materials [27]. □

ASSUMPTION 3.3 (Local smoothness conditions on the loss function). Let $\hat{\boldsymbol{\beta}}_0 = (0, \hat{\boldsymbol{\gamma}}^T)^T$, $\hat{\mathbf{v}} = (1, -\hat{\mathbf{w}}^T)^T$ and $\mathbf{v}^* = (1, -\mathbf{w}^{*T})^T$. Assume that for both $\check{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0$ and $\check{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$, it holds that

$$\mathbf{v}^{*T}\{\nabla \ell(\check{\boldsymbol{\beta}}) - \nabla \ell(\boldsymbol{\beta}^*) - \nabla^2 \ell(\boldsymbol{\beta}^*)(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\} = o_{\mathbb{P}}(n^{-1/2}), \tag{3.1}$$

and $(\hat{\mathbf{v}} - \mathbf{v}^*)^T(\nabla \ell(\check{\boldsymbol{\beta}}) - \nabla \ell(\boldsymbol{\beta}^*)) = o_{\mathbb{P}}(n^{-1/2})$.

In this assumption, we implicitly assume that the loss function is second- order differentiable. This assumption essentially quantifies the smoothness of $\ell(\boldsymbol{\beta})$ around a small neighborhood of $\boldsymbol{\beta}^*$. It is easy to check that (3.1) always holds if $\ell(\boldsymbol{\beta})$ is a quadratic function of $\boldsymbol{\beta}$. For nonquadratic loss functions, for example, the negative log-likelihood in Poisson regression, applying the Taylor expansion, we find that (3.1) reduces to

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{v}^{*T}\boldsymbol{Q}_i \exp(\boldsymbol{\beta}^{*T}\boldsymbol{Q}_i) R(\check{\Delta}^T \boldsymbol{Q}_i)\{\check{\Delta}^T \boldsymbol{Q}_i\}^2,$$

where $\check{\Delta} = \check{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ and $R(\cdot)$ is the remainder. Then it can be further bounded from above by the prediction error $n^{-1}\sum_{i=1}^n (\check{\Delta}^T \boldsymbol{Q}_i)^2$ up to a constant. Similarly, we can show that the last part of Assumption 3.3 is also related to both prediction errors $n^{-1}\sum_{i=1}^n \{(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \boldsymbol{Q}_i\}^2$ and $n^{-1}\sum_{i=1}^n \{(\hat{\mathbf{v}} - \mathbf{v}^*)^T \boldsymbol{Q}_i\}^2$. The following lemma shows that Assumption 3.3 holds for all five models considered in Section 2.

LEMMA 3.3. *Assumption 3.3 holds for linear models* (*Example* 1), *logistic models* (*Example* 2), *Poisson models* (*Example* 3), *Gaussian graphical models* (*Example* 4) *and additive hazards models* (*Example* 5).

PROOF. A detailed proof is shown in the supplementary materials [27]. □

ASSUMPTION 3.4 (Central limit theorem for the score function). Let $\boldsymbol{\Sigma}^* = \lim_{n\to\infty} \mathrm{Var}_{\boldsymbol{\beta}^*}(n^{1/2}\nabla \ell(\boldsymbol{\beta}^*))$. It holds that

$$\sqrt{n}\mathbf{v}^{*T}\nabla \ell(\boldsymbol{\beta}^*)/\sqrt{\sigma_s^*} \rightsquigarrow N(0,1) \qquad \text{where } \sigma_s^* = \mathbf{v}^{*T}\boldsymbol{\Sigma}^*\mathbf{v}^*$$

and $\sigma_s^* \geq C$ for some constant $C > 0$.

Assumption 3.4 is the central limit theorem for a linear combination of the score functions, which can be obtained by verifying the Lindeberg's condition. When $\ell(\boldsymbol{\beta})$ is the negative log-likelihood function, we have the information identity or the second Bartlett identity: $\boldsymbol{\Sigma}^* = \mathbf{I}^*$ [19]. However, for general loss functions, $\boldsymbol{\Sigma}^*$ and $\mathbf{I}^*$ could be different. The following lemma shows that Assumption 3.4 holds for all five models considered in Section 2.

LEMMA 3.4. *Assumption 3.4 holds for linear models* (*Example* 1), *logistic models* (*Example* 2), *Poisson models* (*Example* 3), *Gaussian graphical models* (*Example* 4) *and additive hazards models* (*Example* 5).

PROOF. A detailed proof is shown in the supplementary materials [27]. □

Given the decorrelated score function, we define the test statistic as

$$\hat{U}_n = n^{1/2}\hat{S}(0, \hat{\boldsymbol{\gamma}})/\sqrt{\hat{\sigma}_s}, \tag{3.2}$$

where $\hat{\sigma}_s$ is a consistent estimator of σ_s^*. The following theorem shows that the decorrelated score function $n^{1/2}\hat{S}(0, \hat{\boldsymbol{\gamma}})$ and the associated test statistic $\hat{U}_n$ are asymptotically normal.

THEOREM 3.1. *Under Assumptions 3.1–3.4, if* $(\eta_1(n)+\eta_2(n))\sqrt{\log d} = o(1)$, *then we have*

$$n^{1/2}\hat{S}(0, \hat{\boldsymbol{\gamma}})\sigma_s^{*-1/2} \rightsquigarrow N(0, 1), \tag{3.3}$$

and for any $t \in \mathbb{R}$,

$$\lim_{n\to\infty}\left|\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \Phi(t)\right| = 0. \tag{3.4}$$

PROOF. Recall that $\hat{\boldsymbol{\beta}}_0 = (0, \hat{\boldsymbol{\gamma}}^T)^T$, $\hat{\mathbf{v}} = (1, -\hat{\mathbf{w}}^T)^T$ and $\mathbf{v}^* = (1, -\mathbf{w}^{*T})^T$. By the definition of $\hat{S}(0, \hat{\boldsymbol{\gamma}})$, we have the following decomposition:

$$\begin{aligned} &n^{1/2}\left|\hat{S}(\hat{\boldsymbol{\beta}}_0) - S(\boldsymbol{\beta}^*)\right| \\ &\quad = n^{1/2}\left|\hat{\mathbf{v}}^T\nabla\ell(\hat{\boldsymbol{\beta}}_0) - \mathbf{v}^{*T}\nabla\ell(\boldsymbol{\beta}^*)\right| \\ &\quad \le n^{1/2}\left|\mathbf{v}^{*T}\{\nabla\ell(\hat{\boldsymbol{\beta}}_0) - \nabla\ell(\boldsymbol{\beta}^*)\}\right| + n^{1/2}\left|(\hat{\mathbf{v}} - \mathbf{v}^*)^T\nabla\ell(\hat{\boldsymbol{\beta}}_0)\right| \\ &\quad := I_1 + I_2. \end{aligned} \tag{3.5}$$

By Assumption 3.3, we can show that

$$\begin{aligned} |I_1| &\le n^{1/2}\left|\mathbf{v}^{*T}\nabla^2\ell(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*)\right| + o_{\mathbb{P}}(1) \\ &\le n^{1/2}\left\|\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\right\|_1\left\|\nabla^2_{\theta\boldsymbol{\gamma}}\ell(\boldsymbol{\beta}^*) - \mathbf{w}^{*T}\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\ell(\boldsymbol{\beta}^*)\right\|_\infty + o_{\mathbb{P}}(1). \end{aligned}$$

By Assumptions 3.1 and 3.2, we have $|I_1| \lesssim \eta_1(n)\sqrt{\log d} + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$. For I_2, Assumption 3.3 yields

$$|I_2| \le n^{1/2}\left|(\hat{\mathbf{v}} - \mathbf{v}^*)^T\nabla\ell(\boldsymbol{\beta}^*)\right| + o_{\mathbb{P}}(1) \le n^{1/2}\left\|\hat{\mathbf{v}} - \mathbf{v}^*\right\|_1\left\|\nabla\ell(\boldsymbol{\beta}^*)\right\|_\infty + o_{\mathbb{P}}(1).$$

By Assumptions 3.1 and 3.2, we have $|I_2| \lesssim \eta_2(n)\sqrt{\log d} + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$. Together with (3.5), the bounds for I_1 and I_2 imply $n^{1/2}|\hat{S}(\hat{\boldsymbol{\beta}}_0) - S(\boldsymbol{\beta}^*)| = o_{\mathbb{P}}(1)$. In

addition, $n^{1/2}S(\boldsymbol{\beta}^*)\sigma_s^{*-1/2} \rightsquigarrow N(0,1)$ by Assumption 3.4. By $\sigma_s^* \geq C$ in Assumption 3.4, we obtain that

$$n^{1/2}\big|\hat{S}(0,\hat{\boldsymbol{\gamma}})\sigma_s^{*-1/2} - S\big(0,\boldsymbol{\gamma}^*\big)\sigma_s^{*-1/2}\big| = o_{\mathbb{P}}(1).$$

This completes the proof by applying the Slutsky's theorem. □

REMARK 5. Based on the test statistic $\hat{U}_n$, the score test with significance level α, for the null hypothesis $H_0 : \theta^* = 0$ versus the two-sided alternative $H_1 : \theta^* \neq 0$ is given by

$$T_n = \begin{cases} 0 & \text{if } |\hat{U}_n| \leq \Phi^{-1}(1-\alpha/2), \\ 1 & \text{if } |\hat{U}_n| > \Phi^{-1}(1-\alpha/2), \end{cases} \tag{3.6}$$

where $\Phi(\cdot)$ is the cdf of a standard normal distribution. Given the value of T_n, we reject the null hypothesis if and only if $T_n = 1$. The type I error of T_n, that is the probability of rejecting H_0 (i.e., $T_n = 1$) when H_0 is true, can be controlled by α asymptotically. This is $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\beta}^*}(T_n = 1) = \alpha$.

3.2. *Theoretical results for optimal confidence regions.* The following theorem shows that the one-step estimator $\tilde{\theta}$ is asymptotically normal.

THEOREM 3.2. *Under Assumptions 3.1–3.4, if* $(\eta_1(n)+\eta_2(n))\sqrt{\log d} = o(1)$, $\hat{I}_{\theta|\boldsymbol{\gamma}}$ *is consistent for* $I^*_{\theta|\boldsymbol{\gamma}}$ *and* $I^*_{\theta|\boldsymbol{\gamma}} \geq C$ *for some constant* $C > 0$, *then*

$$n^{1/2}\big(\tilde{\theta} - \theta^*\big)I^*_{\theta|\boldsymbol{\gamma}}/\sigma_s^{*1/2} = -S\big(\boldsymbol{\beta}^*\big)/\sigma_s^{*1/2} + o_{\mathbb{P}}(1) \rightsquigarrow N(0,1), \tag{3.7}$$

where σ_s^* *is defined in Assumption 3.4.*

PROOF. A detailed proof is provided in Appendix B.1. □

Since $\hat{I}_{\theta|\boldsymbol{\gamma}}$ is consistent for $I^*_{\theta|\boldsymbol{\gamma}}$ and $\hat{\sigma}_s$ is consistent for σ_s^*, we can construct a $(1-\alpha)\times 100\%$ confidence interval for θ^* as $[\tilde{\theta} - n^{-1/2}\Phi^{-1}(1-\alpha/2)\hat{I}^{-1}_{\theta|\boldsymbol{\gamma}}\hat{\sigma}_s^{1/2}, \tilde{\theta} + n^{-1/2}\Phi^{-1}(1-\alpha/2)\hat{I}^{-1}_{\theta|\boldsymbol{\gamma}}\hat{\sigma}_s^{1/2}]$. In addition, if $\ell(\boldsymbol{\beta})$ is the negative log-likelihood, then $I^*_{\theta|\boldsymbol{\gamma}} = \sigma_s^*$ and, therefore, (3.7) implies that the asymptotic variance of $\tilde{\theta}$ is identical to the inverse of the partial information matrix $I^*_{\theta|\boldsymbol{\gamma}}$, which is also known as the efficient information in the presence of nuisance parameters; see Chapter 25 of [35]. This implies that the one-step estimator $\tilde{\theta}$ is semiparametrically efficient. A similar criterion on optimality under the linear model is considered by [34].

4. Examples. Given Lemmas 3.1–3.4, in this section, we summarize the inferential results for linear regression, logistic regression, Poisson regression, Gaussian graphical model and additive hazards model.

4.1. *Example* 1: *Linear regression model.* Recall that the linear regression model is given by $Y_i = \theta^* Z_i + \boldsymbol{\gamma}^{*T} \boldsymbol{X}_i + \varepsilon_i$, where $Z_i \in \mathbb{R}$, $\boldsymbol{X}_i \in \mathbb{R}^{d-1}$, and the error ε_i satisfies $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}(\varepsilon_i^2) = \sigma^2$ for $i = 1, \ldots, n$. Let $\boldsymbol{Q}_i = (Z_i, \boldsymbol{X}_i^T)^T$. For simplicity, we first assume that the variance σ^2 is known. Later, we will show that the same results are obtained if σ^2 is estimated. Recall that we assume the noise ε_i is sub-Gaussian, the minimum and maximum eigenvalues of $\mathbb{E}(\boldsymbol{Q}_i^{\otimes 2})$ are bounded away from 0 and infinity by constants, and $\boldsymbol{Q}_i$ is a sub-Gaussian vector. These assumptions are standard for the analysis of high dimensional linear regression models; see [5, 13, 34, 42].

We consider the penalized M-estimator $\hat{\boldsymbol{\beta}}$ in (1.1) with possibly nonconvex penalty functions. As shown by [37, 38] and many others, if the nonconvex penalty function satisfies the conditions (a)–(e) in [38], the penalized M-estimator $\hat{\boldsymbol{\beta}}$ satisfies the rate of convergence in Lemma 3.1. Based on the Gaussian quasi-log-likelihood, the decorrelated score function is

$$S(\theta, \boldsymbol{\gamma}) = -\frac{1}{\sigma^2 n} \sum_{i=1}^{n} (Y_i - \theta Z_i - \boldsymbol{\gamma}^T \boldsymbol{X}_i)(Z_i - \mathbf{w}^T \boldsymbol{X}_i),$$

where $\mathbf{w} = \mathbb{E}_{\boldsymbol{\beta}}(\boldsymbol{X}_i^{\otimes 2})^{-1} \mathbb{E}_{\boldsymbol{\beta}}(Z_i \boldsymbol{X}_i)$. Since the distribution of the design matrix does not depend on $\boldsymbol{\beta}$, we can replace $\mathbb{E}_{\boldsymbol{\beta}}(\cdot)$ by $\mathbb{E}(\cdot)$ for notational simplicity. In practice, under the null hypothesis $H_0 : \theta^* = 0$, the decorrelated score function can be estimated by

$$\hat{S}(0, \hat{\boldsymbol{\gamma}}) = -\frac{1}{\sigma^2 n} \sum_{i=1}^{n} (Y_i - \hat{\boldsymbol{\gamma}}^T \boldsymbol{X}_i)(Z_i - \hat{\mathbf{w}}^T \boldsymbol{X}_i), \tag{4.1}$$

where

$$\hat{\mathbf{w}} = \operatorname{argmin} \|\mathbf{w}\|_1 \qquad \text{s.t.} \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i (Z_i - \mathbf{w}^T \boldsymbol{X}_i) \right\|_\infty \le \lambda'. \tag{4.2}$$

In this example, the information identity holds, that is, $\mathbf{I}^* = \boldsymbol{\Sigma}^*$, where the (partial) information matrices are given by

$$\mathbf{I}^* = \sigma^{-2} \mathbb{E}(\boldsymbol{Q}_i^{\otimes 2}) \quad \text{and} \quad I^*_{\theta|\boldsymbol{\gamma}} = \sigma^{-2} \big(\mathbb{E}(Z_i^2) - \mathbb{E}(Z_i \boldsymbol{X}_i^T) \mathbb{E}(\boldsymbol{X}_i^{\otimes 2})^{-1} \mathbb{E}(\boldsymbol{X}_i Z_i)\big).$$

They can be easily estimated by

$$\hat{\mathbf{I}} = \frac{1}{\sigma^2 n} \sum_{i=1}^{n} \boldsymbol{Q}_i^{\otimes 2} \quad \text{and} \quad \hat{I}_{\theta|\boldsymbol{\gamma}} = \frac{1}{\sigma^2} \left\{ \frac{1}{n} \sum_{i=1}^{n} Z_i^2 - \hat{\mathbf{w}}^T \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i Z_i \right) \right\},$$

respectively. This leads to the score test statistic $\hat{U}_n = n^{1/2} \hat{S}(0, \hat{\boldsymbol{\gamma}}) \hat{I}_{\theta|\boldsymbol{\gamma}}^{-1/2}$. The following corollary of the general results in Theorem 3.1 establishes the asymptotic null distribution of $\hat{U}_n$.

COROLLARY 4.1. *Let* $S = \operatorname{supp}(\boldsymbol{\beta}^*)$ *and* $S' = \operatorname{supp}(\mathbf{w}^*)$ *satisfy* $|S| = s^*$ *and* $|S'| = s'$. *If* $n^{-1/2}(s' \vee s^*)\log d = o(1)$ *and* $\lambda \asymp \lambda' \asymp \sqrt{\frac{\log d}{n}}$, *then under* $H_0 : \theta^* = 0$ *for each* $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \left|\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \Phi(t)\right| = 0. \tag{4.3}$$

PROOF. A detailed proof is shown in the supplementary materials [27]. □

Note that the estimator $\hat{\mathbf{w}}$ in (4.2) has the same form as the Dantzig selector [6]. Using the L_1 penalty, an alternative estimator $\tilde{\mathbf{w}}$ in (2.11) is

$$\tilde{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (Z_i - \mathbf{w}^T \boldsymbol{X}_i)^2 + \lambda' \|\mathbf{w}\|_1 \right\}. \tag{4.4}$$

Under the same conditions in Corollary 4.1, the score test $\hat{U}_n$ with $\hat{\mathbf{w}}$ replaced by $\tilde{\mathbf{w}}$ in (4.4) satisfies $\lim_{n\to\infty} |\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \Phi(t)| = 0$, for any $t \in \mathbb{R}$.

The following corollary of Theorem 3.2 shows that the one-step estimator $\tilde{\theta}$ is asymptotically normal and semiparametrically efficient.

COROLLARY 4.2. *If* $n^{-1/2}(s' \vee s^*)\log d = o(1)$ *and* $\lambda \asymp \lambda' \asymp \sqrt{\frac{\log d}{n}}$, *then* $n^{1/2}(\tilde{\theta} - \theta^*)\hat{I}_{\theta|\boldsymbol{\gamma}}^{1/2} \rightsquigarrow N(0,1)$, *where* $\tilde{\theta}$ *and* $\hat{I}_{\theta|\boldsymbol{\gamma}}$ *are constructed based on either* $\hat{\mathbf{w}}$ *or* $\tilde{\mathbf{w}}$.

PROOF. A detailed proof is shown in the supplementary materials [27]. □

REMARK 6. In Corollary 4.1, we assume the sparsity of $\boldsymbol{\beta}^*$ and $\mathbf{w}^*$. Note that, by the block matrix inversion formula, the sparsity of $\mathbf{w}^*$ is implied by the sparsity of the precision matrix $(\mathbb{E}(\boldsymbol{Q}_i^{\otimes 2}))^{-1}$, assumed in [34]. If $\boldsymbol{Q} = (Z, \boldsymbol{X}^T)^T$ follows a Gaussian graphical model, then $\|\mathbf{w}^*\|_0$ is identical to the degree of the node Z in the graph, which can be much smaller than the maximum degree in the graph. Thus, to conduct the inference on the coefficient of Z, we only require the degree of the node Z to be relatively small. This is a more refined assumption than assuming the whole graph is sparse for the inference on a single component.

In the supplementary materials [27], we show that the weak convergence in (4.3) holds uniformly over $\boldsymbol{\beta}^* \in \Omega_0$, where $\Omega_0 = \{(0, \boldsymbol{\gamma}) : \|\boldsymbol{\gamma}\|_0 \le s^*\}$. This implies the honesty of the proposed test. We also establish the uniform convergence of $\hat{U}_n$ over $\boldsymbol{\beta}^*$ in the space of alternative hypothesis $\Omega_1(\tilde{C}, 1/2)$, where $\Omega_1(\tilde{C}, 1/2) = \{(\theta, \boldsymbol{\gamma}) : \theta = \tilde{C}n^{-1/2}, \|\boldsymbol{\gamma}\|_0 \le s^*\}$, for some constant $\tilde{C}$. This establishes the asymptotic local power of the proposed score test. The decorrelated score inference under the deterministic design is also discussed in the supplementary materials [27].

4.1.1. *Estimation of unknown variance* σ^2. In the previous section, we assume σ^2 is known. In this section, we consider the estimation of σ^2 and the asymptotic properties of the score test with estimated σ^2. With the estimator $\hat{\boldsymbol{\beta}}$, one can estimate σ^2 by $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (Y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{Q}_i)^2$. Consider the following score statistic with $\hat{\sigma}^2$:

$$\tilde{U}_n = -\frac{1}{\hat{\sigma} n^{1/2}} \sum_{i=1}^n (Y_i - \hat{\boldsymbol{\gamma}}^T \boldsymbol{X}_i)(Z_i - \hat{\mathbf{w}}^T \boldsymbol{X}_i)(H_Z - \hat{\mathbf{w}}^T \mathbf{H}_{XZ})^{-1/2},$$

where $H_Z = n^{-1}\sum_{i=1}^n Z_i^2$ and $\mathbf{H}_{XZ} = n^{-1}\sum_{i=1}^n Z_i \boldsymbol{X}_i$. The following corollary characterizes the asymptotic null distribution of $\tilde{U}_n$. In particular, we show that $\tilde{U}_n$ and $\hat{U}_n$ are uniformly asymptotically equivalent, where $\hat{U}_n$ is the score test statistic with known σ^2.

COROLLARY 4.3. *Assume that the conditions in Corollary 4.1 hold and the true parameter satisfies* $\sigma^{*2} \geq C$ *for some constant* $C > 0$. *Then for any* $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \left| \mathbb{P}_{\boldsymbol{\beta}^*}(\tilde{U}_n \leq t) - \Phi(t) \right| = 0.$$

PROOF. A detailed proof is shown in the supplementary materials [27]. □

REMARK 7. Note that Corollary 4.3 holds due to the orthogonality of parameters $\boldsymbol{\beta}$ and σ^2 in the quasi-log-likelihood function. In practice, there exist many alternative estimators such as the scaled Lasso [32],

$$(\tilde{\boldsymbol{\beta}}_{\text{scale}}, \tilde{\sigma}_{\text{scale}}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^d, \sigma>0}{\operatorname{argmin}} \left\{ \frac{1}{2\sigma n} \sum_{i=1}^n (Y_i - \boldsymbol{\beta}^T \boldsymbol{Q}_i)^2 + \frac{\sigma}{2} + \lambda \|\boldsymbol{\beta}\|_1 \right\}.$$

By Theorem 1 and Corollary 1 of [32], we can show that the decorrelated score test with $(\tilde{\boldsymbol{\beta}}_{\text{scale}}, \tilde{\sigma}_{\text{scale}})$ is asymptotically equivalent to $\tilde{U}_n$ and $\hat{U}_n$ and, therefore, has the same type I error and local asymptotic power. Moreover, the estimator $(\tilde{\boldsymbol{\beta}}_{\text{scale}}, \tilde{\sigma}_{\text{scale}})$ has the additional advantage of being tuning insensitive. We refer to [2, 32] for more details.

4.1.2. *Comparison with existing inferential methods*. In this subsection, we compare our decorrelated score test method for linear models to the desparsifying estimator in [34] and the debiased estimator in [13]. From the methodological perspective, the proposed decorrelated score method is new and different from [34] and [13]. Our method directly aims to construct a test statistic to test the validity of the null hypothesis. In contrast, both [34] and [13] focused on how to correct the bias of the Lasso estimator and construct an asymptotically normal estimator. In addition, our framework allows a variety of methods for estimating $\mathbf{w}^*$, including the Dantzig selector (4.2) and the Lasso type estimator (4.4). Van de Geer et al. [34] only considered the Lasso estimator and [13] proposed to solve an alternative

constrained optimization problem. The proposed one-step estimator with the Lasso estimator (4.4) is identical to the desparsifying estimator in [34].

We now comment on the assumptions and results. We do not assume the noise is Gaussian. Thus, our Theorem 4.1 and Corollary 4.2 are comparable to Theorem 2.4 of [34] and Theorem 4.1 of [13]. In terms of estimation efficiency, the proposed one-step estimator and the desparsifying estimator are both semiparametrically efficient, that is, $n^{1/2}(\tilde{\theta} - \theta^*) \rightsquigarrow N(0, I^{*-1}_{\theta|\boldsymbol{\gamma}})$. This follows from Corollary 4.2 and $\hat{I}_{\theta|\boldsymbol{\gamma}} - I^*_{\theta|\boldsymbol{\gamma}} = o_{\mathbb{P}}(1)$ (shown in the proof of Corollary 4.1). However, it is unclear whether the debiased estimator is semiparametrically efficient, because the sparsity assumption on the precision matrix $(\mathbb{E}(\boldsymbol{Q}_i^{\otimes 2}))^{-1}$ is relaxed. We refer to Remark 6 for further discussion on the sparsity assumption. Finally, we comment on the model misspecification. When the linear model assumption is invalid, the results in [34] and [13] are not directly applicable. However, the proposed decorrelated score test is robust to the model misspecification in the sense that it can infer the oracle parameter with theoretical guarantees; see Section 5 for details.

4.2. *Example* 2: *Logistic regression model.* Under the logistic regression model, the estimated decorrelated score function reduces to

$$\hat{S}(0, \hat{\boldsymbol{\gamma}}) = -\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \frac{\exp(\hat{\boldsymbol{\gamma}}^T \boldsymbol{X}_i)}{1 + \exp(\hat{\boldsymbol{\gamma}}^T \boldsymbol{X}_i)}\right)(Z_i - \hat{\mathbf{w}}^T \boldsymbol{X}_i),$$

where $\hat{\boldsymbol{\beta}}$ is taken as the penalized M-estimator in (1.1) with the L_1-penalty and $\hat{\mathbf{w}}$ in the context of logistic regression models is given by

$$\hat{\mathbf{w}} = \operatorname{argmin} \|\mathbf{w}\|_1 \qquad \text{s.t.} \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{Q}_i)}{[1 + \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{Q}_i)]^2}(Z_i - \mathbf{w}^T \boldsymbol{X}_i)\boldsymbol{X}_i \right\|_\infty \le \lambda'.$$

Alternatively, we can also use the estimator $\tilde{\mathbf{w}}$ in (2.11) with the L_1-penalty,

$$\tilde{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{Q}_i)}{[1 + \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{Q}_i)]^2}(Z_i - \mathbf{w}^T \boldsymbol{X}_i)^2 + \lambda' \|\mathbf{w}\|_1 \right\}.$$

In this example, the information identity also holds. Thus, we can calculate the Fisher information matrix as $\mathbf{I}^* = \mathbb{E}_{\boldsymbol{\beta}^*}(\exp(\boldsymbol{\beta}^{*T} \boldsymbol{Q}_i)\boldsymbol{Q}_i^{\otimes 2}/(1 + \exp(\boldsymbol{\beta}^{*T} \boldsymbol{Q}_i))^2)$. The partial Fisher information matrix is $I^*_{\theta|\boldsymbol{\gamma}} = \mathbb{E}_{\boldsymbol{\beta}^*}(Z_i(Z_i - \mathbf{w}^{*T} \boldsymbol{X}_i)\exp(\boldsymbol{\beta}^{*T} \boldsymbol{Q}_i)/(1 + \exp(\boldsymbol{\beta}^{*T} \boldsymbol{Q}_i))^2)$, which is estimated by

$$\hat{I}_{\theta|\boldsymbol{\gamma}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{Q}_i)}{[1 + \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{Q}_i)]^2} Z_i(Z_i - \hat{\mathbf{w}}^T \boldsymbol{X}_i).$$

Thus, the decorrelated score test statistic is $\hat{U}_n = n^{1/2}\hat{S}(0, \hat{\boldsymbol{\gamma}})\hat{I}^{-1/2}_{\theta|\boldsymbol{\gamma}}$.

Recall that, for the theoretical development, we assume $\lambda_{\min}(\mathbf{I}^*) \geq \kappa^2$ for some constant $\kappa > 0$, $\|\boldsymbol{Q}_i\|_\infty \leq K$, and $|\mathbf{w}^{*T}\boldsymbol{X}_i| \leq K$ for some constant $K > 0$. It is easy to see that the conditions in Theorem 3.3 of [34] imply our conditions. More importantly, our conditions are strictly weaker, because we do not require that the mean regression effect is bounded as in [34] [i.e., $\max_{1\leq i\leq n} |\boldsymbol{\beta}^{*T}\boldsymbol{Q}_i| = O(1)$]. This is accomplished by utilizing a more refined self-concordance property for logistic regressions [1]. We refer to the supplementary materials [27] for further discussion. The following corollary is an application of Theorems 3.1 and 3.2 in the context of logistic regression. In particular, we can see that the one-step estimator $\tilde{\theta}$ defined in (2.13) is semiparametrically efficient.

COROLLARY 4.4. *With* $\lambda \asymp \lambda' \asymp \sqrt{\frac{\log d}{n}}$, *if* $n^{-1/2}(s' \vee s^*)\log d = o(1)$, *then under the null hypothesis* $H_0 : \theta^* = 0$, *for each* $t \in \mathbb{R}$, $\lim_{n\to\infty} |\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \leq t) - \Phi(t)| = 0$, *and* $n^{1/2}(\tilde{\theta} - \theta^*)\hat{I}_{\theta|\boldsymbol{\gamma}}^{1/2} \rightsquigarrow N(0,1)$, *where* $\hat{U}_n$ *and* $\tilde{\theta}$ *are constructed based on either* $\hat{\mathbf{w}}$ *or* $\tilde{\mathbf{w}}$.

PROOF. A detailed proof is shown in the supplementary materials [27]. □

4.3. *Example* 3: *Poisson regression model*. Under the Poisson regression model, the estimated decorrelated score function reduces to

$$\hat{S}(0, \hat{\boldsymbol{\gamma}}) = -\frac{1}{n}\sum_{i=1}^{n}(Y_i - \exp(\hat{\boldsymbol{\gamma}}^T\boldsymbol{X}_i))(Z_i - \hat{\mathbf{w}}^T\boldsymbol{X}_i),$$

where $\hat{\mathbf{w}}$ in the context of Poisson regression is given by

$$\hat{\mathbf{w}} = \operatorname{argmin} \|\mathbf{w}\|_1 \qquad \text{s.t.} \left\| \frac{1}{n}\sum_{i=1}^{n} \exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{Q}_i)(Z_i - \mathbf{w}^T\boldsymbol{X}_i)\boldsymbol{X}_i \right\|_\infty \leq \lambda'.$$

Recall that, for the theoretical development, we assume $\lambda_{\min}(\mathbf{I}^*) \geq \kappa^2$ for some constant $\kappa > 0$, $\|\boldsymbol{Q}_i\|_\infty \leq K$, $|\mathbf{w}^{*T}\boldsymbol{X}_i| \leq K$ and $|\boldsymbol{\beta}^{*T}\boldsymbol{Q}_i| \leq K$ for some constant $K > 0$. It is seen that the conditions in Theorem 3.3 of [34] imply our conditions. Compared to the logistic regression, we need the extra technical condition $|\boldsymbol{\beta}^{*T}\boldsymbol{Q}_i| \leq K$, because $\exp(\boldsymbol{\beta}^{*T}\boldsymbol{Q}_i)$ may diverge to infinity very fast as $\boldsymbol{\beta}^{*T}\boldsymbol{Q}_i$ increases. The following corollary is an application of Theorems 3.1 and 3.2 in the context of Poisson regression.

COROLLARY 4.5. *With* $\lambda \asymp \lambda' \asymp \sqrt{\frac{\log d}{n}}$, *if* $n^{-1/2}(s' \vee s^*)\log d = o(1)$, *then under the null hypothesis*, *for each* $t \in \mathbb{R}$, $\lim_{n\to\infty} |\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \leq t) - \Phi(t)| = 0$, *and* $n^{1/2}(\tilde{\theta} - \theta^*)\hat{I}_{\theta|\boldsymbol{\gamma}}^{1/2} \rightsquigarrow N(0,1)$, *where* $\hat{U}_n = n^{1/2}\hat{S}(0, \hat{\boldsymbol{\gamma}})\hat{I}_{\theta|\boldsymbol{\gamma}}^{-1/2}$.

PROOF. A detailed proof is shown in the supplementary materials [27]. □

REMARK 8 (Extension to more generalized linear models). The results in Corollary 4.5 can be further extended to a broader class of generalized linear models, where the p.d.f. of Y given the covariates $\boldsymbol{Q}$ is $f(Y \mid \boldsymbol{Q}) = h(Y)\exp(Y\boldsymbol{\beta}^{*T}\boldsymbol{Q} - b(\boldsymbol{\beta}^{*T}\boldsymbol{Q}))$. Here, $h(\cdot)$ and $b(\cdot)$ are two known univariate functions. In the supplementary materials [27], we show that Corollary 4.5 holds if the following condition on $b(\cdot)$ holds. Let $b''(\cdot)$ be a continuous function. For some K_1, K_2 and any $t \in [K_1 - \varepsilon, K_2 + \varepsilon]$ with some constant $\varepsilon > 0$, and a sequence t_1 satisfying $|t_1 - t| = o(1)$, it holds that $0 < b''(t) \le C$ and $|b''(t_1) - b''(t)| \le C|t_1 - t|b''(t)$ for some constant $C > 0$.

In the supplementary materials [27], we show that the logistic regression, exponential regression and Poisson regression models all satisfy this condition. It is easy to check that this condition is equivalent to the conditions in [34]; see the supplementary materials [27] for the detailed results on generalized linear models.

4.4. *Example* 4: *Gaussian graphical model*. To estimate the precision matrix of a Gaussian graphical model, [21] proposed the following SCIO estimator $\hat{\boldsymbol{\beta}} = \arg\min \ell(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$, where $\ell(\boldsymbol{\beta})$ is a quadratic loss defined in (2.1). Given the loss function $\ell(\boldsymbol{\beta})$, the decorrelated score function reduces to $\hat{S}(\boldsymbol{\beta}) = \hat{\mathbf{v}}^T(\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \mathbf{e}_k)$, where $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix and $\hat{\mathbf{v}} = (1, -\hat{\mathbf{w}}^T)^T$. Let $\hat{\boldsymbol{\Sigma}}_{12} \in \mathbb{R}^{d-1}$ be the first row of $\hat{\boldsymbol{\Sigma}}$ with the first element $\hat{\Sigma}_{11}$ removed, and $\hat{\boldsymbol{\Sigma}}_{22}$ be a $(d-1) \times (d-1)$ submatrix of $\hat{\boldsymbol{\Sigma}}$ with the first row and column removed. In this example, $\hat{\mathbf{w}}$ is defined as

$$\hat{\mathbf{w}} = \operatorname{argmin} \|\mathbf{w}\|_1 \qquad \text{s.t. } \|\hat{\boldsymbol{\Sigma}}_{12} - \mathbf{w}^T\hat{\boldsymbol{\Sigma}}_{22}\|_\infty \le \lambda'.$$

With straightforward algebra, we can show that the asymptotic variance of the decorrelated score function is $\sigma_s^* = (\Theta_{kk}^*\Theta_{11}^* + \Theta_{k1}^{*2})/\Theta_{11}^{*2}$ and $I_{\theta|\boldsymbol{\gamma}}^* = 1/\Theta_{11}^*$. Based on the SCIO estimator, one can estimate Θ_{kk}^*, Θ_{11}^* and Θ_{k1}^* by $\hat{\Theta}_{kk}$, $\hat{\Theta}_{11}$ and $\hat{\Theta}_{k1}$, respectively. This leads to the plug-in estimators $\hat{\sigma}_s$ and $\hat{I}_{\theta|\boldsymbol{\gamma}}$. The resulting score test statistic becomes $\hat{U}_n = n^{1/2}\hat{S}(0, \hat{\boldsymbol{\gamma}})/\sqrt{\hat{\sigma}_s}$. Let $s^* = \max_{1\le j\le d}\sum_{k=1}^d I(\Theta_{jk} \ne 0)$ denote the maximum degree of the graph.

COROLLARY 4.6. *With* $\lambda \asymp \lambda' \asymp \sqrt{\frac{\log d}{n}}$, *if* $n^{-1/2}s^*\log d = o(1)$, *then under the null hypothesis, for each* $t \in \mathbb{R}$, $\lim_{n\to\infty}|\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \Phi(t)| = 0$ *and* $n^{1/2}(\tilde{\theta} - \theta^*)\hat{I}_{\theta|\boldsymbol{\gamma}}/\hat{\sigma}_s^{1/2} \rightsquigarrow N(0,1)$, *where* $\tilde{\theta}$ *is the one-step estimator.*

PROOF. A detailed proof is shown in the supplementary materials [27]. □

The inferential problems for high dimensional Gaussian graphical models have been studied by several authors. For instance, [12] extended the desparsifying method to the log-likelihood for Gaussian graphical model. Their analysis required the irrepresentable condition and the assumption $s^{*3}(\log d)/n = o(1)$. In addition,

nodewise regression based approaches are considered by [20, 30]. Compared to these existing methods, our assumption $n^{-1/2}s^* \log d = o(1)$ is weaker than [12] and is identical to [20, 30]. In addition, we find that the asymptotic variance of $n^{1/2}(\tilde{\theta} - \theta^*)$ is $\Theta^*_{kk}\Theta^*_{11} + \Theta^{*2}_{k1}$, which matches the inverse of the information lower bound for Gaussian graphical models. Thus, our estimator has the same asymptotic efficiency as [12, 20, 30] and is semiparametrically efficient.

4.5. *Example* 5: *Additive hazards model*. In this section, we apply the general results in Theorems 3.1 and 3.2 to the additive hazards model. Recall that the inference on the regression coefficient $\boldsymbol{\beta}$ is conducted based on the loss function $\ell(\boldsymbol{\beta})$ in (2.3). To estimate $\boldsymbol{\beta}$, [18, 23] proposed the estimator $\hat{\boldsymbol{\beta}}$ in (1.1) with both convex and nonconvex penalty functions. For simplicity, we focus on the Lasso type estimator. The estimated decorrelated score function has the form $\hat{S}(\boldsymbol{\beta}) = \hat{\mathbf{v}}^T(\mathbf{V}\boldsymbol{\beta} - \boldsymbol{b})$, where $\boldsymbol{b}$ and $\mathbf{V}$ are defined in (2.2) and $\hat{\mathbf{v}} = (1, -\hat{\mathbf{w}}^T)^T$ with

$$\hat{\mathbf{w}} = \operatorname{argmin} \|\mathbf{w}\|_1 \qquad \text{s.t. } \|\mathbf{V}_{\theta\boldsymbol{\gamma}} - \mathbf{w}^T\mathbf{V}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\|_\infty \le \lambda',$$

where $\mathbf{V}_{\theta\boldsymbol{\gamma}}$ and $\mathbf{V}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ are partitions of $\mathbf{V}$ corresponding to $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma}^T)^T$. By taking derivatives, it can be show that $\mathbf{I}^* = \mathbf{V}^*$, where $\mathbf{V}^*$ is defined in (2.4). As shown in [17], $\nabla\ell(\boldsymbol{\beta}^*)$ is a martingale integral. Applying the martingale theory, we find that $\boldsymbol{\Sigma}^*$ in Assumption 3.4 has the form of $\boldsymbol{\Sigma}^* = \mathbf{W}^*$, where $\mathbf{W}^*$ is defined in (2.4). Notice that $\mathbf{V}$ does not depend on the unknown parameters, and it is a consistent estimator of $\mathbf{V}^*$. In addition, $\mathbf{W}^*$ can be also consistently estimated by $\mathbf{W}$, where

$$\mathbf{W} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{\boldsymbol{Q}_i(t) - \bar{\boldsymbol{Q}}(t)\}^{\otimes 2}\, dN_i(t).$$

Thus, we can estimate $I^*_{\theta|\boldsymbol{\gamma}}$ by $\hat{I}_{\theta|\boldsymbol{\gamma}} = \mathbf{V}_{\theta\theta} - \hat{\mathbf{w}}^T\mathbf{V}_{\boldsymbol{\gamma}\theta}$, and estimate the asymptotic variance σ^*_s in Assumption 3.4 by $\hat{\sigma}_s = \hat{\mathbf{v}}^T\mathbf{W}\hat{\mathbf{v}}$. The resulting score test statistic is $\hat{U}_n = n^{1/2}\hat{S}(0, \hat{\boldsymbol{\gamma}})/\sqrt{\hat{\sigma}_s}$. The following corollary is an application of Theorems 3.1 and 3.2 to the additive hazards model.

COROLLARY 4.7. *With* $\lambda \asymp \lambda' \asymp \sqrt{\frac{\log d}{n}}$, *if* $n^{-1/2}(s' \vee s^*)\log d = o(1)$, *under the null hypothesis*, *for each* $t \in \mathbb{R}$, $\lim_{n\to\infty} |\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \Phi(t)| = 0$, *and* $n^{1/2}(\tilde{\theta} - \theta^*)\hat{I}_{\theta|\boldsymbol{\gamma}}/\hat{\sigma}_s^{1/2} \rightsquigarrow N(0, 1)$, *where* $\tilde{\theta}$ *is the one-step estimator.*

PROOF. A detailed proof is shown in the supplementary materials [27]. □

Recently, [44] proposed a variance reduced partial profiling method to construct valid confidence interval for θ in the additive hazards model. However, their method requires some strong conditions which are not needed in our decorrelated score framework. For instance, let $\eta > 0$ be a pre-specified value. Define the set

$$S_1 = \left\{ j \in \{1, \ldots, d-1\} : \left| \mathbb{E}\left[\int_0^\tau Y(t)Z(t)X_j(t)\, dt \right] \right| > \eta \right\}.$$

The condition (C5) of [44] required the cardinality of S_1 to increase with n in a suitable rate. Such a condition imposes stringent assumptions on the correlation between $Z(t)$ and $X_j(t)$, which is not needed in our framework.

5. Model misspecification. In the previous sections, an implicit assumption is that the probability model for Y_i given the covariate $\boldsymbol{Q}_i$ is correctly specified. In this section, we establish the theoretical properties of the decorrelated score test, if the true probability distribution $\mathbb{P}^*$ does not belong to the assumed statistical model $\mathcal{P} = \{\mathbb{P}_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \Omega\}$. To ease presentation, we let $\ell(\boldsymbol{\beta})$ denote the negative log-likelihood of the assumed model. The method and theory can be straightforwardly extended to inference based on general loss functions. We define the Kullback–Leibler divergence as

$$\mathrm{KL}(\boldsymbol{\beta}) = \mathbb{E}^*\left\{\log \frac{f^*(Y_i, \boldsymbol{Q}_i)}{f(Y_i, \boldsymbol{Q}_i; \boldsymbol{\beta})}\right\},$$

where $f^*(Y_i, \boldsymbol{Q}_i)$ is the true density function of $(Y_i, \boldsymbol{Q}_i)$, and $f(Y_i, \boldsymbol{Q}_i; \boldsymbol{\beta})$ is the density corresponding to the model $\mathbb{P}_{\boldsymbol{\beta}}$. Here, we use $\mathbb{P}^*(\cdot)$ and $\mathbb{E}^*(\cdot)$ to denote the probability and the expectation with respect to the true density function $f^*(Y_i, \boldsymbol{Q}_i)$. Let $\boldsymbol{\beta}^o$ denote the oracle parameter (or least false parameter) that minimizes the Kullback–Leibler divergence, that is, $\boldsymbol{\beta}^o = \operatorname{argmin}_{\boldsymbol{\beta}} \mathrm{KL}(\boldsymbol{\beta})$, where $\boldsymbol{\beta}^o = (\theta^o, \boldsymbol{\gamma}^o)$. Note that, if the model is correctly specified, we have $f^*(Y_i, \boldsymbol{Q}_i) = f(Y_i, \boldsymbol{Q}_i; \boldsymbol{\beta}^*)$ and the oracle parameter reduces to $\boldsymbol{\beta}^*$. Although, under the misspecified model, the true distribution is not estimable, it is often of interest to understand the behavior of the oracle parameter. In particular, assume that the inferential problem can be formulated as testing $H_0^o : \theta^o = 0$ versus $H_1^o : \theta^o \neq 0$. Similarly, we define $\mathbf{I}^o = \mathbb{E}^*(\nabla^2 \ell(\boldsymbol{\beta}^o))$, and $\mathbf{w}^{oT} = \mathbf{I}^o_{\theta\boldsymbol{\gamma}} \mathbf{I}^{o-1}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$. Denote $\boldsymbol{\Sigma}^o = \lim_{n\to\infty} \mathrm{Var}^*(n^{1/2} \nabla \ell(\boldsymbol{\beta}^o))$. The theoretical properties of the decorrelated score function under misspecified models are shown in the following theorem.

THEOREM 5.1. *Assume that the Assumptions 3.1–3.4 hold with $\boldsymbol{\beta}^*$, $\mathbf{w}^*$ and $\boldsymbol{\Sigma}^*$ replaced by $\boldsymbol{\beta}^o$, $\mathbf{w}^o$ and $\boldsymbol{\Sigma}^o$. Under the null hypothesis $H_0 : \theta^o = 0$, if $(\eta_1(n) + \eta_2(n))\sqrt{\log d} = o(1)$, then*

$$n^{1/2} \hat{S}(0, \hat{\boldsymbol{\gamma}}) / \sqrt{\sigma_s^o} \rightsquigarrow N(0, 1), \tag{5.1}$$

where $\sigma_s^o = \mathbf{v}^{oT} \boldsymbol{\Sigma}^o \mathbf{v}^o$ with $\mathbf{v}^o = (1, -\mathbf{w}^{oT})^T$. In addition, the decorrelated score test statistic $\hat{U}_n^o = n^{1/2} \hat{S}(0, \hat{\boldsymbol{\gamma}}) / \sqrt{\hat{\sigma}_s}$, where $\hat{\sigma}_s$ is a consistent estimator of σ_s^o, satisfies for any $t \in \mathbb{R}$

$$\lim_{n\to\infty} \left|\mathbb{P}^*\big(\hat{U}_n^o \leq t\big) - \Phi(t)\right| = 0.$$

The proof is similar to that of Theorem 3.1. We omit it for simplicity. As seen in Theorem 5.1, in the misspecified model, we need to standardize the score function

$\hat{S}(0, \hat{\boldsymbol{\gamma}})$ by $\sqrt{\mathbf{v}^{oT} \boldsymbol{\Sigma}^o \mathbf{v}^o}$, which reduces to the root of the partial information, that is, $I^{*1/2}_{\theta|\boldsymbol{\gamma}}$ under the correctly specified model. In addition, the proposed decorrelated score test statistic $\hat{U}^o_n$ has the same form as $\hat{U}_n$ defined in (3.2), even if the specified model is incorrect. This implies that the proposed test statistic $\hat{U}^o_n$ has the desired robustness property under model misspecification [40]. Specifically, the test based on $\hat{U}^o_n$ can correctly control the type I error no matter the model is correct or not.

As an illustration of the general results in Theorem 5.1, we now consider the linear regression under model misspecification. Since the linear model assumption is no longer true, we cannot use the simple identity $\varepsilon_i = Y_i - \boldsymbol{\beta}^{*T} \boldsymbol{Q}_i$. This makes the arguments in [13] and [34] not directly applicable to misspecified models. Assume that $\hat{\boldsymbol{\beta}}$ is the Lasso estimator. By definition, the oracle parameter $\boldsymbol{\beta}^o$ is defined as $\boldsymbol{\beta}^o = \operatorname{argmin}_{\boldsymbol{\beta}} \mathbb{E}^*(Y_i - \boldsymbol{\beta}^T \boldsymbol{Q}_i)^2$, and the decorrelated score function for testing $\theta^o = 0$ is

$$\hat{S}(0, \hat{\boldsymbol{\gamma}}) = -\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\boldsymbol{\gamma}}^T \boldsymbol{X}_i)(Z_i - \hat{\mathbf{w}}^T \boldsymbol{X}_i).$$

By definition, $\boldsymbol{\Sigma}^o = \mathbb{E}^*(\boldsymbol{Q}_i^{\otimes 2}(Y_i - \boldsymbol{\beta}^{oT} \boldsymbol{Q}_i)^2)$, which can be estimated by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Q}_i^{\otimes 2} (Y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{Q}_i)^2.$$

This leads to the decorrelated score test statistic $\hat{U}^o_n = n^{1/2} \hat{S}(0, \hat{\boldsymbol{\gamma}}) / \sqrt{\hat{\mathbf{v}}^T \hat{\boldsymbol{\Sigma}} \hat{\mathbf{v}}}$, where $\hat{\mathbf{v}} = (1, -\hat{\mathbf{w}}^T)^T$ and $\hat{\mathbf{w}}$ is defined in (4.2). We can obtain the following corollary of Theorem 5.1.

COROLLARY 5.1. *Assume that* (1) $2\kappa \le \lambda_{\min}(\mathbb{E}(\boldsymbol{Q}_i^{\otimes 2})) \le \lambda_{\max}(\mathbb{E}(\boldsymbol{Q}_i^{\otimes 2})) \le 2/\kappa$ *and* $\lambda_{\min}(\boldsymbol{\Sigma}^o) \ge 2\kappa$ *for some constant* $\kappa > 0$, (2) $Y_i - \boldsymbol{\gamma}^{oT} \boldsymbol{X}_i$ *is sub-Gaussian, and* $\boldsymbol{Q}_i$ *is a sub-Gaussian vector. Denote* $s' = \|\mathbf{w}^o\|_0$ *and* $s^* = \|\boldsymbol{\beta}^o\|_0$. *If* $n^{-1} s^* (\log(nd))^5 = o(1)$, $n^{-1/2}(s' \vee s^*) \log d = o(1)$, *and* $\lambda \asymp \lambda' \asymp \sqrt{\frac{\log d}{n}}$, *then under* $H_0^o : \theta^o = 0$, *for each* $t \in \mathbb{R}$,

$$\lim_{n \to \infty} \left| \mathbb{P}^*\left(\hat{U}^o_n \le t\right) - \Phi(t) \right| = 0.$$

PROOF. A detailed proof is shown in the supplementary materials [27]. □

6. Numerical results. In this section, we conduct simulation studies to investigate the finite sample performance of the proposed score test. In particular, we simulate the response from the following three models: the linear regression with the standard Gaussian noise, the logistic regression and Poisson regression. To generate the covariates, we simulate $n = 200$ independent samples from a multivariate Gaussian distribution $N_d(\mathbf{0}, \boldsymbol{\Sigma})$, where $d = 100, 200, 500$ and $\boldsymbol{\Sigma}$ is a Toeplitz matrix with $\Sigma_{jk} = \rho^{|j-k|}$. Here, we consider four possible values for ρ, that is, 0.25,

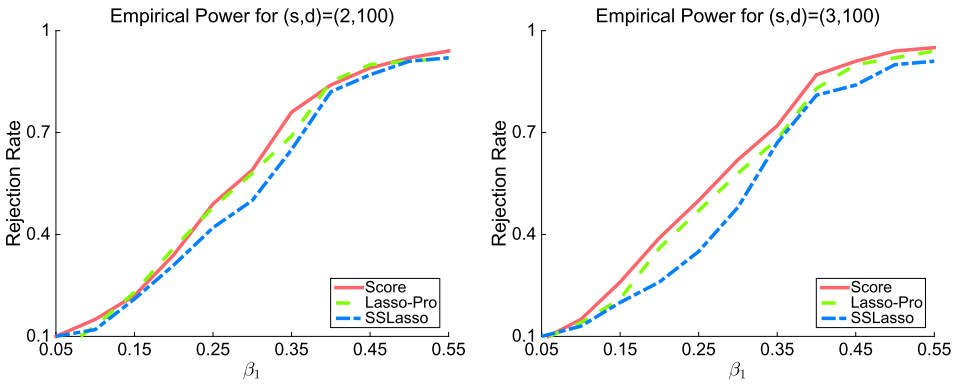

FIG. 2. *Power of the decorrelated score test, Lasso-Pro and SSLasso for linear regression.*

0.4, 0.6 and 0.75. The true value $\boldsymbol{\beta}^*$ satisfies $\|\boldsymbol{\beta}^*\|_0 = s$, with $s = 2$ and 3. We consider two scenarios for generating $\boldsymbol{\beta}^*$ on its support set S. In the first setting, we set $\boldsymbol{\beta}^*_S = (1, \ldots, 1)$, which is a Dirac measure. In the second setting, we generate each component of $\boldsymbol{\beta}^*_S$ from a uniform distribution on $[0, 2]$. Our goal is to test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. To check the validity of the type I error, we set $\beta^*_1 = 0$.

The tuning parameters λ and λ' are chosen by cross-validations. In the linear regression, we compare the performance of the score test with the desparsifying method (Lasso-Pro) in [34] and the de-bias method (SSLasso) in [13]. Both of their methods are equivalent to certain types of Wald tests. The type I errors of the three tests are reported in Tables 1. We find that all three tests have similar performance and their type I errors are close to the desired significance level, which is consistent with the asymptotic equivalence among these three tests. To evaluate the power of the tests, we regenerate the data with the values of β^*_1 ranging from 0 to 0.55. The power of the three tests is shown in Figure 2. We find that the decorrelated score test is slightly more powerful than the existing tests. This agrees with the statistical literature that the score test can be more powerful than the Wald test. To conclude this section, we note that the decorrelated score test also performs well in the logistic and Poisson regressions; see Table 2 for the type I errors and Figure 3 for the power. Further simulation results on the power of the proposed test are shown in the supplementary materials [27].

7. Discussion. In this paper, we propose a general framework for high dimensional inference based on the decorrelated score function. It can be used to test statistical hypotheses and construct confidence intervals. To broaden the applicability of the method, the theory is presented under a general setting. We note that the inferential problems for many high dimensional models can be analyzed by using the current framework. For example, [9] provided analysis to the high dimensional proportional hazards model. Unlike the additive hazards model, the

TABLE 1
Averaged type I error of the decorrelated score test, Lasso-Pro and SSLasso for the linear regression at 5% significance level

			$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
Method	s	d	Dirac	Unif	Dirac	Unif	Dirac	Unif	Dirac	Unif
Score	2	100	5.3%	4.9%	5.1%	4.8%	5.2%	5.5%	5.3%	5.0%
		200	5.1%	4.8%	5.3%	4.8%	5.9%	5.6%	4.7%	5.2%
		500	5.7%	5.7%	5.8%	5.8%	5.4%	5.7%	4.2%	4.3%
	3	100	5.1%	5.3%	5.2%	4.9%	5.0%	4.8%	4.4%	4.6%
		200	4.7%	4.8%	5.1%	5.4%	5.3%	4.9%	5.1%	4.8%
		500	4.4%	4.1%	4.3%	4.0%	4.1%	4.3%	4.0%	4.1%
Lasso-Pro	2	100	5.1%	5.2%	5.0%	4.7%	5.4%	5.1%	4.9%	5.1%
		200	5.3%	4.9%	4.8%	5.1%	5.4%	5.1%	4.9%	5.3%
		500	5.6%	5.7%	5.3%	4.7%	5.1%	4.6%	3.9%	4.1%
	3	100	5.0%	4.9%	5.2%	4.8%	5.3%	5.2%	4.7%	4.6%
		200	5.4%	5.3%	5.3%	5.2%	4.7%	5.6%	5.4%	5.5%
		500	5.5%	5.9%	5.1%	4.6%	4.7%	5.3%	6.2%	6.3%
SSLasso	2	100	5.0%	5.1%	5.2%	4.8%	4.8%	4.7%	5.2%	5.4%
		200	5.2%	4.7%	4.6%	5.4%	4.7%	5.1%	5.2%	4.8%
		500	5.4%	5.5%	4.5%	4.4%	4.5%	4.8%	6.2%	5.9%
	3	100	5.4%	5.3%	4.9%	4.7%	5.1%	5.0%	5.1%	4.9%
		200	5.3%	5.2%	4.9%	4.8%	5.3%	4.8%	4.5%	4.7%
		500	5.8%	5.6%	5.5%	5.7%	5.3%	5.6%	6.5%	6.1%

nonlinearity structure of the proportional hazards model poses additional technical challenges. We refer to [9] for a comprehensive investigation.

From a technical perspective, the sparsity assumption on $\mathbf{w}^*$ can be relaxed. The reason we assume $\mathbf{w}^*$ to be sparse is that we need the high dimensional vector $\mathbf{w}^*$ to be consistently estimated with a sufficiently fast rate of convergence.

One possible extension is to consider the weak sparsity case which imposes a certain decaying rate on the ordered entries of $\mathbf{w}^*$. This direction is left for further investigation.

APPENDIX A: UNIFORM CONVERGENCE OF SCORE TEST

In this appendix, we establish the general results for the uniform convergence of the decorrelated score test under the null and alternative hypotheses. In particular, we focus on the case that the information identity holds, for example, $\ell(\boldsymbol{\beta})$ is the negative log-likelihood. We note that the results can be easily extended to the general loss functions.

TABLE 2
Average type I error of the decorrelated score test for the logistic and Poisson regressions at 5% significance level

			$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
Method	s	d	Dirac	Unif	Dirac	Unif	Dirac	Unif	Dirac	Unif
Logistic	2	100	5.4%	5.3%	4.8%	4.9%	5.0%	4.9%	4.8%	5.1%
		200	5.1%	4.5%	4.9%	5.4%	4.6%	4.7%	4.4%	4.2%
		500	3.7%	4.2%	4.7%	4.4%	6.7%	5.9%	6.9%	6.1%
	3	100	5.6%	5.2%	5.4%	5.5%	4.8%	4.5%	4.7%	5.0%
		200	4.3%	4.5%	4.7%	4.9%	5.3%	5.5%	4.8%	4.6%
		500	3.6%	3.4%	3.6%	4.1%	3.7%	3.2%	5.5%	5.2%
Poisson	2	100	5.6%	5.0%	5.8%	5.7%	4.3%	4.9%	6.0%	6.1%
		200	5.5%	4.6%	5.9%	6.2%	6.0%	5.8%	6.2%	6.2%
		500	7.0%	6.8%	7.4%	6.6%	6.5%	6.0%	7.1%	7.4%
	3	100	5.3%	4.7%	5.8%	5.2%	5.6%	6.0%	5.8%	6.0%
		200	6.1%	6.5%	5.7%	6.8%	6.2%	6.1%	5.8%	6.4%
		500	6.7%	6.7%	7.0%	6.6%	7.3%	7.0%	7.3%	6.8%

A.1. Uniform weak convergence under the null hypothesis. Although in the previous section the limiting distribution of the score test statistic $\hat{U}_n$ is established, the convergence is shown under the fixed probability distribution $\mathbb{P}_{\boldsymbol{\beta}^*} = \mathbb{P}_{(0,\boldsymbol{\gamma}^*)}$. However, in practice, $\boldsymbol{\gamma}^*$ is unknown. To guarantee that the convergence properties are not affected by the values of $\boldsymbol{\gamma}^*$, it is of interest to strengthen

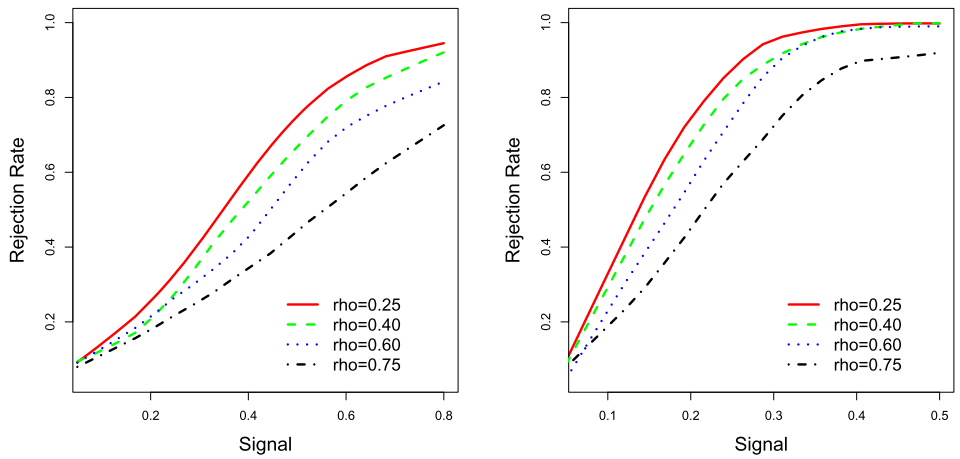

FIG. 3. *Power of the decorrelated score test for the logistic regression (left panel) and Poisson regression (right panel) with* $d = 200$, $s = 3$ *and* $\boldsymbol{\beta}^*_S = (1, \ldots, 1)$.

the weak convergence results in Theorem 3.1 to the weak convergence uniformly over the values of $\boldsymbol{\gamma}^*$. In particular, consider the following parameter space:

$$\Omega_0 = \{(0, \boldsymbol{\gamma}) : \|\boldsymbol{\gamma}\|_0 \le s^*, \text{for some } s^* \ll n\}.$$

To ensure that the parameter $\boldsymbol{\beta}$ can be still consistently estimated, we assume Ω_0 only contains sparse parameters. Similarly, to study the weak convergence uniformly over $\boldsymbol{\beta}^* \in \Omega_0$, we impose the following conditions.

ASSUMPTION A.1. It holds that $\lim_{n\to\infty} \inf_{\boldsymbol{\beta}^*\in\Omega_0} \mathbb{P}_{\boldsymbol{\beta}^*}(\mathcal{F}_1^{\boldsymbol{\beta}^*}) = 1$ and $\lim_{n\to\infty} \inf_{\boldsymbol{\beta}^*\in\Omega_0} \mathbb{P}_{\boldsymbol{\beta}^*}(\mathcal{F}_2^{\boldsymbol{\beta}^*}) = 1$, where $\mathcal{F}_1^{\boldsymbol{\beta}^*} = \{\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 \lesssim \eta_1(n)\}$ and $\mathcal{F}_2^{\boldsymbol{\beta}^*} = \{\|\hat{\mathbf{w}} - \mathbf{w}^*\|_1 \lesssim \eta_2(n)\}$, for some $\eta_1(n)$ and $\eta_2(n) \to 0$, as $n \to \infty$.

ASSUMPTION A.2. Assume that $\lim_{n\to\infty} \inf_{\boldsymbol{\beta}^*\in\Omega_0} \mathbb{P}_{\boldsymbol{\beta}^*}(\mathcal{F}_3^{\boldsymbol{\beta}^*}) = 1$, where $\mathcal{F}_3^{\boldsymbol{\beta}^*} = \{\|\nabla_{\boldsymbol{\gamma}} \ell(\boldsymbol{\beta}^*)\|_\infty \lesssim \sqrt{\log d/n}\}$, as $n \to \infty$.

ASSUMPTION A.3. Assume that $\lim_{n\to\infty} \inf_{\boldsymbol{\beta}^*\in\Omega_0} \mathbb{P}_{\boldsymbol{\beta}^*}(\mathcal{F}_4^{\boldsymbol{\beta}^*}) = 1$, where

$$\mathcal{F}_4^{\boldsymbol{\beta}^*} = \Big\{ \sup_{v\in[0,1]} \|\nabla^2_{\theta\boldsymbol{\gamma}} \ell(0, \boldsymbol{\gamma}_v) - \hat{\mathbf{w}}^T \nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \ell(0, \boldsymbol{\gamma}_v)\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \Big\}.$$

Here, $\boldsymbol{\gamma}_v = v\boldsymbol{\gamma}^* + (1-v)\hat{\boldsymbol{\gamma}}$ with $v \in [0, 1]$, as $n \to \infty$.

ASSUMPTION A.4. For $\mathbf{v}^* = (1, -\mathbf{w}^{*T})^T$, it holds that

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}^*\in\Omega_0} \sup_{t\in\mathbb{R}} \Big| \mathbb{P}_{\boldsymbol{\beta}^*} \Big(\frac{\sqrt{n}\mathbf{v}^{*T} \nabla \ell(\boldsymbol{\beta}^*)}{\sqrt{\mathbf{v}^{*T} \mathbf{I}^* \mathbf{v}^*}} \le t \Big) - \Phi(t) \Big| = 0.$$

Assume that $C' \le \mathbf{v}^{*T} \mathbf{I}^* \mathbf{v}^* \le 1/C'$, for some $C' > 0$.

These assumptions are verified for linear models in the supplementary materials [27].

Here, $\eta_1(n)$ and $\eta_2(n)$ are deterministic and do not depend on $\boldsymbol{\beta}^*$. Note that Assumptions A.1, A.2, A.3 and A.4 intrinsically play the same role as Assumptions 3.1, 3.2, 3.3 and 3.4.

But, to study the uniform convergence, we need to assume that the events $\mathcal{F}_1^{\boldsymbol{\beta}^*}$, $\mathcal{F}_2^{\boldsymbol{\beta}^*}$, $\mathcal{F}_3^{\boldsymbol{\beta}^*}$ and $\mathcal{F}_4^{\boldsymbol{\beta}^*}$ hold under the distribution $\mathbb{P}_{\boldsymbol{\beta}^*}$ uniformly over $\boldsymbol{\beta}^* \in \Omega_0$. The following theorem establishes the uniform convergence of the score test statistic $\hat{U}_n$ in (3.2).

THEOREM A.1. *Assume that the Assumptions A.1–A.4 hold. It also holds that* $(\eta_1(n) + \eta_2(n))\sqrt{\log d} = o(1)$. *Then*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}^*\in\Omega_0} \sup_{t\in\mathbb{R}} \big| \mathbb{P}_{\boldsymbol{\beta}^*}\big(n^{1/2} \hat{S}(0, \hat{\boldsymbol{\gamma}}) I^{*-1/2}_{\theta|\boldsymbol{\gamma}} \le t\big) - \Phi(t) \big| = 0. \tag{A.1}$$

Moreover, for any $\varepsilon > 0$, *if* $\lim_{n\to\infty} \inf_{\boldsymbol{\beta}^* \in \Omega_0} \mathbb{P}_{\boldsymbol{\beta}^*}(|\hat{I}_{\theta|\boldsymbol{\gamma}} - I^*_{\theta|\boldsymbol{\gamma}}| < \varepsilon) = 1$, *then*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}^* \in \Omega_0} \sup_{t\in\mathbb{R}} \big|\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \Phi(t)\big| = 0. \tag{A.2}$$

PROOF. A detailed proof is shown in Appendix B. □

REMARK 9. Theorem A.1 implies that the type I error of the score test T_n in (3.6) converges to its significance level α uniformly over $\boldsymbol{\beta}^* \in \Omega_0$, that is,

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}^* \in \Omega_0} \sup_{\alpha\in(0,1)} \big|\mathbb{P}_{\boldsymbol{\beta}^*}(T_n = 1) - \alpha\big| = 0.$$

The hypothesis test with such uniform convergence property is called the honest test. See [3, 13, 34] for further examples.

A.2. Uniform weak convergence under the alternative hypothesis. In this section, we consider the power of the score test for detecting the alternative hypothesis. In particular, we are interested in the limiting behavior of T_n under the sequence of local alternative hypothesis $H_{1n} : \theta^* = \tilde{C}n^{-\phi}$, where $\tilde{C}$ is a constant, and ϕ is a positive constant. Consider the following parameter space:

$$\Omega_1(\tilde{C}, \phi) = \big\{(\theta, \boldsymbol{\gamma}) : \theta = \tilde{C}n^{-\phi}, \|\boldsymbol{\gamma}\|_0 \le s^*, \text{for some } s^* \ll n\big\}.$$

The parameter space $\Omega_1(\tilde{C}, \phi)$ describes the local alternatives around the null hypothesis $\theta^* = 0$, in the sense that $\theta^* = \tilde{C}n^{-\phi}$ gradually shrinks to 0 as $n \to \infty$. Similar to Ω_0, we only consider sparse local alternatives. The following theorem characterizes the limiting distributions of the score test statistic $\hat{U}_n = n^{1/2}\hat{S}(0, \hat{\boldsymbol{\gamma}})\hat{I}^{-1/2}_{\theta|\boldsymbol{\gamma}}$, with respect to different values of ϕ.

THEOREM A.2. *Assume that the Assumptions A.1–A.4 with* Ω_0 *replaced by* $\Omega_1(\tilde{C}, \phi)$ *hold. In addition, we assume uniformly over* $\boldsymbol{\beta}^* \in \Omega_1(\tilde{C}, \phi)$,

$$\sqrt{n}\big|S(\theta^*, \boldsymbol{\gamma}^*) - S\big(0, \boldsymbol{\gamma}^*\big) - \theta^* \mathbf{I}^*_{\theta|\boldsymbol{\gamma}}\big| = o_{\mathbb{P}}(1),$$

and $\hat{I}_{\theta|\boldsymbol{\gamma}} - I^*_{\theta|\boldsymbol{\gamma}} = o_{\mathbb{P}}(1)$. *If* $(\eta_1(n) + \eta_2(n))\sqrt{\log d} = o(1)$, *then*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}^* \in \Omega_1(\tilde{C},\phi)} \sup_{t\in\mathbb{R}} \big|\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \Phi(t)\big| = 0 \qquad \text{if } \phi > 1/2, \tag{A.3}$$

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}^* \in \Omega_1(\tilde{C},\phi)} \sup_{t\in\mathbb{R}} \big|\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \Phi\big(t + \tilde{C}I^{*1/2}_{\theta|\boldsymbol{\gamma}}\big)\big| = 0 \qquad \text{if } \phi = 1/2, \tag{A.4}$$

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}^* \in \Omega_1(\tilde{C},\phi)} \mathbb{P}_{\boldsymbol{\beta}^*}\big(|\hat{U}_n| \le t\big) = 0 \qquad \text{if } \phi < 1/2. \tag{A.5}$$

Here, (A.5) *holds for any fixed* $t \in \mathbb{R}$ *and* $\tilde{C} \neq 0$.

PROOF. A detailed proof is shown in Appendix B. □

REMARK 10. This theorem implies that the score test statistic $\hat{U}_n$ has distinct limiting behaviors in terms of the magnitude of ϕ. In particular, (A.3) implies that $\hat{U}_n \rightsquigarrow N(0,1)$ if $\phi > 1/2$ and (A.4) implies that $\hat{U}_n \rightsquigarrow N(-\tilde{C} I_{\theta|\boldsymbol{\gamma}}^{*1/2}, 1)$ if $\phi = 1/2$. These results agree with the classical Rao's score test for low dimensional parameters.

REMARK 11. Note that the power of the two-sided test T_n in (3.6) is given by the probability of $T_n = 1$ when $\boldsymbol{\beta}^* \in \Omega_1(\tilde{C}, \phi)$. Given the fact that the type I error of T_n can be controlled at level α asymptotically, Theorem A.2 characterizes the uniform asymptotic power of T_n under the local alternative hypothesis $H_{1n} : \theta^* = \tilde{C} n^{-\phi}$. In particular, Theorem A.2 implies

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}^* \in \Omega_1(\tilde{C},\phi)} \sup_{\alpha\in(0,1)} \left|\mathbb{P}_{\boldsymbol{\beta}^*}(T_n = 1) - \alpha\right| = 0 \qquad \text{if } \phi > 1/2, \tag{A.6}$$

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}^* \in \Omega_1(\tilde{C},\phi)} \sup_{\alpha\in(0,1)} \left|\mathbb{P}_{\boldsymbol{\beta}^*}(T_n = 1) - \psi_\alpha)\right| = 0 \qquad \text{if } \phi = 1/2, \tag{A.7}$$

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta}^* \in \Omega_1(\tilde{C},\phi)} \mathbb{P}_{\boldsymbol{\beta}^*}(T_n = 1) = 1 \qquad \text{if } \phi < 1/2, \tag{A.8}$$

where $\psi_\alpha = 1 - \Phi(\Phi^{-1}(1-\alpha/2) + \tilde{C} I_{\theta|\boldsymbol{\gamma}}^{*1/2}) + \Phi(-\Phi^{-1}(1-\alpha/2) + \tilde{C} I_{\theta|\boldsymbol{\gamma}}^{*1/2})$. Here, (A.8) holds for any $\alpha \in [\delta, 1)$ with some constant $\delta > 0$ and $\tilde{C} \neq 0$. In particular, (A.6) implies that the test T_n has no power beyond the type I error to distinguish H_0 from H_{1n} if $\phi > 1/2$. Moreover, it is seen that $\psi_\alpha > \alpha$ for any $\tilde{C} \neq 0$. Hence, (A.7) shows that the test T_n is asymptotically unbiased. That means the proposed score test T_n has asymptotic power larger than the type I error for detecting $H_{1n} : \theta^* = \tilde{C} n^{-1/2}$. Finally, (A.8) implies that the minimal power of T_n increases to 1 as $n \to \infty$, if $\phi < 1/2$.

REMARK 12. Recall that the hypothesis test T_n is for $H_0 : \theta^* = 0$ versus $H_1 : \theta^* \neq 0$, which is two-sided. To test the one-sided alternative hypothesis, say $H_1' : \theta^* > 0$, with the significance level α, we can define the score test T_n', such that $T_n' = 1$ if and only if $\hat{U}_n < -\Phi^{-1}(1-\alpha)$. Theorem A.1 shows that the type I error of T_n' converges to its significance level α uniformly. In addition, by Theorem A.2, the uniform asymptotic power of T_n' under the local alternative hypothesis $H_{1n} : \theta^* = \tilde{C} n^{-1/2}$ for some $\tilde{C} > 0$ is given by (A.7) with ψ_α replaced by $\psi_\alpha' = 1 - \Phi(\Phi^{-1}(1-\alpha) - \tilde{C} I_{\theta|\boldsymbol{\gamma}}^{*1/2})$.

APPENDIX B: PROOFS OF MAIN RESULTS

B.1. Proof of Theorem 3.2. Recall that we have defined $\hat{\mathbf{v}} = (1, -\hat{\mathbf{w}}^T)^T$ and $\mathbf{v}^* = (1, -\mathbf{w}^{*T})^T$. Our goal is to show that

$$n^{1/2}\big|(\tilde{\theta} - \theta^*)I^*_{\theta|\boldsymbol{\gamma}}/\sigma_s^{*1/2} + \mathbf{v}^{*T}\nabla\ell(\boldsymbol{\beta}^*)/\sigma_s^{*1/2}\big| = o_{\mathbb{P}}(1). \tag{B.1}$$

By the definition of $\tilde{\theta}$, we have the following decomposition:

$$\begin{aligned}
&n^{1/2}\big|(\tilde{\theta} - \theta^*)I^*_{\theta|\boldsymbol{\gamma}} + \mathbf{v}^{*T}\nabla\ell(\boldsymbol{\beta}^*)\big| \\
&\quad = n^{1/2}\big|(\hat{\theta} - \theta^*)I^*_{\theta|\boldsymbol{\gamma}} - I^*_{\theta|\boldsymbol{\gamma}}\hat{I}^{-1}_{\theta|\boldsymbol{\gamma}}\hat{\mathbf{v}}^T\nabla\ell(\hat{\boldsymbol{\beta}}) + \mathbf{v}^{*T}\nabla\ell(\boldsymbol{\beta}^*)\big| \\
&\quad \le n^{1/2}\big|(\hat{\theta} - \theta^*)I^*_{\theta|\boldsymbol{\gamma}} - \mathbf{v}^{*T}\big(\nabla\ell(\hat{\boldsymbol{\beta}}) - \nabla\ell(\boldsymbol{\beta}^*)\big)\big| \\
&\qquad + n^{1/2}\big|(\hat{\mathbf{v}} - \mathbf{v}^*)^T\nabla\ell(\hat{\boldsymbol{\beta}})\big| + n^{1/2}\big|(I^*_{\theta|\boldsymbol{\gamma}}\hat{I}^{-1}_{\theta|\boldsymbol{\gamma}} - 1)\hat{\mathbf{v}}^T\nabla\ell(\hat{\boldsymbol{\beta}})\big| \\
&\quad := I_1 + I_2 + I_3.
\end{aligned}$$

The proof of Theorem 3.1 implies that $n^{1/2}\hat{\mathbf{v}}^T\nabla\ell(\hat{\boldsymbol{\beta}})/\sigma_s^{*1/2} = \mathcal{O}_{\mathbb{P}}(1)$. Thus, by the consistency of $\hat{I}_{\theta|\boldsymbol{\gamma}}$, we have $I_3/\sigma_s^{*1/2} = o_{\mathbb{P}}(1)$. Following the same proof of Theorem 3.1, it is easy to show that

$$|I_2| \lesssim \eta_2(n)\sqrt{\log d} + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1).$$

It remains to control the term I_1. By the smoothness condition in Assumption 3.3, we can show that

$$\begin{aligned}
|I_1| &\le n^{1/2}\big|(\hat{\theta} - \theta^*)I^*_{\theta|\boldsymbol{\gamma}} - \mathbf{v}^{*T}\nabla^2\ell(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\big| + o_{\mathbb{P}}(1) \\
&\le n^{1/2}\big|(\hat{\theta} - \theta^*)I^*_{\theta|\boldsymbol{\gamma}} - (\hat{\theta} - \theta^*)\big(\nabla^2_{\theta\theta}\ell(\boldsymbol{\beta}^*) - \mathbf{w}^{*T}\nabla^2_{\boldsymbol{\gamma}\theta}\ell(\boldsymbol{\beta}^*)\big)\big| \\
&\qquad + n^{1/2}\big|(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\big(\nabla^2_{\theta\boldsymbol{\gamma}}\ell(\boldsymbol{\beta}^*) - \mathbf{w}^{*T}\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\ell(\boldsymbol{\beta}^*)\big)\big| + o_{\mathbb{P}}(1) \\
&\lesssim n^{1/2}\big\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\big\|_1\|\mathbf{T}\|_\infty + o_{\mathbb{P}}(1),
\end{aligned}$$

where $\mathbf{T} = [I^*_{\theta|\boldsymbol{\gamma}} - (\nabla^2_{\theta\theta}\ell(\boldsymbol{\beta}^*) - \mathbf{w}^{*T}\nabla^2_{\boldsymbol{\gamma}\theta}\ell(\boldsymbol{\beta}^*)), \nabla^2_{\theta\boldsymbol{\gamma}}\ell(\boldsymbol{\beta}^*) - \mathbf{w}^{*T}\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\ell(\boldsymbol{\beta}^*)]$ is a d dimensional vector. By Assumption 3.2, $\|\mathbf{T}\|_\infty \lesssim \sqrt{\log d/n}$. Thus,

$$|I_1| \lesssim \eta_1(n)\sqrt{\log d} + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1).$$

This completes the proof of (B.1).

B.2. Proof of Theorem A.1. We first present the following lemma.

LEMMA B.1. *Under the Assumption A.1–A.4, it also holds that* $(\eta_1(n) + \eta_2(n))\sqrt{\log d} = o(1)$. *Then*

$$\lim_{n\to\infty}\sup_{\boldsymbol{\beta}^*\in\Omega_0}\sup_{t\in\mathbb{R}}\big|\mathbb{P}_{\boldsymbol{\beta}^*}\big(n^{1/2}\hat{S}(0, \hat{\boldsymbol{\gamma}})I^{*-1/2}_{\theta|\boldsymbol{\gamma}} \le t\big) - \Phi(t)\big| = 0.$$

PROOF. A detailed proof is shown in the supplementary materials [27]. □

We now present the proof of Theorem A.1.

PROOF OF THEOREM A.1. Let ψ_n denote a sequence converging to 0 and satisfying $|\hat{I}_{\theta|\boldsymbol{\gamma}} - I^*_{\theta|\boldsymbol{\gamma}}| \lesssim \psi_n$. Denote $U_n = n^{1/2}\hat{S}(0, \hat{\boldsymbol{\gamma}})I^{*-1/2}_{\theta|\boldsymbol{\gamma}}$. To show (A.2), for any t and a sequence of positive $\delta_n \to 0$ to be determined later, we have

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \Phi(t) &= \{\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \mathbb{P}_{\boldsymbol{\beta}^*}(U_n \le t + \delta_n)\} \\ &\quad + \{\mathbb{P}_{\boldsymbol{\beta}^*}(U_n \le t + \delta_n) - \Phi(t + \delta_n)\} + \{\Phi(t + \delta_n) - \Phi(t)\} \\ &:= I_1 + I_2 + I_3.\end{aligned}$$

In the following, we first show that $\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}^*\in\Omega_0} \sup_{t\in\mathbb{R}} I_1 \le 0$. By the triangle inequality, it is easily seen that

$$\begin{aligned}\sup_{t\in\mathbb{R}} I_1 &\le \mathbb{P}_{\boldsymbol{\beta}^*}\big(|\hat{U}_n - U_n| \ge \delta_n\big) = \mathbb{P}_{\boldsymbol{\beta}^*}\big(|U_n||1 - \hat{I}^{-1/2}_{\theta|\boldsymbol{\gamma}} I^{*1/2}_{\theta|\boldsymbol{\gamma}}| \ge \delta_n\big) \\ &\le \mathbb{P}_{\boldsymbol{\beta}^*}\big(|U_n| \ge \delta_n^{-1}\big) + \mathbb{P}_{\boldsymbol{\beta}^*}\big(|1 - \hat{I}^{-1/2}_{\theta|\boldsymbol{\gamma}} I^{*1/2}_{\theta|\boldsymbol{\gamma}}| \ge \delta_n^2\big).\end{aligned} \tag{B.2}$$

The first term of (B.2) can be bounded by

$$\mathbb{P}_{\boldsymbol{\beta}^*}\big(|U_n| \ge \delta_n^{-1}\big) \le \big|\mathbb{P}_{\boldsymbol{\beta}^*}\big(|U_n| \ge \delta_n^{-1}\big) - \mathbb{P}\big(|N| \ge \delta_n^{-1}\big)\big| + \mathbb{P}\big(|N| \ge \delta_n^{-1}\big),$$

where $N \sim N(0, 1)$. By Lemma B.1,

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}^*\in\Omega_0} \sup_{\delta_n\in\mathbb{R}} \big|\mathbb{P}_{\boldsymbol{\beta}^*}\big(|U_n| \ge \delta_n^{-1}\big) - \mathbb{P}\big(|N| \ge \delta_n^{-1}\big)\big| = 0.$$

The tail bound for the standard normal distribution yields $\mathbb{P}(|N| \ge \delta_n^{-1}) \le 2\frac{\delta_n}{\sqrt{2\pi}} \exp(-\frac{1}{2\delta_n^2}) \to 0$, as $\delta_n \to 0$. Thus, we can show that $\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}^*\in\Omega_0} \sup_{\delta_n\in\mathbb{R}} \mathbb{P}_{\boldsymbol{\beta}^*}(|U_n| \ge \delta_n^{-1}) \le 0$. That means, the first term of (B.2) is bounded above by 0. For the second term of (B.2), we have

$$\mathbb{P}_{\boldsymbol{\beta}^*}\big(|1 - \hat{I}^{-1/2}_{\theta|\boldsymbol{\gamma}} I^{*1/2}_{\theta|\boldsymbol{\gamma}}| \ge \delta_n^2\big) = \mathbb{P}_{\boldsymbol{\beta}^*}\bigg(\frac{|\hat{I}_{\theta|\boldsymbol{\gamma}} - I^*_{\theta|\boldsymbol{\gamma}}|}{(\hat{I}^{1/2}_{\theta|\boldsymbol{\gamma}} + I^{*1/2}_{\theta|\boldsymbol{\gamma}})\hat{I}^{1/2}_{\theta|\boldsymbol{\gamma}}} \ge \delta_n^2\bigg).$$

By the assumption $C' \le I^*_{\theta|\boldsymbol{\gamma}}$, we can show that

$$|\hat{I}_{\theta|\boldsymbol{\gamma}} - I^*_{\theta|\boldsymbol{\gamma}}|\{(\hat{I}^{1/2}_{\theta|\boldsymbol{\gamma}} + I^{*1/2}_{\theta|\boldsymbol{\gamma}})\hat{I}^{1/2}_{\theta|\boldsymbol{\gamma}}\}^{-1} \lesssim |\hat{I}_{\theta|\boldsymbol{\gamma}} - I^*_{\theta|\boldsymbol{\gamma}}| \lesssim \psi_n,$$

since $\hat{I}_{\theta|\boldsymbol{\gamma}} \ge C' - \psi_n \ge C'/2$ for n large enough. Hence, with $\delta_n = C\psi_n^{1/2}$, for some sufficiently large constant C, we obtain that the second term of (B.2), $\mathbb{P}_{\boldsymbol{\beta}^*}(|1 - \hat{I}^{-1/2}_{\theta|\boldsymbol{\gamma}} I^{*1/2}_{\theta|\boldsymbol{\gamma}}| \ge \delta_n^2) = 0$, for n large enough. As a result, $\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}^*\in\Omega_0} \sup_{t\in\mathbb{R}} I_1 \le 0$. By Lemma B.1, we can show that

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}^*\in\Omega_0} \sup_{t\in\mathbb{R}} I_2 \le \limsup_{n\to\infty} \sup_{\boldsymbol{\beta}^*\in\Omega_0} \sup_{t'\in\mathbb{R}} \big|\mathbb{P}_{\boldsymbol{\beta}^*}\big(U_n \le t'\big) - \Phi\big(t'\big)\big| = 0.$$

Finally, $I_3 \le (2\pi)^{-1/2}\delta_n$, which implies that $\limsup_{n\to\infty}\sup_{\boldsymbol{\beta}^*\in\Omega_0}\sup_{t\in\mathbb{R}} I_3 \le 0$. Combining these results, we obtain

$$\limsup_{n\to\infty}\sup_{\boldsymbol{\beta}^*\in\Omega_0}\sup_{t\in\mathbb{R}}\{\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \Phi(t)\} \le 0.$$

Similar arguments yields the bound for the minimum,

$$\liminf_{n\to\infty}\inf_{\boldsymbol{\beta}^*\in\Omega_0}\inf_{t\in\mathbb{R}}\{\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \Phi(t)\} \ge 0.$$

This completes the proof of (A.2). □

B.3. Proof of Theorem A.2. We start from the Lemma B.2.

LEMMA B.2. *Under the Assumptions in Theorem A.2, we have*

$$\lim_{n\to\infty}\sup_{\boldsymbol{\beta}^*\in\Omega_1(\tilde{C},\phi)}\sup_{t\in\mathbb{R}}|\mathbb{P}_{\boldsymbol{\beta}^*}(n^{1/2}\hat{S}(0,\hat{\boldsymbol{\gamma}})I_{\theta|\boldsymbol{\gamma}}^{*-1/2} \le t) - \Phi(t)| = 0, \tag{B.3}$$

$$\lim_{n\to\infty}\sup_{\boldsymbol{\beta}^*\in\Omega_1(\tilde{C},\phi)}\sup_{t\in\mathbb{R}}|\mathbb{P}_{\boldsymbol{\beta}^*}(n^{1/2}\hat{S}(0,\hat{\boldsymbol{\gamma}})I_{\theta|\boldsymbol{\gamma}}^{*-1/2} \le t) - \Phi(t+\tilde{C}I_{\theta|\boldsymbol{\gamma}}^{*1/2})| = 0, \tag{B.4}$$

for $\phi > 1/2$ and $\phi = 1/2$, respectively. For any fixed $t \in \mathbb{R}$ and $\tilde{C} \ne 0$,

$$\lim_{n\to\infty}\sup_{\boldsymbol{\beta}^*\in\Omega_1(\tilde{C},\phi)}\mathbb{P}_{\boldsymbol{\beta}^*}(|n^{1/2}\hat{S}(0,\hat{\boldsymbol{\gamma}})I_{\theta|\boldsymbol{\gamma}}^{*-1/2}| \le t) = 0 \qquad \text{if } \phi < 1/2. \tag{B.5}$$

PROOF. A detailed proof is shown in the supplementary materials [27]. □

Given Lemma B.2, we now prove Theorem A.2.

PROOF OF THEOREM A.2. The proof is similar to that of Theorem A.1. To highlight the difference, we only present the proofs of (A.4) and (A.5). Let ψ_n denote a sequence converging to 0 and satisfying $|\hat{I}_{\theta|\boldsymbol{\gamma}} - I_{\theta|\boldsymbol{\gamma}}^*| \lesssim \psi_n$. Denote $U_n = n^{1/2}\hat{S}(0,\hat{\boldsymbol{\gamma}})I_{\theta|\boldsymbol{\gamma}}^{*-1/2}$. For any t and a sequence of positive $\delta_n \to 0$ to be determined later,

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \Phi(t+\tilde{C}I_{\theta|\boldsymbol{\gamma}}^{*1/2}) &= \{\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \le t) - \mathbb{P}_{\boldsymbol{\beta}^*}(U_n \le t+\delta_n)\} \\ &\quad + \{\mathbb{P}_{\boldsymbol{\beta}^*}(U_n \le t+\delta_n) - \Phi(t+\tilde{C}I_{\theta|\boldsymbol{\gamma}}^{*1/2}+\delta_n)\} \\ &\quad + \{\Phi(t+\tilde{C}I_{\theta|\boldsymbol{\gamma}}^{*1/2}+\delta_n) - \Phi(t+\tilde{C}I_{\theta|\boldsymbol{\gamma}}^{*1/2})\} \\ &:= I_1 + I_2 + I_3.\end{aligned}$$

By the triangle inequality, it is easily seen that

$$\begin{aligned}\sup_{t\in\mathbb{R}} I_1 &\le \mathbb{P}_{\boldsymbol{\beta}^*}(|\hat{U}_n - U_n| \ge \delta_n) = \mathbb{P}_{\boldsymbol{\beta}^*}(|U_n||1 - \hat{I}_{\theta|\boldsymbol{\gamma}}^{-1/2}I_{\theta|\boldsymbol{\gamma}}^{*1/2}| \ge \delta_n) \\ &\le \mathbb{P}_{\boldsymbol{\beta}^*}(|U_n| \ge \delta_n^{-1}) + \mathbb{P}_{\boldsymbol{\beta}^*}(|1 - \hat{I}_{\theta|\boldsymbol{\gamma}}^{-1/2}I_{\theta|\boldsymbol{\gamma}}^{*1/2}| \ge \delta_n^2).\end{aligned} \tag{B.6}$$

The first term of (B.6) can be further bounded by

$$\mathbb{P}_{\boldsymbol{\beta}^*}\big(|U_n| \geq \delta_n^{-1}\big) \leq \big|\mathbb{P}_{\boldsymbol{\beta}^*}\big(|U_n| \geq \delta_n^{-1}\big) - \mathbb{P}\big(\big|N - \tilde{C} I_{\theta|\boldsymbol{\gamma}}^{*1/2}\big| \geq \delta_n^{-1}\big)\big| + \mathbb{P}\big(\big|N - \tilde{C} I_{\theta|\boldsymbol{\gamma}}^{*1/2}\big| \geq \delta_n^{-1}\big),$$

where $N \sim N(0,1)$. By Lemma B.2,

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}^* \in \Omega_1(\tilde{C},\phi)} \sup_{\delta_n \in \mathbb{R}} \big|\mathbb{P}_{\boldsymbol{\beta}^*}\big(|U_n| \geq \delta_n^{-1}\big) - \mathbb{P}\big(\big|N - \tilde{C} I_{\theta|\boldsymbol{\gamma}}^{*1/2}\big| \geq \delta_n^{-1}\big)\big| = 0,$$

and the tail bound for the standard normal distribution yields

$$\begin{aligned}\mathbb{P}\big(\big|N - \tilde{C} I_{\theta|\boldsymbol{\gamma}}^{*1/2}\big| \geq \delta_n^{-1}\big) &\leq \mathbb{P}\big(|N| \geq \delta_n^{-1} - |\tilde{C}| I_{\theta|\boldsymbol{\gamma}}^{*1/2}\big) \\ &\leq \frac{2}{\sqrt{2\pi}(\delta_n^{-1} - |\tilde{C}| I_{\theta|\boldsymbol{\gamma}}^{*1/2})} \exp\bigg(-\frac{(\delta_n^{-1} - |\tilde{C}| I_{\theta|\boldsymbol{\gamma}}^{*1/2})^2}{2}\bigg) \\ &\to 0,\end{aligned}$$

as $\delta \to 0$, uniformly over $\boldsymbol{\beta}^*$, due to $I_{\theta|\boldsymbol{\gamma}}^* \leq C''$. Thus, the first term of (B.6) is bounded above by 0. For the second term of (B.6), we have

$$\mathbb{P}_{\boldsymbol{\beta}^*}\big(\big|1 - \hat{I}_{\theta|\boldsymbol{\gamma}}^{-1/2} I_{\theta|\boldsymbol{\gamma}}^{*1/2}\big| \geq \delta_n^2\big) = \mathbb{P}_{\boldsymbol{\beta}^*}\bigg(\frac{|\hat{I}_{\theta|\boldsymbol{\gamma}} - I_{\theta|\boldsymbol{\gamma}}^*|}{(\hat{I}_{\theta|\boldsymbol{\gamma}}^{1/2} + I_{\theta|\boldsymbol{\gamma}}^{*1/2})\hat{I}_{\theta|\boldsymbol{\gamma}}^{1/2}} \geq \delta_n^2\bigg).$$

Applying the similar arguments, we can show that the second term of (B.6) goes to 0, and $\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}^* \in \Omega_0} \sup_{t\in\mathbb{R}} I_1 \leq 0$. By Lemma B.2, we can obtain that $\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}^* \in \Omega_1(\tilde{C},\phi)} \sup_{t\in\mathbb{R}} I_2$ is less than

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}^* \in \Omega_1(\tilde{C},\phi)} \sup_{t' \in \mathbb{R}} \big|\mathbb{P}_{\boldsymbol{\beta}^*}\big(U_n \leq t'\big) - \Phi\big(t' + \tilde{C} I_{\theta|\boldsymbol{\gamma}}^{*1/2}\big)\big| = 0.$$

Finally, $I_3 \leq (2\pi)^{-1/2}\delta_n$, which implies $\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}^* \in \Omega_1(\tilde{C},\phi)} \sup_{t\in\mathbb{R}} I_3 \leq 0$. Combining these results, we obtain

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}^* \in \Omega_1(\tilde{C},\phi)} \sup_{t\in\mathbb{R}} \big\{\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \leq t) - \Phi\big(t + \tilde{C} I_{\theta|\boldsymbol{\gamma}}^{*1/2}\big)\big\} \leq 0.$$

Similar arguments yield the following lower bound:

$$\liminf_{n\to\infty} \inf_{\boldsymbol{\beta}^* \in \Omega_1(\tilde{C},\phi)} \inf_{t\in\mathbb{R}} \big\{\mathbb{P}_{\boldsymbol{\beta}^*}(\hat{U}_n \leq t) - \Phi\big(t + \tilde{C} I_{\theta|\boldsymbol{\gamma}}^{*1/2}\big)\big\} \geq 0.$$

This completes the proof of (A.4). For (A.5), since $\sup_{\boldsymbol{\beta}^* \in \Omega_1(\tilde{C},\phi)} |\hat{I}_{\theta|\boldsymbol{\gamma}} - I_{\theta|\boldsymbol{\gamma}}^*| = o_{\mathbb{P}}(1)$ and $C' \leq I_{\theta|\boldsymbol{\gamma}}^*$, we have $|\hat{I}_{\theta|\boldsymbol{\gamma}}/I_{\theta|\boldsymbol{\gamma}}^* - 1| \leq 3$, for n large enough (not depending on $\boldsymbol{\beta}^*$). Given any $t \in \mathbb{R}$, for n sufficiently large,

$$\mathbb{P}_{\boldsymbol{\beta}^*}\big(|\hat{U}_n| \leq t\big) = \mathbb{P}_{\boldsymbol{\beta}^*}\big(|U_n| \leq t\big(\hat{I}_{\theta|\boldsymbol{\gamma}}/I_{\theta|\boldsymbol{\gamma}}^*\big)^{1/2}\big) \leq \mathbb{P}_{\boldsymbol{\beta}^*}\big(|U_n| \leq 2t\big).$$

Hence, by (B.5) in Lemma B.2, we finally obtain

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}^*\in\Omega_1(\tilde{C},\phi)} \mathbb{P}_{\boldsymbol{\beta}^*}\big(|\hat{U}_n| \le t\big) \le \lim_{n\to\infty} \sup_{\boldsymbol{\beta}^*\in\Omega_1(\tilde{C},\phi)} \mathbb{P}_{\boldsymbol{\beta}^*}\big(|U_n| \le 2t\big) = 0. \qquad \square$$

Acknowledgments. We thank the Editor, the Associate Editor and the referees for their helpful comments, which significantly improve the paper. We also thank Ethan X. Fang for helping with the numerical studies, and Heather Battey for useful comments.

SUPPLEMENTARY MATERIAL

Supplement to "A general theory of hypothesis tests and confidence regions for sparse high dimensional models" (DOI: 10.1214/16-AOS1448SUPP; .pdf). The supplementary materials contain additional technical details, simulation results and proofs.

REFERENCES

[1] BACH, F. (2010). Self-concordant analysis for logistic regression. *Electron. J. Stat.* **4** 384–414. MR2645490

[2] BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324

[3] BELLONI, A., CHERNOZHUKOV, V. and WEI, Y. (2013). Honest confidence regions for logistic regression with a large number of controls. Preprint. Available at arXiv:1304.3969.

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

[5] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

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

[7] COX, D. R. and HINKLEY, D. V. (1979). *Theoretical Statistics*. CRC Press, Boca Raton, FL.

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

[9] FANG, E. X., NING, Y. and LIU, H. (2014). Testing and confidence intervals for high dimensional proportional hazards model. Preprint. Available at arXiv:1412.5158.

[10] GODAMBE, V. P. and KALE, B. K. (1991). Estimating functions: An overview. In *Estimating Functions*. *Oxford Statist. Sci. Ser.* **7** 3–20. Oxford Univ. Press, New York. MR1163993

[11] GODFREY, L. G. (1991). *Misspecification Tests in Econometrics*: *The Lagrange Multiplier Principle and Other Approaches*. *Econometric Society Monographs* **16**. Cambridge Univ. Press, Cambridge. Reprint of the 1988 original. MR1113260

[12] JANKOVA, J. and VAN DE GEER, S. (2013). Confidence intervals for high-dimensional inverse covariance estimation. Preprint. Available at arXiv:1403.6752.

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

[14] JAVANMARD, A. and MONTANARI, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inform. Theory* **60** 6522–6554. MR3265038

[15] KNIGHT, K. and FU, W. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

[16] LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2013). Exact inference after model selection via the Lasso. Preprint. Available at arXiv:1311.6238.

[17] LIN, D. Y. and YING, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81** 61–71. MR1279656

[18] LIN, W. and LV, J. (2013). High-dimensional sparse additive hazards regression. *J. Amer. Statist. Assoc.* **108** 247–264. MR3174617

[19] LINDSAY, B. (1982). Conditional score functions: Some optimality results. *Biometrika* **69** 503–512. MR0695197

[20] LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* **41** 2948–2978. MR3161453

[21] LIU, W. and LUO, X. (2012). High-dimensional sparse precision matrix estimation via sparse column inverse operator. Preprint. Available at arXiv:1203.3896.

[22] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the Lasso. *Ann. Statist.* **42** 413–468. MR3210970

[23] MARTINUSSEN, T. and SCHEIKE, T. H. (2009). Covariate selection for the semiparametric additive risk model. *Scand. J. Stat.* **36** 602–619. MR2572578

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

[25] MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523

[26] MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681. MR2750584

[27] NING, Y. and LIU, H. (2016). Supplement to "A general theory of hypothesis tests and confidence regions for sparse high dimensional models." DOI:10.1214/16-AOS1448SUPP.

[28] PORTNOY, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12** 1298–1309. MR0760690

[29] PORTNOY, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13** 1403–1417. MR0811499

[30] REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. MR3346695

[31] SMALL, C. G. and MCLEISH, D. L. (2011). *Hilbert Space Methods in Probability and Statistical Inference* **920**. John Wiley & Sons, New York.

[32] SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

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

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

[35] VAN DER VAART, A. W. (2000). *Asymptotic Statistics* **3**. Cambridge Univ. Press, Cambridge.

[36] VOORMAN, A., SHOJAIE, A. and WITTEN, D. (2014). Inference in high dimensions with the penalized score test. Preprint. Available at arXiv:1401.2678.

[37] WANG, L., KIM, Y. and LI, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *Ann. Statist.* **41** 2505–2536. MR3127873

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

[39] WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689

[40] WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163

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

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

[43] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

[44] ZHONG, P.-S., HU, T. and LI, J. (2015). Tests for coefficients in high-dimensional additive hazard models. *Scand. J. Stat.* **42** 649–664. MR3391684

DEPARTMENT OF STATISTICAL SCIENCE
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853
USA
E-MAIL: yn265@cornell.edu

DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: hanliu@princeton.edu