The Annals of Statistics
2017, Vol. 45, No. 6, 2299–2327
https://doi.org/10.1214/16-AOS1483

A LIKELIHOOD RATIO FRAMEWORK FOR HIGH-DIMENSIONAL SEMIPARAMETRIC REGRESSION

BY YANG NING*, TIANQI ZHAO† AND HAN LIU†

*Cornell University** *and Princeton University*†

We propose a new inferential framework for high-dimensional semiparametric generalized linear models. This framework addresses a variety of challenging problems in high-dimensional data analysis, including incomplete data, selection bias and heterogeneity. Our work has three main contributions: (i) We develop a regularized statistical chromatography approach to infer the parameter of interest under the proposed semiparametric generalized linear model without the need of estimating the unknown base measure function. (ii) We propose a new likelihood ratio based framework to construct post-regularization confidence regions and tests for the low dimensional components of high-dimensional parameters. Unlike existing post-regularization inferential methods, our approach is based on a novel directional likelihood. (iii) We develop new concentration inequalities and normal approximation results for U-statistics with unbounded kernels, which are of independent interest. We further extend the theoretical results to the problems of missing data and multiple datasets inference. Extensive simulation studies and real data analysis are provided to illustrate the proposed approach.

1. Introduction. Modern data are characterized by their high dimensionality, complexity and heterogeneity. More specifically, the datasets usually contain (1) a large number of explanatory variables, (2) complex sampling and missing value schemes due to design or incapability of contacting study subjects and (3) heterogeneity due to the combination of different data sources. To handle these challenges, regularization based methods are proposed. For instance, the L_1-regularized maximum likelihood estimation for linear models is proposed by [36] and the nonconvex penalized maximum likelihood estimation is considered by [11]. During the past decades, these methods enjoy great success in handling high- dimensional data. However, the existing framework is not flexible enough to handle more challenging settings with incomplete data, complex sampling, and multiple heterogeneous datasets. To motivate our study, consider the following two examples.

EXAMPLE 1 (Missing data and selection bias). Given a univariate random variable Y and a d dimensional random vector $\boldsymbol{X}$, assume that Y given $\boldsymbol{X}$ follows

Received December 2015; revised May 2016.

MSC2010 subject classifications. Primary 62E20, 62G20; secondary 62G10.

Key words and phrases. Post-regularization inference, likelihood ratio test, semiparametric sparsity, confidence interval, nonconvex penalty, U-statistics.

from a generalized linear model with the canonical link,

$$p(y \mid \boldsymbol{x}) = \exp\{\boldsymbol{x}^T\boldsymbol{\beta} \cdot y - b(\boldsymbol{x}^T\boldsymbol{\beta}, f) + \log f(y)\}, \tag{1.1}$$

where $\boldsymbol{\beta}$ is a d-dimensional unknown parameter, $f(\cdot)$ is a known base measure function and $b(\cdot,\cdot)$ is a normalizing function. Let $(Y_1, \boldsymbol{X}_1), \dots, (Y_n, \boldsymbol{X}_n)$ denote n independent copies of $(Y, \boldsymbol{X})$. In high-dimensional data analysis, the samples $(Y_1, \boldsymbol{X}_1), \dots, (Y_n, \boldsymbol{X}_n)$ may contain missing values or they are observed after some unknown selection process. To account for the effect of missingness or selection bias, we introduce an indicator variable δ_i, whose value is 1 if $(Y_i, \boldsymbol{X}_i)$ is completely observed or selected, and 0 otherwise. Due to the selection effect, the standard penalized maximum likelihood estimator under model (1.1) with only selected data (i.e., $\delta_i = 1$) is often inconsistent for $\boldsymbol{\beta}$. To account for the missing data and selection bias, we need to develop a new framework to infer the high-dimensional parameter $\boldsymbol{\beta}$.

EXAMPLE 2 (Multiple datasets inference with heterogeneity). Modern datasets are often collected by aggregating multiple data sources. Analysis of such types of data has been studied in the fields of multitask learning in machine learning [1, 22] and seemingly unrelated regression in econometrics [33]. In the multitask learning setting, each dataset corresponds to a learning task. More specifically, assume that the data in the tth task, $t = 1, \dots, T$ are i.i.d. copies of $(Y_{(t)}, \boldsymbol{X}_{(t)})$, which follows from (1.1), that is,

$$p(y_{(t)} \mid \boldsymbol{x}_{(t)}) = \exp\{\boldsymbol{x}_{(t)}^T\boldsymbol{\beta}_t \cdot y_{(t)} - b(\boldsymbol{x}_{(t)}^T\boldsymbol{\beta}_t, f_t) + \log f_t(y_{(t)})\}, \tag{1.2}$$

where $\boldsymbol{\beta}_t$ is a task-specific regression parameter. Most of the existing literature only focuses on the analysis of homogeneous datasets that means $f_t(\cdot) = f(\cdot)$ for any $t = 1, \dots, T$. However, the aggregated data are often highly heterogeneous. For instance, the learning tasks obtained from different areas may contain classification for binary responses as well as regression for continuous and count responses, which implies different forms of $f_t(\cdot)$ in (1.2). Thus, to take into account data heterogeneity, we need a new inferential procedure for $\boldsymbol{\beta}_t$ that does not depend on the knowledge of $f_t(\cdot)$.

To handle the above challenges, we propose a new semiparametric model, which takes the form (1.1) but with both $\boldsymbol{\beta}$ and $f(\cdot)$ as unknown parameters. It naturally handles data with missing values, complex sampling and heterogeneity. This paper contains three major contributions.

Our first contribution is to provide a new regularized statistical chromatography procedure to directly estimate the finite dimensional regression parameter $\boldsymbol{\beta}$ and leave the nonparametric component $f(\cdot)$ as a nuisance. In particular, we model the data at a more refined granularity of rank and order statistics, so that sophisticated conditioning arguments and the structure of exponential family distributions can be exploited to separate the parameter of interest and nuisance component (thus the whole procedure is named “statistical chromatography”). Once the parameter of interest and nuisance parameter are separated, we eliminate the nuisance component to construct a pseudo-likelihood of rank statistics and exploit lower order approximation to speed up computation.

Our second contribution is to develop a new likelihood ratio inferential framework for low-dimensional parameters under the high-dimensional model. In particular, we propose a directional likelihood ratio statistic for hypothesis testing and a maximum directional likelihood estimator for confidence regions in the high-dimensional setting. Compared to the existing post-regularization inferential methods, our procedure has two important features: (1) We allow general regularized estimators including nonconvex regularized estimators and pseudo-likelihood; and (2) We do not need any signal strength assumption for model selection consistency. Our third contribution is to develop new technical tools for studying high-dimensional inference related to U-statistics. First, we prove a concentration inequality in Lemma A.3 for U-statistics with unbounded kernels with subexponential decay. A more general maximal inequality is shown in Lemma F.2 of Supplementary Material [29], which plays the key role to derive improved rates of convergence for multiple datasets inference problems. Second, to apply the central limit theorem for U-statistics, we provide the theoretical justification of the Hájek projection in increasing dimensions for normal approximation. More details are provided in Lemma A.5. These U-statistic results are of independent interest.

Comparison with related works: The proposed model is closely related to the proportional likelihood ratio model [7, 23]. However, unlike their model we do not require the density assumption for the nonparametric function. The proposed estimation procedure is related to the permutation based test [16] and the second-order approximation reduces to the pairwise likelihood considered by [7, 18]. To the best of our knowledge, the proposed estimation method dates back to the original work by [18], in which a pairwise likelihood method is used to eliminate the nonparametric function. We follow their idea and generalize it to the missing data and multitask learning problems. Our investigation mainly focuses on the theoretical properties in high-dimensional regimes, which have not been studied before.

In the literature, a marginal rank likelihood method is proposed to eliminate the nuisance functions in the linear transformation model [30] and the copula model [14]. However, unlike the marginal rank likelihood, our likelihood function can be viewed as a conditional rank likelihood constructed by the conditional rank probability given the order statistics. To handle high-dimensional data with missing values, [34] proposed an expectation-maximization algorithm. When the explanatory variables are missing completely at random (MCAR), Loh and Wainwright [20] developed the theory of a nonconvex optimization approach. Compared with these works, we consider a much broader class of missing data mechanisms.

In the linear models, the estimation, prediction error bounds and variable selection consistency for the L_1-regularized estimator have been well studied by

[5, 6, 24, 27, 42]. More recently, the estimation bounds and oracle properties for the nonconvex regularized estimator are established by [12, 21, 38], among others. In addition to these estimation results, significant progress has been made toward understanding the high-dimensional inference (e.g., constructing confidence intervals or testing hypotheses) under the generalized linear models. Examples include [2–4, 15, 37, 41]. All these procedures lead to asymptotically normally distributed estimators that can be used to construct Wald-type statistics. Other related inferential procedures include the data-splitting method [26, 39], stability selection [25, 32], L_2 confidence set [28] and conditional inference [19, 35]. Under a stronger oracle property, the asymptotic normality of nonconvex estimators is established by [11].

This paper proposes a new directional likelihood based method for constructing confidence regions and testing hypotheses in high dimensions. Compared to the existing work on high-dimensional inference under the generalized linear model, our method and theory are different in the following three aspects. First, our proposed semiparametric model is much more sophisticated than the generalized linear model. In particular, the U-statistic structure due to the statistical chromatography leads to additional technical challenge (see the third contribution above) and requires more refined analysis to control the variability of the estimated nuisance parameters in the proposed directional likelihood function. Second, from the hypothesis testing perspective, our main inferential tool is a new directional likelihood ratio test, whereas the existing methods mainly focus on the Wald or score- type tests. Third, we can conduct the inference based on local solutions of a nonconvex regularized problem, while the method in [37] based on inverting the Karush–Kuhn–Tucker condition may not be directly applicable.

The rest of this paper is organized as follows. In Section 2, we formally define the proposed semiparametric model. In Section 3, we introduce the main ideas of regularized statistical chromatography, along with the directional likelihood based inference for hypothesis tests and confidence regions. In Section 4, we analyze the theoretical properties of the obtained confidence regions and establish the asymptotic distributions of the directional likelihood ratio test statistics. Section 5 contains both simulation and real data analysis results. The last section includes remarks and discussions. The proofs of main results are shown in the Appendix.

Notation: For positive sequences a_n and b_n, we write $a_n \lesssim b_n$, if $a_n/b_n = O(1)$. We denote $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Denote $X_n \rightsquigarrow X$ for some random variable X if X_n converges weakly to X. For $\mathbf{v} = (v_1, \dots, v_d)^T \in \mathbb{R}^d$, 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 $\mathbf{M}$, let $\|\mathbf{M}\|_2$, $\|\mathbf{M}\|_\infty$, $\|\mathbf{M}\|_1$ and $\|\mathbf{M}\|_{L_1}$ be the spectral, elementwise supreme, elementwise L_1 and matrix L_1 norms of $\mathbf{M}$. For two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, we write $\mathbf{M}_1 \le \mathbf{M}_2$ if $\mathbf{M}_2 - \mathbf{M}_1$ is positive semidefinite. For $S \subseteq \{1, \dots, d\}$, let $\mathbf{v}_S = \{v_j : j \in S\}$ and S^c be the complement of S. The gradient and subgradient of a function $f(\boldsymbol{x})$ are denoted by $\nabla f(\boldsymbol{x})$ and $\partial f(\boldsymbol{x})$, respectively.

For a univariate function $f(x)$, its derivative can also be represented by $f'(x)$. Let $\nabla_S f(\boldsymbol{x})$ denote the gradient of $f(\boldsymbol{x})$ with respect to $\boldsymbol{x}_S$. Let $\mathbf{I}_d$ be the d by d identity matrix. Let $\lfloor k \rfloor$ denote the largest integer less than k. Throughout the paper, we use bold letters to denote vectors and matrices and unbold letters to denote scalars. We use the following definition of subexponential random variables.

DEFINITION 1.1. A random variable Y is subexponential if there exist constants $C, C' > 0$, such that $\mathbb{P}(|Y| \geq \delta) \leq C' \exp(-C\delta)$, for any $\delta > 0$.

2. The semiparametric generalized linear model. We first define a semiparametric natural exponential family model, which further leads to the definition of the semiparametric generalized linear model.

DEFINITION 2.1 (Semiparametric natural exponential family). A random variable $Y \in \mathcal{Y} \subseteq \mathbb{R}$ satisfies the semiparametric natural exponential family (spEF) with parameters (θ, f), if its density satisfies

$$p(y; \theta, f) = \exp\{\theta \cdot y - b(\theta, f) + \log f(y)\}, \tag{2.1}$$

where $f(\cdot)$ is an unknown base measure, θ is an unknown canonical parameter, and $b(\theta, f) = \log \int_{\mathcal{Y}} \exp(\theta \cdot y) f(y)\, dy < \infty$ is the log-partition function.

The spEF extends the classical natural exponential family by treating the base measure $f(y)$ as an infinite dimensional parameter. By choosing a suitable base measure, the spEF recovers the whole class of natural exponential family distributions. However, the spEF suffers from the identifiability issue. For instance, spEF(θ, f) is identical to spEF$(\theta, c \cdot f)$, where c is any positive constant. To address this problem, we need to impose some identifiability conditions, such as $f(y_0) = 1$, for some $y_0 \in \mathcal{Y}$, or $\int_{\mathcal{Y}} f(y) \cdot dy = 1$ if $f(y)$ is integrable. Later, we can see that these identifiability conditions will not affect our inference procedures. We now define the semiparametric generalized linear model.

DEFINITION 2.2 (Semiparametric generalized linear model). Given a vector of d-dimensional covariates $\boldsymbol{X} = (X_1, \dots, X_d)^T$ and response $Y \in \mathbb{R}$, assume Y given $\boldsymbol{X}$ follows the semiparametric natural exponential family

$$p(y \mid \boldsymbol{x}) = \exp\{\theta(\boldsymbol{x}) \cdot y - b(\theta(\boldsymbol{x}), f) + \log f(y)\} \quad \text{and} \quad \theta(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}, \tag{2.2}$$

where $b(\cdot, \cdot)$ is the log-partition function and $\boldsymbol{\beta}$ is a d-dimensional parameter. We say that Y given $\boldsymbol{X}$ follows the semiparametric generalized linear model (GLM) with parameters $(\boldsymbol{\beta}, f)$.

Note that we directly set $\theta(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}$ in (2.2), because we implicitly adopt the canonical link, that is, we choose a link function g such that $g^{-1}(\cdot) = b'(\cdot, f)$. Compared with the classical generalized linear models (GLMs), the proposed model contains unknown parameters $\boldsymbol{\beta}$ and $f(\cdot)$, where $\boldsymbol{\beta}$ characterizes the covariate effect, and $f(\cdot)$ determines the distribution in the natural exponential family. For instance, the linear regression with standard Gaussian noise has $f(y) = \exp(-y^2/2)$; the logistic regression has $f(y) = 1$; and the Poisson regression has $f(y) = 1/y!$. Thus, these GLMs are parametric submodels of the semiparametric generalized linear model.

REMARK 1. Some exponential family distributions, such as the normal distribution, involve dispersion parameters. In this case, the semiparametric natural exponential family can be written as

$$p(y; \boldsymbol{\theta}, \tau, f) = \exp\{[\theta \cdot y - b(\theta, f)]/a(\tau) + \log f(y; \tau)\},$$

where $f(\cdot;\cdot)$ is an unknown positive function, θ is the natural parameter, $a(\tau)$ is a known function of the dispersion parameter τ and $b(\theta, f)$ is the log-partition function. Then, with $\theta(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$, the semiparametric generalized linear model reduces to

$$p(y \mid \boldsymbol{x}; \boldsymbol{\beta}, \tau, f) = \exp\{\bar{\boldsymbol{\beta}}^T\boldsymbol{x} \cdot y - \bar{b}(\bar{\boldsymbol{\beta}}^T\boldsymbol{x}, \tau, f) + \log f(y; \tau)\},$$

where $\bar{\boldsymbol{\beta}} = \boldsymbol{\beta}/a(\tau)$ and $\bar{b}(\bar{\boldsymbol{\beta}}^T\boldsymbol{x}, \tau, f) = b(a(\tau)\bar{\boldsymbol{\beta}}^T\boldsymbol{x}, f)/a(\tau)$. Hence, with the new reparametrization $\bar{\boldsymbol{\beta}}$, the proposed model is identical to (2.2), except that we allow $\bar{b}(\cdot)$ and $f(\cdot;\cdot)$ to depend on the dispersion parameter τ. Later, we will see that this dependence does not lead to any extra level of difficulty in terms of inference on $\bar{\boldsymbol{\beta}}$.

The semiparametric generalized linear model has broad applicability to address the challenging problems involving complex and heterogeneous data. In the following, we illustrate how the semiparametric model can be used to handle the missing data and selection bias problems in Example 1 and heterogeneous multitask learning problem in Example 2.

Revisit of Example 1: Missing data and selection bias. Recall that Y_i given $\boldsymbol{X}_i$ follows the GLM in (1.1) and we are interested in making inference on $\boldsymbol{\beta}$. To account for the missing data and selection effect, we assume that the selection indicator δ_i given Y_i and $\boldsymbol{X}_i$ satisfies the following decomposable selection model.

DEFINITION 2.3 (Decomposable selection model). The missing data or selection model is decomposable, if there exist two nonnegative functions $g_1(\cdot)$ and $g_2(\cdot)$ such that $\mathbb{P}(\delta_i = 1 \mid Y_i, \boldsymbol{X}_i) = g_1(Y_i) \cdot g_2(\boldsymbol{X}_i)$, where $\int g_1(y) \cdot dy = 1$ and $\int g_2(\boldsymbol{x}) \cdot d\boldsymbol{x} = 1$.

Under the assumption of MCAR, the missing data model satisfies $\mathbb{P}(\delta_i = 1|Y_i, \boldsymbol{X}_i) = \mathbb{P}(\delta_i = 1)$, which implies that MCAR is decomposable. Indeed, the decomposable model is much more general. Consider the following partition of covariates $\boldsymbol{X}_i = (\boldsymbol{X}_{io}, \boldsymbol{X}_{im})$, and assume that $(Y_i, \boldsymbol{X}_{im})$ are subject to missingness. It is seen that the missing at random (MAR), defined by $\mathbb{P}(\delta_i = 1|Y_i, \boldsymbol{X}_i) = \mathbb{P}(\delta_i = 1|\boldsymbol{X}_{io})$, is also decomposable. So is the outcome dependent sampling model [17]. In addition, the decomposable model can be missing not at random (MNAR). For instance, if Y_i is subject to missingness and the missing mechanism only depends on the potentially unobserved value of Y_i, then the missing data pattern is not at random but is still decomposable. Thus, the decomposable selection model is a very flexible nonparametric model for missing data and selection bias. In general, the functions $g_1(\cdot)$ and $g_2(\cdot)$ may not be identifiable. Later, we will see that this nonidentifiability issue can be handled by using the proposed method.

To specify the likelihood based on the selected data, we derive the probability density function of Y_i given $\boldsymbol{X}_i$ and $\delta_i = 1$. Using the Bayes formula,

$$p(y_i \mid \boldsymbol{x}_i, \delta_i = 1) = \mathbb{P}(\delta_i = 1 \mid y_i, \boldsymbol{x}_i) \cdot p(y_i \mid \boldsymbol{x}_i)/T_i(\boldsymbol{x}_i),$$

where $T_i(\boldsymbol{x}_i) = \int \mathbb{P}(\delta_i = 1 \mid y_i, \boldsymbol{x}_i) p(y_i \mid \boldsymbol{x}_i)\, dy_i$ and $(y_i, \boldsymbol{x}_i)$ is the observed value of $(Y_i, \boldsymbol{X}_i)$. Under the generalized linear model in (1.1) and the decomposable selection model, we obtain

$$p(y_i \mid \boldsymbol{x}_i, \delta_i = 1) = \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta} \cdot y_i - b(\boldsymbol{x}_i^T \boldsymbol{\beta}, f^m) + \log f^m(y_i)\}, \tag{2.3}$$

where $f^m(y) = g_1(y)f(y)$. Hence, if Y_i given $\boldsymbol{X}_i$ follows the GLM (1.1) or more generally the semiparametric version (2.2) and the selection model is decomposable, then Y_i given $\boldsymbol{X}_i$ and $\delta_i = 1$ satisfies (2.2) with the same unknown parameter $\boldsymbol{\beta}$ and the unknown based measure $f^m(y) = g_1(y)f(y)$. We call this the invariance property of semiparametric GLMs under the decomposable selection model. Hence, the inference on $\boldsymbol{\beta}$ with missing data and selection bias is equivalent to the inference problem under the semiparametric GLM (2.2).

Revisit of Example 2: Multiple datasets inference with heterogeneity. In Example 2 of Section 1, to take into account of data heterogeneity, we can assume that the based measure function $f_t(\cdot)$ is a task-specific unknown function. Thus, the multiple datasets inference with heterogeneity can be handled by the semiparametric GLM framework, and an inferential method that is invariant to $f(\cdot)$ under the model (2.2) is needed.

3. Semiparametric inference.

In this section, we consider how to construct confidence intervals and perform hypothesis tests for a single component of $\boldsymbol{\beta}$ under the semiparametric GLM. The extension to the confidence regions and tests for multidimensional components of $\boldsymbol{\beta}$ is standard and is deferred to the Supplementary Material [29].

3.1. *Regularized statistical chromatography.*

Due to the presence of the unknown function $f(\cdot)$, the likelihood of the semiparametric GLM is complicated, making likelihood based inference of $\boldsymbol{\beta}$ intractable. To handle this problem, we propose a new procedure called statistical chromatography to extract information on $\boldsymbol{\beta}$.

For $i = 1, \dots, n$, suppose that the data $(Y_i, \boldsymbol{X}_i)$ are i.i.d. By the discriminative modeling approach, the probability distribution of the data is $p(\boldsymbol{y}, \mathbf{x}; \boldsymbol{\beta}, f) = p(\boldsymbol{y} \mid \mathbf{x}; \boldsymbol{\beta}, f) \cdot p(\mathbf{x})$, where $\boldsymbol{y} = (y_1, \dots, y_n)$ and $\mathbf{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ are the observed values of $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ and $\mathbf{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$. Since the marginal distribution of $\mathbf{X}$ does not involve $\boldsymbol{\beta}$ or f, we only focus on the first conditional distribution $p(\boldsymbol{y} \mid \mathbf{x}; \boldsymbol{\beta}, f)$. However, its dependence on $\boldsymbol{\beta}$ and f is still intertwined and the inference on $\boldsymbol{\beta}$ is hindered by the nuisance parameter f. To tackle this problem, we need to further separate the parameters $\boldsymbol{\beta}$ and f in the conditional likelihood. To this end, we decompose $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ into $\mathbf{R} = (R_1, \dots, R_n)$ and $\boldsymbol{Y}_{(\cdot)} = (Y_{(1)}, \dots, Y_{(n)})$, which denote the rank and order statistics of $\boldsymbol{Y}$, respectively. Let $\boldsymbol{r}$ and $\boldsymbol{y}_{(\cdot)}$ denote the observed values of $\mathbf{R}$ and $\boldsymbol{Y}_{(\cdot)}$, respectively. Thus, we have

$$p(\boldsymbol{y} \mid \mathbf{x}; \boldsymbol{\beta}, f) = \mathbb{P}(\mathbf{R} = \boldsymbol{r} \mid \mathbf{x}, \boldsymbol{y}_{(\cdot)}; \boldsymbol{\beta}) \cdot p(\boldsymbol{y}_{(\cdot)} \mid \mathbf{x}; \boldsymbol{\beta}, f), \tag{3.1}$$

where by the definition of conditional probabilities we can show that

$$\begin{aligned} \mathbb{P}(\mathbf{R} = \mathbf{r} \mid \mathbf{x}, \boldsymbol{y}_{(\cdot)}; \boldsymbol{\beta}) &= \frac{\prod_{i=1}^n p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}, f)}{\sum_{\pi \in \Xi} \prod_{i=1}^n p(y_{\pi(i)} \mid \boldsymbol{x}_i; \boldsymbol{\beta}, f)} \\ &= \frac{\exp(\sum_{i=1}^n \boldsymbol{\beta}^T \boldsymbol{x}_i \cdot y_i)}{\sum_{\pi \in \Xi} \exp(\sum_{i=1}^n \boldsymbol{\beta}^T \boldsymbol{x}_i \cdot y_{\pi(i)})}, \end{aligned} \tag{3.2}$$

where Ξ is the set of all one-to-one maps from $\{1, \dots, n\}$ to $\{1, \dots, n\}$. The intuition behind the data decomposition is that the rank statistic given the order statistic has no information on f. Mathematically, the product $\prod_{i=1}^n f(y_i)$ appearing in both numerator and denominator of (3.2) only depends on $\boldsymbol{Y}_{(\cdot)}$ and is eliminated. Since we separate parameters $\boldsymbol{\beta}$ and f at a more refined granularity of rank and order statistics, we call this procedure as statistical chromatography.

Given the chromatography decomposition in (3.1), one may opt to only keep the conditional probability (3.2) for estimation and inference of $\boldsymbol{\beta}$. However, the probability in (3.2) is computationally intensive due to the combinatorial nature of permutations. To this end, we consider a surrogate of $\mathbb{P}(\mathbf{R} = \boldsymbol{r} \mid \mathbf{x}, \boldsymbol{y}_{(\cdot)}; \boldsymbol{\beta})$ using the kth order information. For notational simplicity, we only present $k = 2$, and leave the discussion for $k > 2$ to the Supplementary Material [29]. For any i and j, let $\mathbf{R}_{ij}^L$ denote the local rank statistic of Y_i and Y_j among the pair (Y_i, Y_j) [i.e., $\mathbf{R}_{ij}^L = (1, 2)$ or $(2, 1)$]. Instead of considering the full conditional probability in (3.2), we study the product of all possible combinations of the local rank conditional probability,

$$\begin{aligned} &\prod_{i<j} \mathbb{P}(\mathbf{R}_{ij}^L = \boldsymbol{r}_{ij}^L \mid \boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{y}_{(i,j)}^L; \boldsymbol{\beta}) \\ &\quad = \prod_{i<j} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i y_i + \boldsymbol{\beta}^T \boldsymbol{x}_j y_j)}{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i y_i + \boldsymbol{\beta}^T \boldsymbol{x}_j y_j) + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i y_j + \boldsymbol{\beta}^T \boldsymbol{x}_j y_i)}, \end{aligned} \tag{3.3}$$

where $\boldsymbol{Y}_{(i,j)}^{L} = (\min(Y_i, Y_j), \max(Y_i, Y_j))$, and $\boldsymbol{y}_{(i,j)}^{L}$ and $\boldsymbol{r}_{ij}^{L}$ are the observed values of $\boldsymbol{Y}_{(i,j)}^{L}$ and $\mathbf{R}_{ij}^{L}$, respectively. Applying the logarithmic transformation to (3.3), we obtain the function

$$\ell(\boldsymbol{\beta}) = -\binom{n}{2}^{-1} \sum_{1\le i<j\le n} \log\bigl(1 + R_{ij}(\boldsymbol{\beta})\bigr), \tag{3.4}$$

where $R_{ij}(\boldsymbol{\beta}) = \exp\{-(y_i - y_j) \cdot \boldsymbol{\beta}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\}$. It is also known as the pairwise log-likelihood, which has been considered by [7, 8, 18]. In high dimensions, we may add a regularization term to $\ell(\boldsymbol{\beta})$, which leads to the regularized chromatography approach.

3.2. *Confidence interval and hypothesis test*: *A likelihood ratio approach*. Given the composite log-likelihood (3.4), we consider the problem of testing a pre-specified component of $\boldsymbol{\beta}$. Without loss of generality, assume that $\boldsymbol{\beta}$ can be partitioned as $\boldsymbol{\beta} = (\alpha, \boldsymbol{\gamma}^T)^T$, where $\alpha \in \mathbb{R}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{d-1}$. Now, we consider the null hypothesis $H_0 : \alpha = \alpha_0$, and treat $\boldsymbol{\gamma}$ as a $(d-1)$-dimensional nuisance parameter. Let $\boldsymbol{\beta}^*$ be the true value of $\boldsymbol{\beta}$. It is well known that the classical likelihood ratio test is not directly applicable to testing the null hypothesis H_0, when the nuisance parameter $\boldsymbol{\gamma}$ is high dimensional. In what follows, we propose a new directional likelihood function and the corresponding likelihood ratio test for H_0, which provides valid inferential results in high-dimensional settings.

Specifically, we define the directional likelihood function for α as

$$\widehat{\ell}(\alpha) = \ell\bigl(\alpha, \widehat{\boldsymbol{\gamma}} + (\widehat{\alpha} - \alpha)\widehat{\mathbf{w}}\bigr), \tag{3.5}$$

where $\widehat{\boldsymbol{\beta}} := (\widehat{\alpha}, \widehat{\boldsymbol{\gamma}})$ is an initial estimator for $\boldsymbol{\beta}^*$, and $\widehat{\mathbf{w}}$ is an estimator for

$$\mathbf{w}^{*T} := \mathbf{H}_{\alpha\boldsymbol{\gamma}}(\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}})^{-1} \in \mathbb{R}^{d-1} \qquad \text{where } \mathbf{H} = -\mathbb{E}\bigl\{\nabla^2\ell(\boldsymbol{\beta}^*)\bigr\}. \tag{3.6}$$

Here, the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{w}}$ will be introduced later and $\mathbf{H}_{\alpha\boldsymbol{\gamma}}$ and $\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ are the corresponding partitions of $\mathbf{H}$. Later, we can show that the directional likelihood function $\widehat{\ell}(\alpha)$ can be treated as a standard likelihood function for a single unknown parameter α. For instance, we define the maximum directional likelihood estimator as

$$\widehat{\alpha}^P = \operatorname*{argmax}_{\alpha\in\mathbb{R}} \widehat{\ell}(\alpha). \tag{3.7}$$

To test the null hypothesis $H_0 : \alpha^* = \alpha_0$, we define the maximum directional likelihood ratio test (DLRT) statistic as

$$\Lambda_n = 2n\bigl\{\widehat{\ell}\bigl(\widehat{\alpha}^P\bigr) - \widehat{\ell}(\alpha_0)\bigr\}. \tag{3.8}$$

In the following, we explain the intuition behind the directional likelihood (3.5) based on the geometry of submodels in the semiparametric literature and the orthogonality property for nuisance parameters. We note that a similar orthogonality property has been used by [3, 4] for the post-selection inference. We leave the detailed comparison and discussion to Remark 2.

Given the likelihood function $\ell(\boldsymbol{\beta})$, we consider a parametrization for a surface $S \subset \mathbb{R}^{d+1}$, in which the coordinates of points can be expressed as $(\boldsymbol{\beta}, \ell(\boldsymbol{\beta})) \in \mathbb{R}^{d+1}$. Consider two smooth functions $\alpha(\cdot) \in \mathbb{R}$ and $\boldsymbol{\gamma}(\cdot) \in \mathbb{R}^{d-1}$, satisfying $\alpha(0) = \alpha^*$, $\alpha'(0) \neq 0$ and $\boldsymbol{\gamma}(0) = \boldsymbol{\gamma}^*$. Define a smooth curve $\boldsymbol{\delta} : I \to \mathbb{R}^{d+1}$, which maps $t \in I$ to $(\alpha(t), \boldsymbol{\gamma}(t), \ell_c(t))$, where I is an interval in $\mathbb{R}$ containing a small neighborhood of 0 and $\ell_c(t) = \ell(\alpha(t), \boldsymbol{\gamma}(t))$. Note that the curve $\boldsymbol{\delta}$ is within the surface S and passes through the true values $(\alpha^*, \boldsymbol{\gamma}^*, \ell(\boldsymbol{\beta}^*))$ when $t = 0$. Since the curve $\boldsymbol{\delta}$ is determined by the form of $(\alpha(t), \boldsymbol{\gamma}(t))$, we need to decide how to choose $(\alpha(t), \boldsymbol{\gamma}(t))$ such that the likelihood $\ell_c(t)$ along the curve has desired properties. Taking the derivative with respect to t, the score function of $\ell_c(t)$ at $t = 0$ is given by

$$S_c(\alpha^*, \boldsymbol{\gamma}^*) := \frac{d\ell_c(t)}{dt}\bigg|_{t=0} = \alpha'(0) \cdot \nabla_\alpha \ell(\alpha^*, \boldsymbol{\gamma}^*) + [\boldsymbol{\gamma}'(0)]^T \cdot \nabla_{\boldsymbol{\gamma}} \ell(\alpha^*, \boldsymbol{\gamma}^*).$$

To construct a valid test statistic, the key insight is to ensure that $S_c(\alpha, \boldsymbol{\gamma})$ is robust to the perturbation of the high-dimensional nuisance parameter $\boldsymbol{\gamma}$. Mathematically, we require the following orthogonality property, that is, $\mathbb{E}[\nabla_{\boldsymbol{\gamma}} S_c(\alpha^*, \boldsymbol{\gamma}^*)] = 0$; see Remark 2 for further discussion. This implies $\alpha'(0)\mathbf{H}_{\alpha\boldsymbol{\gamma}} + [\boldsymbol{\gamma}'(0)]^T \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} = 0$, which is equivalent to $\boldsymbol{\gamma}'(0)/\alpha'(0) = -\mathbf{w}^*$ by (3.6). Thus, for t in a small neighborhood of 0, the Taylor theorem implies

$$\alpha(t) = \alpha^* + \alpha'(0)t + o(t) \quad \text{and} \quad \gamma_j(t) = \gamma_j^* - \alpha'(0)w_j^* t + o(t),$$

where $1 \le j \le d - 1$. Ignoring the higher order terms, this gives $\ell_c(t) = \ell(\alpha^* + \alpha'(0)t, \boldsymbol{\gamma}^* - \alpha'(0)\mathbf{w}^* t)$. Finally, a reparametrization of $\ell_c(t)$ with $\alpha := \alpha^* + \alpha'(0)t$ yields a function $\bar{\ell}_c(\alpha)$ of α, defined as

$$\bar{\ell}_c(\alpha) := \ell_c\left(\frac{\alpha - \alpha^*}{\alpha'(0)}\right) = \ell(\alpha, \boldsymbol{\gamma}^* + (\alpha^* - \alpha)\mathbf{w}^*).$$

Replacing α^*, $\boldsymbol{\gamma}^*$ and $\mathbf{w}^*$ by the corresponding estimators $\widehat{\alpha}$, $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\mathbf{w}}$, the function $\bar{\ell}_c(\alpha)$ becomes the directional likelihood in (3.5). This gives the geometric intuition on how the directional likelihood is derived. When $\ell(\boldsymbol{\beta})$ is the log-likelihood function, the curve $(\alpha(t), \boldsymbol{\gamma}(t))$ corresponds to the least favorable curve up to a reparametrization [31].

Next, we consider how to obtain estimators $\widehat{\alpha}$, $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\mathbf{w}}$ in the directional likelihood (3.5). To estimate $\boldsymbol{\beta}^*$, our proposed framework allows a wide class of estimators $\widehat{\boldsymbol{\beta}} = (\widehat{\alpha}, \widehat{\boldsymbol{\gamma}})$ including the regularized estimators with nonconvex (or folded concave) penalty functions; see Remark 3. To estimate the $(d - 1)$-dimensional vector $\mathbf{w}^*$, we use the following Lasso-type estimator:

$$\widehat{\mathbf{w}} = \arg\max_{\mathbf{w}} \left\{ \frac{1}{2}\mathbf{w}^T \nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \ell(\widehat{\boldsymbol{\beta}})\mathbf{w} - \mathbf{w}^T \nabla^2_{\boldsymbol{\gamma}\alpha} \ell(\widehat{\boldsymbol{\beta}}) - \lambda_1 \|\mathbf{w}\|_1 \right\}, \tag{3.9}$$

where $\lambda_1 \geq 0$ is a tuning parameter.

To analyze the semiparametric GLM, one technical challenge is that $\nabla\ell(\boldsymbol{\beta})$ is a high-dimensional U-statistic with a possibly unbounded kernel function, that is,

$$\nabla\ell(\boldsymbol{\beta}) = \frac{2}{n(n-1)} \cdot \sum_{1\leq i<j\leq n} \frac{R_{ij}(\boldsymbol{\beta}) \cdot (y_i - y_j) \cdot (\boldsymbol{x}_i - \boldsymbol{x}_j)}{1 + R_{ij}(\boldsymbol{\beta})}.$$

To decouple the correlation between summands in $\nabla\ell(\boldsymbol{\beta})$, we resort to the Hájek projection [13] and define

$$\widehat{\mathbf{U}}_n = \frac{2}{n}\sum_{i=1}^n \mathbf{g}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}^*) \qquad \text{where } \mathbf{g}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{n}{2} \cdot \mathbb{E}\{\nabla\ell(\boldsymbol{\beta}) \mid y_i, \boldsymbol{x}_i\}. \tag{3.10}$$

By definition, $2n^{-1}\mathbf{g}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}^*)$ is the projection of $\nabla\ell(\boldsymbol{\beta}^*)$ onto the σ-field generated by $(y_i, \boldsymbol{x}_i)$, and we sum over all samples to construct $\widehat{\mathbf{U}}_n$. We therefore approximate the U-statistic $\nabla\ell(\boldsymbol{\beta}^*)$ by the sum of independent random variables $\widehat{\mathbf{U}}_n$. Let $\boldsymbol{\Sigma} = \mathbb{E}\{(\mathbf{g}_i^*)^{\otimes 2}\}$ denote the variance of $\mathbf{g}_i^*$, where $\mathbf{g}_i^* = \mathbf{g}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}^*)$. In Theorem 4.1, we prove

$$n^{1/2} \cdot (\widehat{\alpha}^P - \alpha^*) \rightsquigarrow N(0, 4 \cdot \sigma^2 \cdot H^{-2}_{\alpha|\boldsymbol{\gamma}}),$$

where $\sigma^2 = \Sigma_{\alpha\alpha} - 2\mathbf{w}^{*T}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\alpha} + \mathbf{w}^{*T}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathbf{w}^*$, $H_{\alpha|\boldsymbol{\gamma}} = H_{\alpha\alpha} - \mathbf{H}_{\alpha\boldsymbol{\gamma}}\mathbf{H}^{-1}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathbf{H}_{\boldsymbol{\gamma}\alpha}$ and $\Sigma_{\alpha\alpha}$, $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\alpha}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ are corresponding partitions of $\boldsymbol{\Sigma}$. To construct confidence intervals and Wald-type hypothesis test, one needs to estimate the asymptotic variance, which depends on the unknown covariance and Hessian matrices $\boldsymbol{\Sigma}$ and $\mathbf{H}$. By exploiting the U-statistic structure of $\nabla\ell(\boldsymbol{\beta})$, we can estimate $\boldsymbol{\Sigma}$ by

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \cdot \sum_{i=1}^n \left\{ \frac{1}{n-1} \sum_{j=1, j\neq i}^n \frac{R_{ij}(\widehat{\boldsymbol{\beta}}) \cdot (y_i - y_j) \cdot (\boldsymbol{x}_i - \boldsymbol{x}_j)}{1 + R_{ij}(\widehat{\boldsymbol{\beta}})} \right\}^{\otimes 2}. \tag{3.11}$$

Thus, we define $\widehat{\sigma}^2 = \widehat{\Sigma}_{\alpha\alpha} - 2\widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}\alpha} + \widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\widehat{\mathbf{w}}$. Moreover, we can estimate $H_{\alpha|\boldsymbol{\gamma}}$ by $\widehat{H}_{\alpha|\boldsymbol{\gamma}} = -\nabla^2_{\alpha\alpha}\ell(\widehat{\boldsymbol{\beta}}) + \widehat{\mathbf{w}}^T\nabla^2_{\boldsymbol{\gamma}\alpha}\ell(\widehat{\boldsymbol{\beta}})$. Therefore, a two-sided confidence interval for α^* with $(1-\xi)$ coverage probability is given by $[\widehat{\alpha}^P - \zeta n^{-1/2}, \widehat{\alpha}^P + \zeta n^{-1/2}]$, where $\zeta = 2\widehat{\sigma}\widehat{H}^{-1}_{\alpha|\boldsymbol{\gamma}}\Phi^{-1}(1-\xi/2)$.

In addition, to test the null hypothesis $H_0 : \alpha^* = \alpha_0$, Theorem 4.2 shows that the maximum directional likelihood ratio test statistic Λ_n in (3.8) satisfies $(4\sigma^2)^{-1}H_{\alpha|\boldsymbol{\gamma}}\Lambda_n \rightsquigarrow \chi^2_1$. Hence our test with the significance level ξ is

$$\psi_{\mathrm{DLRT}}(\xi) = \mathbb{1}\{(4 \cdot \widehat{\sigma}^2)^{-1} \cdot \widehat{H}_{\alpha|\boldsymbol{\gamma}} \cdot \Lambda_n \geq \chi^2_{1\xi}\}, \tag{3.12}$$

where $\chi^2_{1\xi}$ is the $(1-\xi)$th quantile of a χ^2_1 random variable. The null hypothesis is rejected if and only if $\psi_{\mathrm{DLRT}}(\xi) = 1$, and the associated p-value is given by

$P_{\mathrm{DLRT}} = 1 - \chi_1^2((4\widehat{\sigma}^2)^{-1}\widehat{H}_{\alpha|\boldsymbol{\gamma}}\Lambda_n)$, where $\chi_1^2(\cdot)$ is the c.d.f. of a chi-squared distribution with degree of freedom 1. In Corollary 4.2, we prove that the proposed test can control the type I error asymptotically, that is, $\lim_{n\to\infty}\mathbb{P}(\psi_{\mathrm{DLRT}}(\xi) = 1 \mid H_0) = \xi$ and the p-value is asymptotically uniformly distributed, that is, $P_{\mathrm{DLRT}} \rightsquigarrow$ Uniform[0, 1], under H_0.

REMARK 2 (Orthogonality condition). Recall that the orthogonality condition plays an important role in deciding the direction of the curve $\boldsymbol{\delta}$ at $t = 0$ in our geometric interpretation. Under the GLM and the median regression, [4] and [3] developed an alternative method based on a similar orthogonality property, called immunization, to perform post-selection inference. For instance, in the context of the logistic regression model, the key idea of [4] is to construct an instrument $z_i = z(\boldsymbol{x}_i) \in \mathbb{R}$ such that the orthogonality condition $\nabla_{\boldsymbol{\gamma}}\mathbb{E}[\{y_i - G(\boldsymbol{\beta}^T\boldsymbol{x}_i)\}z_i] = 0$ holds, where $G(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$. Their test statistic for $H_0 : \alpha^* = \alpha_0$ is given by $T_n = n^{-1}\sum_{i=1}^n\{y_i - G(\widehat{\boldsymbol{\beta}}_0^T\boldsymbol{x}_i)\}\widehat{z}_i$, where $\widehat{\boldsymbol{\beta}}_0 = (\alpha_0, \widehat{\boldsymbol{\gamma}})$ for some regularized estimator $\widehat{\boldsymbol{\gamma}}$ and $\widehat{z}_i$ is an estimate of z_i. They proved that under regularity conditions $n^{1/2}T_n$ is asymptotically normal with mean 0 and the variance can be consistently estimated. Our likelihood ratio method is different in the following two aspects. First, while our procedure also relies on a similar orthogonality condition, we do not explicit construct the instrumental variable z_i in our testing procedure. Second, our test statistic is different. Namely, their test statistic T_n is based on the sample version of the moment condition $\mathbb{E}[\{y_i - G(\boldsymbol{\beta}^T\boldsymbol{x}_i)\}z_i] = 0$, whereas our test statistic Λ_n in (3.8) is based on the ratio of the directional likelihood.

4. Main results.

We first prove the asymptotic normality of the maximum directional likelihood estimator $\widehat{\alpha}^P$ in (3.7). We then derive the limiting distribution of Λ_n as well as the validity of the maximum directional likelihood ratio test in (3.12) under the null hypothesis $H_0 : \alpha^* = \alpha_0$.

In the following, we present some regularity conditions. Recall that we define $\mathbf{g}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}^*)$ and $\mathbf{H}$ in (3.10) and (3.6), respectively. Denote

$$\boldsymbol{\Sigma} = \mathbb{E}\{\mathbf{g}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}^*)^{\otimes 2}\}, \qquad H_{\alpha|\boldsymbol{\gamma}} = H_{\alpha\alpha} - \mathbf{H}_{\alpha\boldsymbol{\gamma}}\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\mathbf{H}_{\boldsymbol{\gamma}\alpha}.$$

ASSUMPTION 4.1. Assume that Y is subexponential which satisfies Definition 1.1, and $\|\boldsymbol{X}\|_\infty \le m$ for a positive constant m. Assume that $c \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le c'$, and $c \le \lambda_{\min}(\mathbf{H}) \le \lambda_{\max}(\mathbf{H}) \le c'$, for some constants $c, c' > 0$.

It is easily seen that the subexponential condition holds for most commonly used GLMs in practice. Following [37], we assume the bounded covariates for simplicity. It can be easily relaxed to the sub-Gaussian or subexponential assumptions. Note that $\boldsymbol{\Sigma}$ can be interpreted as the second moment of the Hájek projection, which approximates the asymptotic variance of $\nabla\ell(\boldsymbol{\beta}^*)$, and $H_{\alpha|\boldsymbol{\gamma}}$ is known as the partial information matrix for α in the literature. This condition assumes that the eigenvalues of $\mathbf{\Sigma}$ and $\mathbf{H}$ are lower and upper bounded by positive constants. They are standard regularity conditions even for low dimensional models.

The following main theorem establishes the asymptotic normality of the maximum directional likelihood estimator $\widehat{\alpha}^P$. Let $s = \|\boldsymbol{\beta}^*\|_0$ and $s_1 = \|\mathbf{w}^*\|_0$, where $\mathbf{w}^*$ is defined in (3.6).

THEOREM 4.1. *Under the semiparametric GLM in Definition 2.2 and Assumption 4.1, assume that $\widehat{\boldsymbol{\beta}}$ satisfies $\|\widehat{\mathbf{\Delta}}\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{s\log d/n})$, $\|\widehat{\mathbf{\Delta}}\|_1 = \mathcal{O}_{\mathbb{P}}(s\sqrt{\log d/n})$, and $|\widehat{\mathbf{\Delta}}^T\nabla^2\ell(\boldsymbol{\beta}^*)\widehat{\mathbf{\Delta}}| = \mathcal{O}_{\mathbb{P}}(s\log d/n)$, where $\widehat{\mathbf{\Delta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Given any small constant $\delta > 0$, it holds*

$$\lim_{n\to\infty} \frac{\max\{s, s_1\}^2 \cdot \log d}{n^{1/2-\delta}} = 0. \tag{4.1}$$

Then with $\lambda_1 \asymp \log n \cdot \sqrt{\log d/n}$, we have

$$n^{1/2}(\widehat{\alpha}^P - \alpha^*) \rightsquigarrow N(0, 4\sigma^2 H_{\alpha|\boldsymbol{\gamma}}^{-2})$$

$$\text{where } \sigma^2 = \Sigma_{\alpha\alpha} - 2\mathbf{w}^{*T}\mathbf{\Sigma}_{\boldsymbol{\gamma}\alpha} + \mathbf{w}^{*T}\mathbf{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathbf{w}^*.$$

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

Our condition (4.1) essentially requires that $\mathbf{w}^*$ and $\boldsymbol{\beta}^*$ are sufficiently sparse such that the estimation errors of $\mathbf{w}^*$ and $\boldsymbol{\beta}^*$ and the approximation error in the Hájek projection are controllable. Similarly, under the GLM, [37] assumed that the inverse of the Fisher information matrix $\mathbf{\Omega} = \mathbf{H}^{-1}$ is sparse. Let $\mathbf{\Omega}_{*\alpha}$ and $\mathbf{\Omega}_{*\boldsymbol{\gamma}}$ denote the columns of $\mathbf{\Omega}$ corresponding to α and $\boldsymbol{\gamma}$. To see the connections, consider the following block matrix inverse formula, $\mathbf{\Omega}_{*\alpha} = H_{\alpha|\boldsymbol{\gamma}}^{-1}(1, -\mathbf{H}_{\alpha\boldsymbol{\gamma}}\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1})^T$, where $H_{\alpha|\boldsymbol{\gamma}} = H_{\alpha\alpha} - \mathbf{H}_{\alpha\boldsymbol{\gamma}}\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\mathbf{H}_{\boldsymbol{\gamma}\alpha}$. Since $\mathbf{w}^* = \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\mathbf{H}_{\boldsymbol{\gamma}\alpha}$, we have $\|\mathbf{w}^*\|_0 = \|\mathbf{\Omega}_{*\alpha}\|_0 - 1$. Hence, our sparsity assumption on $\mathbf{w}^*$ is implied by the sparsity of $\mathbf{\Omega}$. Moreover, our results reveal that the sparsity of $\mathbf{\Omega}_{*\boldsymbol{\gamma}}$ is not needed for the inference on α.

Under the GLM, [37] and [4] imposed the condition that $\max\{s, s_1\}^2 \cdot \log^k d = o(n)$ for some constant $k > 0$, which is weaker than our condition (4.1). This is mostly due to the technical differences between the composite likelihood derived by the chromatography approach (which has a U-statistic structure) and the likelihood of the generalized linear model.

We also note that, the rate of λ_1 agrees with the conventional $\sqrt{\log d/n}$ rate for tuning parameters up to a $\log n$ factor, due to the subexponential tail of the response variable Y. In particular, if Y is bounded (e.g., 0–1 binary response), the $\log n$ factor can be eliminated so that we have $\lambda_1 \asymp \sqrt{\log d/n}$.

It is seen that our assumptions do not contain any type of minimal signal strength condition on the nonzero components of $\boldsymbol{\beta}^*$. Therefore, unlike the oracle-type results in [11], variable selection consistency is not a priori for our approach and a valid p-value can be produced even if a covariate is not selected in the model.

REMARK 3 (Estimation consistency). Note that Theorem 4.1 requires that the initial estimator $\widehat{\boldsymbol{\beta}}$ satisfies $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{s\log d/n})$, $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s\sqrt{\log d/n})$ and $|(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^T\nabla^2\ell(\boldsymbol{\beta}^*)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)| = \mathcal{O}_{\mathbb{P}}(s\log d/n)$. In high-dimensional settings, we can estimate $\boldsymbol{\beta}$ by maximizing the following penalized composite likelihood function with a generic penalty function $p_\lambda(\cdot)$:

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{argmax}_{\boldsymbol{\beta}\in\mathbb{R}^d}\left\{\ell(\boldsymbol{\beta}) - \sum_{j=1}^{d} p_\lambda(\beta_j)\right\}, \tag{4.2}$$

where $\lambda \geq 0$ is a tuning parameter. In GLMs, [12, 21, 38] showed such conditions hold. We prove that the same conclusion holds for $\widehat{\boldsymbol{\beta}}$ under the semiparametric GLM. To save space, we leave the detailed analysis of the finite sample estimation error bound of $\widehat{\boldsymbol{\beta}}$ with both Lasso penalty and the nonconvex penalty to the Supplementary Material [29]. Here, we emphasize that our inferential framework allows general regularized estimators such as nonconvex penalty functions. Thus, it is more flexible than [37] based on inverting the Karush–Kuhn–Tucker condition for the Lasso estimator.

To apply Theorem 4.1 to construct confidence intervals, one needs to estimate the asymptotic variance $\sigma^2 H_{\alpha|\boldsymbol{\gamma}}^{-2}$, which depends on the unknown covariance matrix $\boldsymbol{\Sigma}$ and $H_{\alpha|\boldsymbol{\gamma}}$. Recall that such an estimator $\widehat{\boldsymbol{\Sigma}}$ is given in (3.11). The following corollary justifies the validity of the confidence interval.

COROLLARY 4.1. *Under the conditions in Theorem* 4.1, *the confidence interval*

$$\mathrm{CI}_\xi = \left\{\alpha\in\mathbb{R} : \left|\alpha - \widehat{\alpha}^P\right| \leq 2\cdot\widehat{\sigma}\cdot\widehat{H}_{\alpha|\boldsymbol{\gamma}}^{-1}\cdot\Phi^{-1}(1-\xi/2)/n^{1/2}\right\}$$

has the asymptotic coverage $1-\xi$, *that is*, $\lim_{n\to\infty}\mathbb{P}(\alpha^*\in\mathrm{CI}_\xi) = 1-\xi$.

PROOF. A detailed proof is shown in the Supplementary Material [29]. □

We note that the estimator $\widehat{\alpha}^P$ is not semiparametrically efficient, because not all information about $\boldsymbol{\beta}$ is retained in the statistical chromatography. Our numerical results seem to suggest that $\widehat{\alpha}^P$ is nearly as efficient as the estimator under the classical generalized linear model. Thus, our method gains model flexibility and computational efficiency without paying much price on the information loss.

Next, we prove the asymptotic distribution of the test statistic Λ_n and the validity of the maximum likelihood ratio test under the same conditions in Theorem 4.1 and Corollary 4.1.

THEOREM 4.2. *Under the conditions in Theorem* 4.1 *and* $\alpha^* = \alpha_0$, *then*

$$(4\cdot\sigma^2)^{-1}\cdot H_{\alpha|\boldsymbol{\gamma}}\cdot\Lambda_n \rightsquigarrow \chi_1^2.$$

PROOF. A detailed proof is shown in the Supplementary Material [29]. □

As before, to apply the theorem in practice, we replace σ^2 and $H_{\alpha|\boldsymbol{\gamma}}$ with their estimators. The following corollary shows that under H_0, type I error of the test $\psi_{\mathrm{DLRT}}(\xi)$ converges to the desired significance level ξ and the p-value is asymptotically uniform.

COROLLARY 4.2. *Suppose the conditions in Corollary 4.1 hold. Then*

$$\lim_{n\to\infty} \mathbb{P}(\psi_{\mathrm{DLRT}}(\xi) = 1 \mid H_0) = \xi \quad \textit{and} \quad P_{\mathrm{DLRT}} \rightsquigarrow \mathrm{Uniform}[0, 1] \qquad \textit{under } H_0,$$

where $\psi_{\mathrm{DLRT}}(\omega)$ is defined in (3.12) and $P_{\mathrm{DLRT}} = 1 - \chi_1^2((4 \cdot \widehat{\sigma}^2)^{-1} \cdot \widehat{H}_{\alpha|\boldsymbol{\gamma}} \cdot \Lambda_n)$ is the associated p-value.

PROOF. A detailed proof is shown in the Supplementary Material [29]. □

Finally, we conclude this section with the following remarks on the extensions to missing data and multiple datasets inference. Due to the space constraint, we defer the detailed results to the Supplementary Material [29].

REMARK 4 (Missing data and multiple datasets inference). In the missing data setup, as shown in equation (2.3), Y given $\boldsymbol{X}$ and $\delta = 1$ satisfies the semiparametric GLM with the same finite dimensional parameter $\boldsymbol{\beta}$ and unknown function $f^m(\cdot)$. The inferential results in this section can be easily extended to the missing data setup; see the Supplementary Material [29] for details. In the multiple datasets inference setup, the sparsity patterns of the d-dimensional parameter $\boldsymbol{\beta}_t^*$ in (1.2) are usually identical across $t = 1, \ldots, T$. To encourage the common sparsity of $\boldsymbol{\beta}_t^*$ and meanwhile account for the heterogeneity of different datasets, we can use similar estimation procedures to (4.2) with the group Lasso penalty. In the Supplementary Material [29], we obtain the finite sample error bounds for parameter estimation and the corresponding inferential results. In particular, by establishing a new maximal inequality for U-statistic with unbounded kernels (i.e., Lemma F.2 of the Supplementary Material [29]), we prove that the group Lasso estimator attains faster rates of convergence than the Lasso estimator. This extends the results in linear models [22] to the more challenging semiparametric setting.

5. Numerical results.

In this section, we provide synthetic and real data examples to back up the theoretical results.

5.1. *Simulation studies.* We conduct simulation studies to assess the finite sample performance of the proposed methods. We generate the outcomes from (1) the linear regression with the standard Gaussian noise or (2) the logistic regression, and the covariates from $N(0, \boldsymbol{\Sigma})$, where $\Sigma_{ij} = 0.6^{|i-j|}$. The true values of $\boldsymbol{\beta}$

are $\boldsymbol{\beta}^*_j = \mu$ for $j = 1, 2, 3$ and $\boldsymbol{\beta}^*_j = 0$ for $j = 4, \dots, d$. Thus, the cardinality of the support set of $\boldsymbol{\beta}^*$ is $s = 3$. The sample size is $n = 100$, the number of covariates is $d = 200$ and the number of simulation replications is 500.

We calculate the ℓ_1-regularized estimator $\widehat{\boldsymbol{\beta}}$ by using the `glmnet` package in `R`. In particular, we determine the regularization parameter λ by minimizing the K-fold cross validated loss function,

$$\mathrm{CV}(\lambda) = \sum_{k=1}^{K} \{\ell(\widehat{\boldsymbol{\beta}}_\lambda^{(-k)}) - \ell^{(-k)}(\widehat{\boldsymbol{\beta}}_\lambda^{(-k)})\},$$

where $\ell^{(-k)}$ stands for the loss function evaluated without the kth subset and similarly $\widehat{\boldsymbol{\beta}}_\lambda^{(-k)}$ stands for the regularized estimator derived without using the kth subset. In the simulation studies, we use 5-fold cross validation. The tuning parameter for the Dantzig selector λ_1 in (3.9) is chosen by $4\sqrt{\log(nd)/n}$. We find that the simulation results are not sensitive to the choice of λ_1. We only present the results with the Lasso penalty. Similar results are observed by using the folded concave penalty based on the LLA algorithm [12].

For the linear regression, we consider the directional likelihood ratio test (DLRT) and the Wald test based on the asymptotic normality of $\widehat{\alpha}^P$, as well as the desparsifying method in [37, 41] and debias method in [15]. Both of these two methods are tailored for the linear regression with the L_2 loss and are optimal for confidence intervals and hypothesis testing. To examine the validity of our tests, we report their type I errors for the null hypothesis $H_0 : \beta_1 = \mu$ with various choices of $\mu \in [0, 1]$ at the 0.05 significance level. The results are summarized in Table 1. We find that, our Wald test and DLRT yield accurate type I errors, which are comparable to the desparsifying and debias methods. In addition, we also compare the powers of these tests. In particular, we test the null hypothesis $H_0 : \beta_1 = 0$, but increase μ from 0 to 1 in the data generating procedure. As shown in the left panel

TABLE 1

Type I errors of the Wald test and directional likelihood ratio test (DLRT), the desparsifying and debias methods for linear and logistic regressions for $H_0 : \alpha = \mu$, at the 0.05 significance level, where $\mu = 0.00, \dots, 1.00$

Model	**Method**	**0.00**	**0.10**	**0.20**	**0.40**	**0.60**	**0.80**	**1.00**
Linear	Wald	0.048	0.066	0.060	0.052	0.054	0.046	0.054
	DLRT	0.040	0.052	0.064	0.042	0.032	0.034	0.040
	Desparsity	0.044	0.054	0.058	0.044	0.058	0.058	0.056
	Debias	0.034	0.030	0.036	0.024	0.028	0.028	0.028
Logistic	Wald	0.054	0.060	0.054	0.054	0.066	0.068	0.038
	DLRT	0.052	0.048	0.058	0.056	0.054	0.050	0.038
	Desparsity	0.052	0.044	0.058	0.046	0.050	0.058	0.058

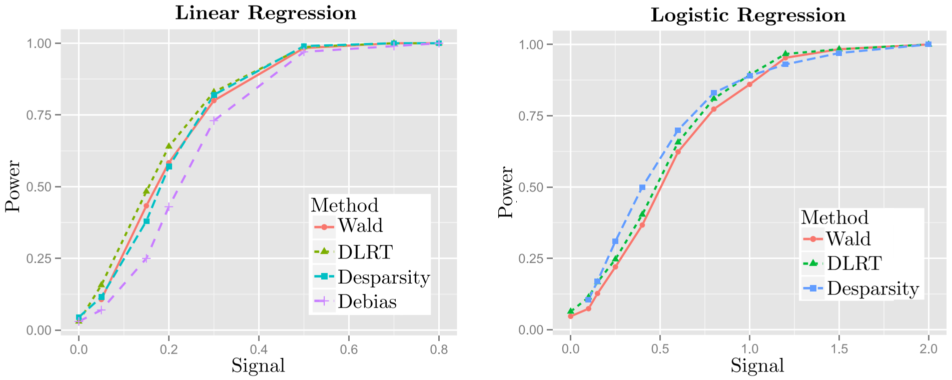

FIG. 1. *Power curves for testing* $H_0 : \beta_1 = 0$ *for the linear* (*left panel*) *and logistic* (*right panel*) *regressions at the* 0.05 *significance level.*

of Figure 1, our Wald test and DLRT based on the semiparametric GLM are nearly as efficient as the desparsifying and debias methods. Such results show that the semiparametric GLM gains model flexibility by losing little inferential efficiency.

For the logistic model, we only consider the desparsifying method, because the debias method is not defined. As shown in Table 1, our proposed tests yield well controlled type I errors. Similarly, the power comparison for testing $H_0 : \beta_1 = 0$ in Figure 1 reveals that our tests under the more flexible semiparametric model are comparable to the desparsifying method. Moreover, the DLRT is more powerful than the remaining two tests, which demonstrates the numerical advantages of the likelihood ratio inference over the Wald-type tests. This observation is also consistent with the literature for low dimensional inference.

To further demonstrate the advantage of the proposed methods, we consider the data with missing values. Similar to the previous data generating procedures, we first simulate the original data Y_i and $\boldsymbol{X}_i$. Then, for the linear regression, we consider the following two scenarios to create missing values: (1) the response Y_i is observed (i.e., $\delta_i = 1$) if and only if $Y_i \leq 0$; and (2) Y_i is always observed if $Y_i \leq 0$ and observed with probability 0.2 if $Y_i > 0$, that is, $\mathbb{P}(\delta_i = 1 \mid Y_i, \boldsymbol{X}_i) = 1 - 0.8I(Y_i > 0)$. For the logistic regression, we also consider two scenarios to create missing values: (1) $\mathbb{P}(\delta_i = 1 \mid Y_i, \boldsymbol{X}_i) = 0.2 + 0.6Y_i$; and (2) $\mathbb{P}(\delta_i = 1 \mid Y_i, \boldsymbol{X}_i) = 0.2 + 0.8Y_i$. Since the desparsifying and debias methods are developed based on the assumption that no missing values exist, we consider the following two practical procedures for handling missing data on Y. The first approach is that we apply the desparsifying and debias methods directly to samples with Y observed, which is known as the complete-case analysis. The second approach is that we apply these two methods to an imputed dataset. More specifically, for those samples with missing values on Y, we impute Y by using the k-nearest neighbors method, implemented by the R function `impute.knn`. The type I errors are shown in Table 2. As expected, for the desparsifying and debias methods,

TABLE 2

Type I errors of the Wald test and directional likelihood ratio test (*DLRT*), *the desparsifying method and debias method based on complete-case analysis* (*CC-*) *and imputation* (*Imp-*) *for linear and logistic regressions with missing data* (*selection bias*) *for* $H_0 : \alpha = \mu$, *at the* 0.05 *significance level, where* $\mu = 0.10, \dots, 0.25$

Scenario	Model	Method	0.10	0.15	0.20	0.25	0.30	0.35
1	Linear	Wald	0.062	0.048	0.064	0.046	0.064	0.050
		DLRT	0.056	0.042	0.060	0.036	0.056	0.048
		CC-Desparsity	0.076	0.156	0.214	0.278	0.334	0.580
		Imp-Desparsity	0.068	0.128	0.176	0.198	0.270	0.448
		CC-Debias	0.126	0.322	0.488	0.662	0.820	0.900
		Imp-Debias	0.108	0.260	0.306	0.438	0.470	0.624
1	Logistic	Wald	0.058	0.064	0.060	0.070	0.078	0.054
		DLRT	0.044	0.052	0.044	0.054	0.052	0.042
		CC-Desparsity	0.296	0.698	0.956	0.988	1.000	1.000
		Imp-Desparsity	0.214	0.582	0.902	0.980	1.000	1.000
2	Linear	Wald	0.060	0.068	0.048	0.060	0.072	0.052
		DLRT	0.060	0.062	0.040	0.048	0.052	0.046
		CC-Desparsity	0.086	0.098	0.164	0.370	0.524	0.660
		Imp-Desparsity	0.080	0.088	0.146	0.236	0.268	0.362
		CC-Debias	0.072	0.152	0.334	0.530	0.728	0.804
		Imp-Debias	0.070	0.096	0.148	0.308	0.376	0.442
2	Logistic	Wald	0.078	0.032	0.050	0.052	0.052	0.060
		DLRT	0.074	0.022	0.040	0.044	0.042	0.046
		CC-Desparsity	0.156	0.422	0.546	0.656	0.768	0.846
		Imp-Desparsity	0.124	0.234	0.340	0.338	0.466	0.514

the type I errors of the complete-case analysis are far from the 0.05 significance level. Although the imputation method shows some advantages over the complete-case analysis, similar patterns are observed. Therefore, in the presence of missing data, the existing methods cannot produce any result that is statistically reliable. In contrast, the type I errors based on the proposed tests are well controlled, and they are robust to the missing data and selection bias. The same conclusion holds under all simulation scenarios.

In summary, our proposed testing procedures under the semiparametric GLM are as competitive as the existing methods even if the assumed model is correct. More importantly, in the presence of missing data or selection bias, the proposed methods significantly outperform the existing ones.

5.2. *Analysis of gene expression data.* In this section, we apply the proposed tests to analyze the AGEMAP (Atlas of Gene Expression in Mouse Aging Project) gene expression data [40]. The dataset contains the expression values for 296 genes

TABLE 3

Significant genes selected by the Wald and directional likelihood ratio tests under the semiparametric GLM, the desparsifying method and debias method based on complete-case analysis (CC-) and imputation (Imp-) for the gene expression data. Here, $M\%$ samples are missing

M	**Wald**	**DLRT**	**CC-Desparsity**	**CC-Debias**	**Imp-Desparsity**	**Imp-Debias**
0	Cdc42	Cdc42	Cdc42	Cdc42	Cdc42	Cdc42
15	Cdc42	Cdc42	–	–	Mapk13	–
25	Cdc42	Cdc42	–	–	Ppp3cb	–
35	Cdc42	Cdc42	–	–	Nfatc3,Ppp3cb	–

belonging to the mouse vascular endothelial growth factor (VEGF) signaling pathway. The sample size is $n = 40$. Among these 296 genes, we are interested in identifying genes that are significantly associated with the target gene Casp9. Thus, we treat the gene Casp9 as the response and the remaining 295 genes as covariates.

Since no missing value presents, we directly apply the desparsifying and debias methods to test $H_0 : \beta_j = 0$ for each $1 \leq j \leq 295$, under the linear model assumption. Similarly, we can assume that the gene Casp9 given the remaining variables follows the semiparametric GLM and the proposed Wald and likelihood ratio tests can be applied. To take into account of the multiplicity of tests, we use the step-down method in the R function `p.adjust` to adjust the p-values. At the 0.05 significance level, all these four methods claim that gene Cdc42 is significant; see the first row of Table 3. This suggests that our tests are as effective as those existing procedures when there are no missing values.

To further illustrate the advantage of our methods in the presence of missing data, we create missing values for the outcome variable Y_i. More specifically, if Y_i is among the top $M\%$ samples, where $M = 0, 15, 25$ and 35, we remove the values of Y_i. Here, $M = 0$ means no missing data is created. This corresponds to the analysis of the original complete data. Similar to that in the simulation studies, the considered missing data mechanism depends on the unobserved values, which makes the analysis challenging.

The results are shown in Table 3, where the results based on the original complete data ($M = 0$) can be used as a benchmark. Based on the incomplete dataset, after the same adjustment for p-values, our Wald and likelihood ratio tests still select gene Cdc42, which are consistent with the results based on the original data. This pattern is preserved, even after 35% data are removed. For the desparsifying and debias methods, similar to the simulation studies, we can either apply them to those samples with only complete data (complete-case analysis) or the full data created by the imputation method. In particular, the CC-Desparsity and the CC-Debias methods consistently select no genes, when there exist missing data. This seems to suggest a lack of power for the existing methods based on the complete-case analysis. In addition, Imp-Desparsity tends to select very different genes at different levels of missing data percentage. They are all different from the benchmark gene Cdc42. Our analysis suggests that the presence of missing values can dramatically change the results of Imp-Desparsity. Finally, Imp-Debias performs similar to CC-Debias and tends to have low powers.

In conclusion, the existing methods based on the imputation methods or complete cases are either very sensitive to the missing data or have low powers. On the other hand, the proposed tests are quite robust and potentially more reliable in the presence of missing data.

6. Discussion. In this paper, we propose a new likelihood ratio inference framework for high-dimensional semiparametric generalized linear models. The proposed model is semiparametric in that the base measure function $f(\cdot)$ is unspecified. This offers extra flexibility to handle the problems with missing data, selection bias and heterogeneity. We note that the proposed model is different from many standard semiparametric models such as the partially linear model. Although in this paper we only consider the likelihood ratio inference for the semiparametric GLM, similar inferential methods can be applied to more general high-dimensional semiparametric models. This is an interesting direction to explore in the future.

Another future direction is to develop the joint confidence intervals for the entire d-dimensional parameter $\boldsymbol{\beta}^*$. Under the GLM, [4] constructed the joint confidence intervals based on a multiplier bootstrap method for approximating maximum of sums of independent high-dimensional random vectors [9]. Under the proposed semiparametric GLM, the gradient of the composite log-likelihood has a U-statistic structure. To construct joint confidence intervals, it may require to extend the high-dimensional multiplier bootstrap method based on sums of independent random vectors to U-statistics. Such extensions are worthy of further investigation.

APPENDIX A: PROOF OF MAIN RESULTS

In this Appendix, we give the proof of Theorem 4.1. The proofs of the remaining results, including Corollary 4.1, and Theorem 4.2 are deferred to the Supplementary Material [29].

We define an unbiased score function as $S(\boldsymbol{\beta}^*) := \nabla_\alpha \ell(\boldsymbol{\beta}^*) - \mathbf{w}^{*T} \nabla_{\boldsymbol{\gamma}} \ell(\boldsymbol{\beta}^*)$, which plays an important role in the proof. The proof of Theorem 4.1 has three steps. First, we show that the first derivative of $\widehat{\ell}(\alpha)$ approximates $S(\boldsymbol{\beta}^*)$. Second, we apply the central limit theorem for a linear combination of high- dimensional U-statistics to conclude the asymptotic normality of $S(\boldsymbol{\beta}^*)$. Finally, we show that the negative Hessian of $\widehat{\ell}(\alpha)$ approximates $H_{\alpha|\boldsymbol{\gamma}}$. For notational simplicity, denote $M := \max_{1\le i<j\le n} \|(y_i - y_j) \cdot (\boldsymbol{x}_i - \boldsymbol{x}_j)\|_\infty$. By Assumption 4.1, we have $M = \mathcal{O}_{\mathbb{P}}(\log n)$.

Step 1: Show the convergence of $\widehat{\ell}'(\alpha^*)$. Define $\widehat{\boldsymbol{\gamma}}(\alpha) := \widehat{\boldsymbol{\gamma}} + (\widehat{\alpha} - \alpha)\widehat{\mathbf{w}}$ and $\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}}(\alpha^*) - \boldsymbol{\gamma}^*$. Moreover, recall that $S(\boldsymbol{\beta}^*) := \nabla_\alpha \ell(\boldsymbol{\beta}^*) - \mathbf{w}^{*T} \nabla_{\boldsymbol{\gamma}} \ell(\boldsymbol{\beta}^*)$. By the chain rule and mean value theorem, we have

$$\widehat{\ell}'(\alpha^*) = \nabla_\alpha \ell(\alpha^*, \widehat{\boldsymbol{\gamma}}(\alpha^*)) - \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\gamma}} \ell(\alpha^*, \widehat{\boldsymbol{\gamma}}(\alpha^*)) = S(\boldsymbol{\beta}^*) + I_1 + I_2, \tag{A.1}$$

where $I_1 := (\mathbf{w}^* - \widehat{\mathbf{w}})^T \nabla_{\boldsymbol{\gamma}} \ell(\boldsymbol{\beta}^*)$ and $I_2 := \{\nabla^2_{\alpha\boldsymbol{\gamma}} \ell(\alpha^*, \bar{\boldsymbol{\gamma}}) - \widehat{\mathbf{w}}^T \nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \ell(\alpha^*, \widetilde{\boldsymbol{\gamma}})\}\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}$. Here, $\bar{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$ are intermediate values between $\boldsymbol{\gamma}^*$ and $\widehat{\boldsymbol{\gamma}}(\alpha^*)$. Thus, the first step of the proof reduces to controlling the two terms I_1 and I_2 in (A.1). In particular, to bound I_1, we need the following Lemma A.1 to bound $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1$ and Lemma A.2 to bound $\|\nabla \ell(\boldsymbol{\beta}^*)\|_\infty$, respectively.

LEMMA A.1. *Under the conditions in Theorem 4.1,*

$$\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}\left(M(s + s_1) \cdot \sqrt{\frac{\log d}{n}}\right).$$

LEMMA A.2. *Assume that Assumption 4.1 holds. Then, for any $C'' > 0$, we have $\|\nabla \ell(\boldsymbol{\beta}^*)\|_\infty \le C'' \cdot \sqrt{\log d/n}$, with probability at least*

$$1 - 2 \cdot d \cdot \exp\left[-\min\left\{\frac{C^2 \cdot C''^2}{2^9 \cdot C'^2 \cdot m^2} \cdot \frac{\log d}{n}, \frac{C \cdot C''}{2^5 \cdot C' \cdot m} \cdot \sqrt{\frac{\log d}{n}}\right\} \cdot k\right], \tag{A.2}$$

where $k = \lfloor n/2 \rfloor$, and C, C' are defined in Definition 1.1.

PROOF. To prove Lemma A.2, the key is to prove a new concentration inequality for U-statistics with subexponential kernel functions. In particular, the following lemma allows the kernel function to be unbounded, which is more general than most of existing concentration results for U-statistics with bounded kernels, such as Theorem 4.1.13 in [10]. The following result can be of independent interest, whose proof is shown in the Supplementary Material [29].

LEMMA A.3. *Let $X_1, \dots, X_n$ be independent random variables. Consider the following U-statistics of order m,*

$$U_n = \binom{n}{m}^{-1} \sum_{i_1 < \cdots < i_m} u(X_{i_1}, \dots, X_{i_m}),$$

where the summation is over all $i_1 < \cdots < i_m$ selected from $\{1, \dots, n\}$ and $\mathbb{E}[u(X_{i_1}, \dots, X_{i_m})] = 0$ for all $i_1 < \cdots < i_m$. Assume that the kernel function $u(X_{i_1}, \dots, X_{i_m})$ is symmetric in the sense that $u(X_{i_1}, \dots, X_{i_m})$ is independent of the order of $X_{i_1}, \dots, X_{i_m}$. If there exist constants L_1 and L_2, such that

$$\mathbb{P}(|u(X_{i_1}, \dots, X_{i_m})| \ge x) \le L_1 \cdot \exp(-L_2 \cdot x), \tag{A.3}$$

for all $i_1 < \cdots < i_m$ and all $x \ge 0$, then

$$\mathbb{P}(|U_n| \ge x) \le 2 \cdot \exp\left[-\min\left\{\frac{L_2^2 \cdot x^2}{8 \cdot L_1^2}, \frac{L_2 \cdot x}{4 \cdot L_1}\right\} \cdot k\right],$$

where $k = \lfloor n/m \rfloor$ is the largest integer less than n/m.

Given the above lemma, we need to verify that the kernel function $\mathbf{h}_{ij}(\boldsymbol{\beta}^*)$ has mean 0, where

$$\mathbf{h}_{ij}(\boldsymbol{\beta}) = \frac{R_{ij}(\boldsymbol{\beta}) \cdot (y_i - y_j) \cdot (\boldsymbol{x}_i - \boldsymbol{x}_j)}{1 + R_{ij}(\boldsymbol{\beta})}, \tag{A.4}$$

and it satisfies (A.3). To show $\mathbb{E}\{\mathbf{h}_{ij}(\boldsymbol{\beta}^*)\} = 0$, let Ξ_{ij} denote the event $\{(Y_{(i)}^L, Y_{(j)}^L) = (y_i, y_j), \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{X}_j = \boldsymbol{x}_j\}$. By (3.3), the conditional distribution of Y_i and Y_j given Ξ_{ij} follows a binomial distribution,

$$\mathbb{P}(Y_i = y_i, Y_j = y_j \mid \Xi_{ij}; \boldsymbol{\beta}) = \bigl[1 + R_{ij}(\boldsymbol{\beta})\bigr]^{-1}, \tag{A.5}$$

and $\mathbb{P}(Y_i = y_j, Y_j = y_i \mid \Xi_{ij}; \boldsymbol{\beta}) = R_{ij}(\boldsymbol{\beta})/[1 + R_{ij}(\boldsymbol{\beta})]$. According to this binomial distribution, the conditional expectation of $\mathbf{h}_{ij}(\boldsymbol{\beta}^*)$ given Ξ_{ij} is

$$\begin{aligned}
&\mathbb{E}\{\mathbf{h}_{ij}(\boldsymbol{\beta}^*) \mid \Xi_{ij}; \boldsymbol{\beta}^*\} \\
&\quad = \frac{R_{ij}(\boldsymbol{\beta}^*)(y_i - y_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)}{1 + R_{ij}(\boldsymbol{\beta}^*)} \mathbb{P}(Y_i = y_i, Y_j = y_j \mid \Xi_{ij}; \boldsymbol{\beta}^*) \\
&\qquad + \frac{R_{ij}^{-1}(\boldsymbol{\beta}^*)(y_j - y_i)(\boldsymbol{x}_i - \boldsymbol{x}_j)}{1 + R_{ij}^{-1}(\boldsymbol{\beta})} \mathbb{P}(Y_i = y_j, Y_j = y_i \mid \Xi_{ij}; \boldsymbol{\beta}^*).
\end{aligned}$$

By plugging (A.5) into above equation, it is easy to verify that $\mathbb{E}\{\mathbf{h}_{ij}(\boldsymbol{\beta}^*) \mid \Xi_{ij}\} = 0$. Finally, $\mathbb{E}\{\mathbf{h}_{ij}(\boldsymbol{\beta}^*)\} = \mathbb{E}[\mathbb{E}\{\mathbf{h}_{ij}(\boldsymbol{\beta}^*) | \Xi_{ij}\}] = 0$. Next, we verify the kernel function satisfies (A.3). Since $R_{ij}(\boldsymbol{\beta}) > 0$ and $\max_{ij} |x_{ij}| \le m$, we have

$$\bigl\|\mathbf{h}_{ij}(\boldsymbol{\beta}^*)\bigr\|_\infty \le \bigl\|(y_i - y_j) \cdot (\boldsymbol{x}_i - \boldsymbol{x}_j)\bigr\|_\infty \le 2 \cdot m \cdot |y_i - y_j|.$$

By the subexponential tail condition on y_i, for any $x > 0$ and $k = 1, \dots, d$,

$$\mathbb{P}\bigl(\bigl|[\mathbf{h}_{ij}(\boldsymbol{\beta}^*)]_k\bigr| > x\bigr) \le \mathbb{P}\bigl(|y_i - y_j| > (2m)^{-1}x\bigr) \le 2C' \exp\{-C(4m)^{-1}x\}.$$

Then we apply Lemma A.3 with $k = \lfloor n/2 \rfloor$ to complete the proof. □

Hence, by Lemma A.1 and Lemma A.2, we can show that

$$|I_1| \le \|\mathbf{w}^* - \widehat{\mathbf{w}}\|_1 \|\nabla_{\boldsymbol{\gamma}} \ell(\boldsymbol{\beta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}\biggl(M(s + s_1) \cdot \sqrt{\frac{\log d}{n}} \cdot \sqrt{\frac{\log d}{n}}\biggr) = o_{\mathbb{P}}\biggl(\frac{1}{\sqrt{n}}\biggr),$$

where the last step follows by the conditions in Theorem 4.1. We further separate I_2 into the following terms: $|I_2| \le I_{21} + I_{22} + I_{23}$, where $I_{21} = |\{\nabla_{\alpha\boldsymbol{\gamma}}^2 \ell(\boldsymbol{\beta}^*) - \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^2 \ell(\boldsymbol{\beta}^*)\}\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}|$, $I_{22} = |\{\nabla_{\alpha\boldsymbol{\gamma}}^2 \ell(\boldsymbol{\beta}^*) - \nabla_{\alpha\boldsymbol{\gamma}}^2 \ell(\alpha^*, \bar{\boldsymbol{\gamma}})\}\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}|$ and $I_{23} = |\widehat{\mathbf{w}}^T \{\nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^2 \ell(\boldsymbol{\beta}^*) - \nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^2 \ell(\alpha^*, \widetilde{\boldsymbol{\gamma}})\}\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}|$. To control the three terms, we first need to bound $\|\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}\|_1$. By the conditions in Theorem 4.1, we have $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s\sqrt{\log d/n})$ and $|\widehat{\alpha} - \alpha^*| = \mathcal{O}_{\mathbb{P}}(s^{1/2}\sqrt{\log d/n})$. Moreover, by the Cauchy–Schwarz inequality, it holds that $\|\mathbf{w}^*\|_1 \le \sqrt{s_1}\|\mathbf{w}^*\|_2 \le \sqrt{s_1}\|\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\mathbf{H}_{\alpha\boldsymbol{\gamma}}^T\|_2 \le$

$\sqrt{s_1}\lambda_{\min}(\mathbf{H})^{-1}\lambda_{\max}(\mathbf{H}) \le \sqrt{s_1}c^{-1}c'$, where the last inequality is by Assumption 4.1. Therefore,

$$\|\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}\|_1 \le \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\| + |\widehat{\alpha} - \alpha^*|\|\widehat{\mathbf{w}}\|_1 = \mathcal{O}_{\mathbb{P}}\biggl(\max\{s, s_1\}\sqrt{\frac{\log d}{n}}\biggr),$$

where we used the fact that $\|\widehat{\mathbf{w}}\|_1 = \|\mathbf{w}^*\|_1 + o_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(s_1^{1/2})$. To control the three terms in I_2, the key step is to quantify the smoothness of the Hessian matrix $\nabla^2\ell(\alpha^*, \boldsymbol{\gamma})$ in a small neighborhood of $\boldsymbol{\gamma}^*$.

LEMMA A.4. *Under the conditions in Theorem 4.1, for any deterministic sequence δ_n such that $M \cdot \delta_n = o(1)$, we have*

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1\le\delta_n} \|\nabla^2\ell(\boldsymbol{\beta}) - \nabla^2\ell(\boldsymbol{\beta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}(M \cdot \delta_n),$$

where $M := \max_{1\le i<j\le n} \|(y_i - y_j)\cdot(\boldsymbol{x}_i - \boldsymbol{x}_j)\|_\infty$.

PROOF. Let $w_{ij} = \exp\{-(y_i - y_j)\cdot \boldsymbol{\Delta}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\}$, where $\boldsymbol{\Delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$. By definition, $R_{ij}(\boldsymbol{\beta}) = R_{ij}(\boldsymbol{\beta}^*)\cdot w_{ij}$. Thus,

$$\nabla^2\ell(\boldsymbol{\beta}) = -\binom{n}{2}^{-1} \sum_{1\le i<j\le n} \frac{u_{ij}\cdot R_{ij}(\boldsymbol{\beta}^*)\cdot(y_i - y_j)^2\cdot(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\otimes 2}}{(1 + R_{ij}(\boldsymbol{\beta}^*))^2},$$

where $u_{ij} = w_{ij}\cdot(1 + R_{ij}(\boldsymbol{\beta}^*))^2(1 + w_{ij}\cdot R_{ij}(\boldsymbol{\beta}^*))^{-2}$. Note that if $w_{ij} \ge 1$, then $(1 + R_{ij}(\boldsymbol{\beta}^*))^2/(1 + w_{ij}\cdot R_{ij}(\boldsymbol{\beta}^*))^2 \le 1$. On the other hand, if $w_{ij} \le 1$,

$$\frac{(1 + R_{ij}(\boldsymbol{\beta}^*))^2}{(1 + w_{ij}\cdot R_{ij}(\boldsymbol{\beta}^*))^2} \le \frac{(1 + R_{ij}(\boldsymbol{\beta}^*))^2}{w_{ij}^2\cdot(1 + R_{ij}(\boldsymbol{\beta}^*))^2} = \frac{1}{w_{ij}^2}.$$

Thus, $u_{ij} \le \max\{w_{ij}, w_{ij}^{-1}\}$. Therefore, for any $1 \le s, t \le d$,

$$\begin{aligned} &|\nabla^2_{st}\ell(\boldsymbol{\beta}) - \nabla^2_{st}\ell(\boldsymbol{\beta}^*)| \\ &\quad = \binom{n}{2}^{-1}\sum_{i<j} \frac{R_{ij}(\boldsymbol{\beta})(y_i - y_j)^2(x_{is} - x_{js})(x_{it} - x_{jt})(u_{ij} - 1)}{(1 + R_{ij}(\boldsymbol{\beta}))^2} \qquad \text{(A.6)} \\ &\quad \le 2^{-1}|\nabla^2_{ss}\ell(\boldsymbol{\beta}^*) + \nabla^2_{tt}\ell(\boldsymbol{\beta}^*)| \max_{i<j}|\max\{w_{ij}, w_{ij}^{-1}\} - 1|. \end{aligned}$$

By Hölder's inequality, we have

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1\le\delta_n} \max_{i<j}|(y_i - y_j)\cdot\boldsymbol{\Delta}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)| \le M\cdot\|\boldsymbol{\Delta}\|_1 = \mathcal{O}_{\mathbb{P}}(M\cdot\delta_n) = o_{\mathbb{P}}(1),$$

and $\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1\le\delta_n}\max_{i<j}|\max\{w_{ij}, w_{ij}^{-1}\} - 1| = \mathcal{O}_{\mathbb{P}}(M\cdot\delta_n)$. Thus, by (A.6),

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1\le\delta_n} \|\nabla^2\ell(\boldsymbol{\beta}) - \nabla^2\ell(\boldsymbol{\beta}^*)\|_\infty \lesssim \{\|\nabla^2\ell(\boldsymbol{\beta}^*) + \mathbf{H}\|_\infty + \|\mathbf{H}\|_\infty\}M\delta_n.$$

By Assumption 4.1, $\|\mathbf{H}\|_\infty$ is bounded. It remains to control $\|\nabla^2\ell(\boldsymbol{\beta}^*)+\mathbf{H}\|_\infty$. Let $\bar{\mathbf{r}}_{ij}=\mathbf{T}_{ij}-\mathbb{E}(\mathbf{T}_{ij})$, where

$$\mathbf{T}_{ij}=\frac{R_{ij}(\boldsymbol{\beta}^*)\cdot(y_i-y_j)^2\cdot(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\otimes 2}}{(1+R_{ij}(\boldsymbol{\beta}^*))^2}.$$

Then $\nabla^2\ell(\boldsymbol{\beta}^*)+\mathbf{H}=-\frac{2}{n(n-1)}\cdot\sum_{i<j}\bar{\mathbf{r}}_{ij}$ is a mean-zero second-order U-statistic with kernel function $\bar{\mathbf{r}}_{ij}$. For any $1\le a,b\le d$, $\bar{\mathbf{r}}_{ij}$ satisfies $[\bar{\mathbf{r}}_{ij}]_{(a,b)}\le 2\cdot M^2$. The Hoeffding inequality yields, for any $x>0$,

$$\mathbb{P}\big(\big|\nabla^2_{ab}\ell(\boldsymbol{\beta}^*)+\mathbf{H}_{a,b}\big|>x\big)\le 2\cdot\exp\biggl(-\frac{k\cdot x^2}{8\cdot M^4}\biggr),$$

where $k=\lfloor n/2\rfloor$. Taking $x=M^2\sqrt{\log d/n}$, by union bound, we get with high probability, $\|\nabla^2\ell(\boldsymbol{\beta}^*)+\mathbf{H}\|_\infty\le M^2\sqrt{\log d/n}$. □

Now we consider these three terms in I_2 one by one. For I_{21}, by Lemmas A.1 and C.2 in the Supplementary Material [29],

$$\begin{aligned}I_{21}&\le\big\|\nabla^2_{\alpha\boldsymbol{\gamma}}\ell(\boldsymbol{\beta}^*)-\mathbf{w}^{*T}\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\ell(\boldsymbol{\beta})\big\|_\infty\|\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}\|_1+\|\widehat{\mathbf{w}}-\mathbf{w}^*\|_1\big\|\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\ell(\boldsymbol{\beta}^*)\big\|_\infty\|\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}\|_1\\&=\mathcal{O}_{\mathbb{P}}\biggl(M\cdot\max\{s,s_1\}\cdot\frac{\log d}{n}+M\cdot\max\{s,s_1\}^2\cdot\frac{\log d}{n}\biggr)=o_{\mathbb{P}}\biggl(\frac{1}{\sqrt{n}}\biggr),\end{aligned}$$

where the last step follows from the scaling condition (4.1). Now, we consider I_{22}. By Lemma A.4 and the fact that $\|\bar{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1\le\|\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}\|_1=\mathcal{O}_{\mathbb{P}}(\max\{s,s_1\}\cdot\sqrt{\log d/n})$, we have

$$\begin{aligned}I_{22}&\le\big\|\nabla^2_{\alpha\boldsymbol{\gamma}}\ell(\boldsymbol{\beta}^*)-\nabla^2_{\alpha\boldsymbol{\gamma}}\ell(\alpha^*,\bar{\boldsymbol{\gamma}})\big\|_\infty\|\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}\|_1\\&=\mathcal{O}_{\mathbb{P}}\biggl(M\max\{s,s_1\}^2\frac{\log d}{n}\biggr)=o_{\mathbb{P}}\biggl(\frac{1}{\sqrt{n}}\biggr),\end{aligned}$$

where the last equality follows from the scaling condition (4.1). Following the similar arguments as in the proof of Lemma A.4, we can prove that

$$I_{23}\le C\biggl(M\cdot\max\{s,s_1\}\cdot\sqrt{\frac{\log d}{n}}\biggr)\cdot\big|\widehat{\mathbf{w}}^T\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\ell(\boldsymbol{\beta}^*)\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}\big|. \tag{A.7}$$

By Lemma A.1 and the similar argument to the proof of Lemma C.2,

$$\begin{aligned}\big|\widehat{\mathbf{w}}^T\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\ell(\boldsymbol{\beta}^*)\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}\big|&\le\big|\mathbf{w}^{*T}\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\ell(\boldsymbol{\beta}^*)\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}\big|+\big|(\widehat{\mathbf{w}}-\mathbf{w}^*)^T\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\ell(\boldsymbol{\beta}^*)\widehat{\boldsymbol{\Delta}}_{\boldsymbol{\gamma}}\big|\\&=\mathcal{O}_{\mathbb{P}}\biggl(\max\{s,s_1\}\cdot\sqrt{\frac{\log d}{n}}+M\cdot\max\{s,s_1\}^2\cdot\frac{\log d}{n}\biggr).\end{aligned}$$

Together with (A.7), we have

$$I_{23}=\mathcal{O}_{\mathbb{P}}\biggl(M\cdot\max\{s,s_1\}^2\cdot\frac{\log d}{n}\biggr)=o_{\mathbb{P}}\biggl(\frac{1}{\sqrt{n}}\biggr).$$

Thus, we have proved the rate of convergence of $n^{1/2}|\widehat{\ell}'(\alpha^*) - S(\boldsymbol{\beta}^*)|$, that is,

$$n^{1/2} \cdot |\widehat{\ell}'(\alpha^*) - S(\boldsymbol{\beta}^*)| = \mathcal{O}_{\mathbb{P}}\biggl(M \cdot \max\{s, s_1\}^2 \cdot \frac{\log d}{\sqrt{n}}\biggr) = o_{\mathbb{P}}(1). \tag{A.8}$$

Step 2: Characterize the limiting distribution of $S(\boldsymbol{\beta}^*)$. We provide the following lemma on the central limit theorem for U-statistics with increasing dimensions.

LEMMA A.5. *Under Assumption* 4.1, *for any* $\mathbf{b} \in \mathbb{R}^d$ *with* $\|\mathbf{b}\|_0 \le \widetilde{s}$ *and* $\|\mathbf{b}\|_2 = 1$, *if* $\widetilde{s}^{3/2} \cdot n^{-1/2} \cdot M^3 = o_{\mathbb{P}}(1)$, *then*

$$\frac{\sqrt{n}}{2} \cdot (\mathbf{b}^T \boldsymbol{\Sigma} \mathbf{b})^{-1/2} \cdot \mathbf{b}^T \nabla \ell(\boldsymbol{\beta}^*) \rightsquigarrow N(0, 1).$$

PROOF. The lemma is proved by using the Hoeffding's decomposition:

$$\begin{aligned}
&\frac{\sqrt{n}}{2} \cdot (\mathbf{b}^T \boldsymbol{\Sigma} \mathbf{b})^{-1/2} \cdot \mathbf{b}^T \nabla \ell(\boldsymbol{\beta}^*) \\
&\quad = (\mathbf{b}^T \boldsymbol{\Sigma} \mathbf{b})^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{b}^T \mathbf{g}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}^*) \\
&\qquad + \frac{\sqrt{n}}{2} (\mathbf{b}^T \boldsymbol{\Sigma} \mathbf{b})^{-1/2} \mathbf{b}^T \{\nabla \ell(\boldsymbol{\beta}^*) - \widehat{\mathbf{U}}_n\},
\end{aligned}$$

where $\mathbf{g}(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}^*)$ and $\widehat{\mathbf{U}}_n$ are defined in (3.10). We can verify that the Lyapunov central limit theorem for independent random variables can be applied for the first term under the assumption that $s_1 = o(n^{1/3-\delta})$. The remaining proof requires more careful calculation of the moment of approximation error $\mathbf{b}^T(\nabla \ell(\boldsymbol{\beta}^*) - \widehat{\mathbf{U}}_n)$ in the Hájek projection, because here we allow the intrinsic dimension $\widetilde{s}$ to scale with n. We defer the detailed proof to the Supplementary Material [29]. □

Since $S(\boldsymbol{\beta}^*)$ is a sparse linear combination of the U-statistic $\nabla \ell(0, \boldsymbol{\gamma}^*)$ and $\|\mathbf{w}^*\|_0 = s_1$, with $\mathbf{b} = (1, -\mathbf{w}^{*T})^T$, Lemma A.5 implies that

$$n^{1/2} S(\boldsymbol{\beta}^*)/(2\sigma) \rightsquigarrow N(0, 1). \tag{A.9}$$

Step3: Show the convergence of $\widehat{\ell}''(\bar{\alpha})$ *for any* $\bar{\alpha}$ *between* 0 *and* $\widehat{\alpha}^P$. We now show that $|\widehat{\ell}_n''(\bar{\alpha}) + H_{\alpha|\boldsymbol{\gamma}}| = o_{\mathbb{P}}(1)$. By chain rule, we have

$$\begin{aligned}
\widehat{\ell}_n''(\bar{\alpha}) &= \nabla_{\alpha\alpha}^2 \ell(\bar{\alpha}, \widehat{\boldsymbol{\gamma}}(\bar{\alpha})) - 2\nabla_{\alpha\boldsymbol{\gamma}}^2 \ell(\bar{\alpha}, \widehat{\boldsymbol{\gamma}}(\bar{\alpha}))^T \widehat{\mathbf{w}} + \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^2 \ell(\bar{\alpha}, \widehat{\boldsymbol{\gamma}}(\bar{\alpha}))^T \widehat{\mathbf{w}} \\
&= (1, -\widehat{\mathbf{w}}^T) \nabla^2 \ell(\bar{\alpha}, \widehat{\boldsymbol{\gamma}}(\bar{\alpha}))(1, -\widehat{\mathbf{w}}^T)^T.
\end{aligned} \tag{A.10}$$

We then decompose $\widehat{\ell}_n''(\bar{\alpha}) + H_{\alpha|\boldsymbol{\gamma}}$ into two terms, namely,

$$\begin{aligned}
\widehat{\ell}_n''(\bar{\alpha}) + H_{\alpha|\boldsymbol{\gamma}} &= [\widehat{\ell}_n''(\bar{\alpha}) - (1, -\widehat{\mathbf{w}}^T) \nabla^2 \ell(\boldsymbol{\beta}^*)(1, -\widehat{\mathbf{w}}^T)^T] \\
&\quad + [(1, -\widehat{\mathbf{w}}^T) \nabla^2 \ell(\boldsymbol{\beta}^*)(1, -\widehat{\mathbf{w}}^T)^T + H_{\alpha|\boldsymbol{\gamma}}] \\
&:= I_3 + I_4.
\end{aligned} \tag{A.11}$$

Let $\bar{\mathbf{\Delta}} = (\bar{\alpha}, \widehat{\boldsymbol{\gamma}}(\bar{\alpha})^T)^T - \boldsymbol{\beta}^*$. We have

$$\|\bar{\mathbf{\Delta}}\|_1 \le |\bar{\alpha} - \alpha^*| + \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 + |\widehat{\alpha} - \bar{\alpha}|\|\widehat{\mathbf{w}}\|_1. \tag{A.12}$$

To control $|\bar{\alpha} - \alpha^*|$, we need a bound on the rate of convergence of the post-regularization estimator $\widehat{\alpha}^P - \alpha^*$. The following lemma serves our purpose.

LEMMA A.6. *Under the conditions in Theorem 4.1, we have*

$$|\widehat{\alpha}^P - \alpha^*| = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{\log n}{n}}\Big).$$

By Lemma A.6, we have $|\bar{\alpha} - \alpha^*| \le |\widehat{\alpha}^P - \alpha^*| = \mathcal{O}_{\mathbb{P}}(\sqrt{\log n/n})$. Moreover, we have $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s \cdot \sqrt{\log d/n})$, $\|\widehat{\mathbf{w}}\|_1 = \|\mathbf{w}^*\|_1 + o_{\mathbb{P}}(1)$ and that

$$|\widehat{\alpha} - \bar{\alpha}| \le |\widehat{\alpha} - \alpha^*| + |\bar{\alpha} - \alpha^*| = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{s \log(d \vee n)}{n}}\Big).$$

Putting together the above results and by (A.12), we conclude that $\|\bar{\mathbf{\Delta}}\|_1 = \mathcal{O}_{\mathbb{P}}(\max\{s, s_1\} \cdot \sqrt{\log(d \vee n)/n})$.

For the first term in (A.11), similar to the proof of Lemma A.4, we get

$$|I_3| \le C\Big(M \cdot \max\{s, s_1\} \cdot \sqrt{\frac{\log(d \vee n)}{n}}\Big) \cdot |\widehat{\mathbf{v}}^T \nabla^2 \ell(\boldsymbol{\beta}^*)\widehat{\mathbf{v}}|, \tag{A.13}$$

where $\widehat{\mathbf{v}} = (1, \widehat{\mathbf{w}}^T)^T$. Let $\mathbf{v}^* = (1, \mathbf{w}^{*T})^T$. By Lemma A.1 and Lemma C.2,

$$\begin{aligned} |\widehat{\mathbf{v}}^T \nabla^2 \ell(\boldsymbol{\beta}^*)\widehat{\mathbf{v}}| &\le |\mathbf{v}^{*T} \nabla^2 \ell(\boldsymbol{\beta}^*)\mathbf{v}^*| + 2|(\widehat{\mathbf{v}} - \mathbf{v}^*)^T \nabla^2 \ell(\boldsymbol{\beta}^*)\mathbf{v}^*| \\ &\quad + |(\widehat{\mathbf{v}} - \mathbf{v}^*)^T \nabla^2 \ell(\boldsymbol{\beta}^*)(\widehat{\mathbf{v}} - \mathbf{v}^*)| \\ &\le |\mathbf{v}^{*T} \mathbf{H} \mathbf{v}^*| + o_{\mathbb{P}}(1). \end{aligned}$$

Therefore, we conclude that

$$|I_3| = \mathcal{O}_{\mathbb{P}}\big(M \cdot \max\{s, s_1\} \cdot \sqrt{\log(d \vee n)/n}\big) = o_{\mathbb{P}}(1). \tag{A.14}$$

We now focus on I_4, which can be decomposed into the following terms: $I_4 = I_{41} - 2I_{42} + I_{43}$, where $I_{41} = \nabla^2_{\alpha\alpha} \ell(\boldsymbol{\beta}^*) + H_{\alpha\alpha}$, $I_{42} = \widehat{\mathbf{w}}^T \nabla^2_{\alpha\boldsymbol{\gamma}} \ell(\boldsymbol{\beta}^*) + \mathbf{w}^{*T} \mathbf{H}_{\alpha\boldsymbol{\gamma}}$ and $I_{43} = \widehat{\mathbf{w}}^T \nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \ell(\boldsymbol{\beta}^*)\widehat{\mathbf{w}} + \mathbf{w}^{*T} \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \mathbf{w}^*$. By the proof of Lemma A.4, we have $\|\nabla^2 \ell(\boldsymbol{\beta}^*) + \mathbf{H}\|_\infty = \mathcal{O}_{\mathbb{P}}(M^2 \cdot \sqrt{\log d/n})$. Hence, $I_{41} = \mathcal{O}_{\mathbb{P}}(M^2 \cdot \sqrt{\log d/n}) = o_{\mathbb{P}}(1)$. For the second term, it holds that $I_{42} = \widehat{\mathbf{w}}^T (\nabla^2_{\alpha\boldsymbol{\gamma}} \ell(\boldsymbol{\beta}^*) + \mathbf{H}_{\alpha\boldsymbol{\gamma}}) - (\widehat{\mathbf{w}} - \mathbf{w}^*)^T \mathbf{H}_{\alpha\boldsymbol{\gamma}}$. We have $|\widehat{\mathbf{w}}^T (\nabla^2_{\alpha\boldsymbol{\gamma}} \ell(\boldsymbol{\beta}^*) + \mathbf{H}_{\alpha\boldsymbol{\gamma}})| \le \|\widehat{\mathbf{w}}\|_1 \|\nabla^2_{\alpha\boldsymbol{\gamma}} \ell(\boldsymbol{\beta}^*) + \mathbf{H}_{\alpha\boldsymbol{\gamma}}\|_\infty = \mathcal{O}_{\mathbb{P}}(M^2 \cdot \sqrt{s_1 \log d/n})$, and $|(\widehat{\mathbf{w}} - \mathbf{w}^*)^T \mathbf{H}_{\alpha\boldsymbol{\gamma}}| \le \|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 \|\mathbf{H}_{\alpha\boldsymbol{\gamma}}\|_\infty = o_{\mathbb{P}}(1)$.

Therefore, we conclude that $|I_{42}| = o_{\mathbb{P}}(1)$. For the term I_{43}, we apply similar arguments to get $I_{43} = \mathcal{O}_{\mathbb{P}}(M(s_1 + s)\sqrt{\log d/n}) = o_{\mathbb{P}}(1)$. Hence, we conclude that $I_4 = o_{\mathbb{P}}(1)$. Together with (A.14), this implies

$$|\widehat{\ell}_n''(\bar{\alpha}) + H_{\alpha|\boldsymbol{\gamma}}| = o_{\mathbb{P}}(1). \tag{A.15}$$

Given (A.8), (A.9), (A.15), we now wrap up the whole proof. By first-order optimality condition, we have $\widehat{\ell}'(\widehat{\alpha}^P) = 0$. Applying mean-value theorem, we get $\widehat{\ell}'(\widehat{\alpha}^P) = \widehat{\ell}'(\alpha^*) + \widehat{\ell}''(\bar{\alpha})(\widehat{\alpha}^P - \alpha^*)$, where $\bar{\alpha}$ is an intermediate value between $\widehat{\alpha}^P$ and α^*. This implies

$$\widehat{\alpha}^P - \alpha^* = \widehat{\ell}''(\bar{\alpha})^{-1}\widehat{\ell}'(\alpha^*). \tag{A.16}$$

Finally, combining (A.16), (A.8), (A.9), (A.15) and applying Slutsky's theorem, we have $n^{1/2}(\widehat{\alpha}^P - \alpha^*) = -H_{\alpha|\boldsymbol{\gamma}}^{-1} \cdot n^{1/2} S(\boldsymbol{\beta}^*) + o_{\mathbb{P}}(1)$. We complete the proof of Theorem 4.1.

Acknowledgments. We thank the Editor, Associate Editor and two referees for their helpful comments, which significantly improves the presentation of this paper. We are also very grateful to Professor Cun-Hui Zhang for sharing his 2011 Oberwolfach report on statistical inference for high-dimensional data. We also thank Nancy Reid and Jing Qin for insightful comments.

SUPPLEMENTARY MATERIAL

Supplement for "A likelihood ratio framework for high-dimensional semiparametric regression" (DOI: 10.1214/16-AOS1483SUPP; .pdf). The supplementary material contain additional technical details, simulation results and proofs.

REFERENCES

[1] ARGYRIOU, A., EVGENIOU, T. and PONTIL, M. (2008). Convex multi-task feature learning. *Mach. Learn.* **73** 243–272.

[2] BELLONI, A., CHEN, D., CHERNOZHUKOV, V. and HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80** 2369–2429. MR3001131

[3] BELLONI, A., CHERNOZHUKOV, V. and KATO, K. (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika* **102** 77–94. MR3335097

[4] BELLONI, A., CHERNOZHUKOV, V. and WEI, Y. (2016). Post-selection inference for generalized linear models with many controls. *J. Bus. Econom. Statist.* **34** 606–619. MR3547999

[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] BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

[7] CHAN, K. C. G. (2013). Nuisance parameter elimination for proportional likelihood ratio models with nonignorable missingness and random truncation. *Biometrika* **100** 269–276. MR3034342

[8] CHEN, Y., NING, Y., HONG, C. and WANG, S. (2014). Semiparametric tests for identifying differentially methylated loci with case-control designs using Illumina arrays. *Genet. Epidemiol.* **38** 42–50.

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

[10] DE LA PEÑA, V. H. and GINÉ, E. (1999). *Decoupling*: *From Dependence to Independence*. Springer, New York. MR1666908

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

[12] FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42** 819–849. MR3210988

[13] HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* **19** 293–325. MR0026294

[14] HOFF, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.* **1** 265–283. MR2393851

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

[16] KALBFLEISCH, J. D. (1978). Likelihood methods and nonparametric tests. *J. Amer. Statist. Assoc.* **73** 167–170. MR0518600

[17] LAWLESS, J. F., KALBFLEISCH, J. D. and WILD, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 413–438. MR1680310

[18] LIANG, K.-Y. and QIN, J. (2000). Regression analysis under non-standard situations: A pairwise pseudolikelihood approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 773–786. MR1796291

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

[20] LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664. MR3015038

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

[22] LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865

[23] LUO, X. and TSAI, W. Y. (2012). A proportional likelihood ratio model. *Biometrika* **99** 211–222. MR2899674

[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] MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

[28] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450

[29] NING, Y., ZHAO, T. and LIU, H. (2017). Supplement to "A likelihood ratio framework for high dimensional semiparametric regression." DOI:10.1214/16-AOS1483SUPP.

[30] PETTITT, A. N. (1982). Inference for the linear model using a likelihood based on ranks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **44** 234–243. MR0676214

[31] SEVERINI, T. A. and WONG, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20** 1768–1802. MR1193312

[32] SHAH, R. D. and SAMWORTH, R. J. (2013). Variable selection with error control: Another look at stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 55–80. MR3008271

[33] SRIVASTAVA, V. K. and GILES, D. E. A. (1987). *Seemingly Unrelated Regression Equations Models: Estimation and Inference. Statistics: Textbooks and Monographs* **80**. Dekker, New York. MR0930104

[34] STÄDLER, N. and BÜHLMANN, P. (2012). Missing values: Sparse inverse covariance estimation and an extension to sparse regression. *Stat. Comput.* **22** 219–235. MR2865066

[35] TAYLOR, J., LOCKHART, R., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). Post-selection adaptive inference for least angle regression and the lasso. arXiv preprint arXiv:1401.3889.

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

[37] 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

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

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

[40] ZAHN, J. M., POOSALA, S., OWEN, A. B., INGRAM, D. K., LUSTIG, A., CARTER, A., WEERARATNA, A. T., TAUB, D. D., GOROSPE, M., MAZAN-MAMCZARZ, K. et al. (2007). AGEMAP: A gene expression database for aging in mice. *PLoS Genet.* **3** 2326–2337.

[41] 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

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

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

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