Submitted to Biometrika (2022), *pp*. 1–15
Printed in Great Britain

Testing generalized linear models with high-dimensional nuisance parameter

BY JINSONG CHEN

College of Applied Health Sciences, University of Illinois at Chicago, 1919 W Taylor St, Chicago, Illinois 60612, U.S.A.

jinsongc@uic.edu

QUEFENG LI

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.

quefeng@email.unc.edu

HUA YUN CHEN

School of Public Health, University of Illinois at Chicago, 2121 W Taylor St, Chicago, Illinois 60612, U.S.A.

hychen@uic.edu

SUMMARY

Generalized linear models often have a high-dimensional nuisance parameters, as seen in applications such as testing gene-environment interactions or gene-gene interactions. In these scenarios, it is essential to test the significance of a high-dimensional sub-vector of the model's coefficients. Although some existing methods can tackle this problem, they often rely on the bootstrap to approximate the asymptotic distribution of the test statistic, and thus are computationally expensive. Here, we propose a computationally efficient test with a closed-form limiting distribution, which allows the parameter being tested to be either sparse or dense. We show that under certain regularity conditions, the type I error of the proposed method is asymptotically correct, and we establish its power under high-dimensional alternatives. Extensive simulations demonstrate the good performance of the proposed test and its robustness when certain sparsity assumptions are violated. We also apply the proposed method to Chinese famine sample data in order to show its performance when testing the significance of gene-environment interactions.

Some key words: Dense parameter; U-statistics; Model misspecification.

1. INTRODUCTION

Testing hypotheses for high-dimensional generalized linear models is a basic task of statistical inference. It is especially important to accurately and efficiently test the significance of a high-dimensional sub-vector of the model coefficients when the nuisance parameter is also high-dimensional. One application of this problem is testing the significance of interaction terms in a high-dimensional generalized linear model–for example, in studies of gene-environment or gene-gene interactions, which contribute to many complex diseases and traits in addition to their own primary effects.

In statistics, a strong heredity condition means that an interaction is significant only if both of the main effects are important (Hamada & Wu, 1992). A naive two-stage application of the strong heredity condition first selects covariates with significant main effects, and then examines the significance of their interactions (Wu et al., 2009). The heredity condition is important in some methods dealing with inter-

action screening in a high-dimensional setting (Bien et al., 2013; Hao & Zhang, 2014). However, the heredity condition might not be realistic in practice: there are many variables that are not important in main effects, but whose interactions affect outcomes and must be accounted for. In many genetic studies, the focus is on testing the interaction between a set of high-dimensional single-nucleotide polymorphisms (SNPs) and an environmental variable; in this scenario, we must include all SNPs in this set into the model as nuisance effects (Barnett et al., 2017). The model under the null hypothesis is thus high-dimensional when the interest is in testing the gene-environment or gene-gene interactions. The method in this paper is suited for this task.

There are a number of approaches in the recent literature dealing with testing parameters in the high-dimensional generalized linear model. This related scholarship can be classified into three categories depending on the dimensions of the testing and nuisance parameters.

Category 1: testing a low-dimensional parameter with a high-dimensional nuisance parameter. For this type of inference, Zhang & Zhang (2014) and van de Geer et al. (2014) proposed desparsified lasso approaches, and Ning & Liu (2017) developed a decorrelated score test. Shi et al. (2019) proposed partial penalized likelihood ratio, score, and Wald tests for testing some low-dimensional linear combinations of parameters for high-dimensional generalized linear models and showed that these tests are asymptotically equivalent. Sun & Zhang (2020) proposed a modified profile likelihood test, where the test statistic was constructed by penalizing only the high-dimensional nuisance parameter. Sur & Candes (2019) considered the maximum likelihood estimate of the logistic regression in which sample and variable sizes become increasingly large in a fixed ratio; they show that the classical maximum likelihood estimate theories no longer work in this regime and develop a theory for high-dimensional logistic regression models.

Category 2: testing a high-dimensional parameter with a low-dimensional nuisance parameter. For this type of inference, Geoman et al. (2011) proposed a scoring procedure that is applicable to testing generalized linear models with canonical link. This approach was modified by Guo & Chen (2016), who retained the power of the original test while obtaining a simpler asymptotic distribution for the test statistic and accommodating a wide range of link functions. Zhang & Cheng (2017) proposed simultaneous inference on a high-dimensional sparse linear model, which requires a desparsified lasso estimator for the full model to construct the test statistics. Barnett et al. (2017) proposed a generalized higher criticism method for testing associations between sets of SNPs and particular disease outcomes under generalized linear models. Ma et al. (2019) considered global and simultaneous hypothesis tests for high-dimensional logistic regression models. In general, the methods in this second category apply only to the case where the model under the null hypothesis is low-dimensional.

Category 3: testing a high-dimensional parameter with a high-dimensional nuisance parameter. This inference, which is the inference of interest here, can be addressed by extending some existing methods. For example, the test developed by Zhu & Bradic (2018) can be used for this type of inference by choosing appropriate loading vectors, although this method only applies to linear regression models. For generalized linear models, an extension of the decorrelated score test by Ning & Liu (2017) can be used, but there are two limitations for this approach. First, it is computationally expensive, as it needs to utilize the multiplier bootstrap procedure (Chernozhukov et al., 2013) to approximate the limiting distribution. Second, it requires the testing parameters to be sparse. Based on our simulation studies in Section 4, these limitations potentially lead to inflation of type I error and loss of power in high-dimensional generalized linear models. Wu et al. (2020) proposed an adaptive interaction sum of powered score test. To maintain high statistical power across a wide range of alternatives, this approach depends on a good choice of power index, but in practice, the optimal choice is unknown and requires computationally expensive ad-hoc methods to find it. In addition, their approach only allows the dimension of testing parameters to be in a polynomial order of the sample size.

For the generalized linear model, there are very few works on testing a high-dimensional parameter with a high-dimensional nuisance parameter. To efficiently fulfill this important inferential task, we propose a new approach that extends the methods proposed by Geoman et al. (2011) and Guo & Chen (2016). Our approach computes the test statistic using estimates from the model under the null hypothesis instead of the full model, which allows the nuisance parameter to be in high dimension. Unlike many existing

methods, our method does not require a sparsity assumption on the testing parameter. This relaxation makes the proposed method applicable to many practical problems where the sparsity assumption may not be reasonable. Importantly, the limiting distribution of our test statistic has a closed form, which is computationally attractive.

The proposed approach has some important applications. For example, it can be used to test whether complex models are necessary, or whether simple models with linear terms only will give accurate results. High-dimensional models, such as partial linear additive models (Lian et al., 2011; Maidman & Wang, 2018) and quadratic regression models (Hao et al., 2018), sometimes require terms in addition to the linear terms; for example, partial linear additive models may require spline functions, and quadratic regression models may require quadratic and interaction terms. However, these additional terms require extra efforts in estimation, and the increased estimation errors may offset the improvement in prediction. Thus, it is crucial to first examine if these additional terms are needed. Our method provides an inferential tool for answering this important question. In other words, the proposed test can serve as a goodness-of-fit test to compare two high-dimensional nested models in practical applications.

However, the proposed method does have some limitations. First, under the alternative hypothesis, the proposed method must perform its estimation for a pseudo parameter that comes from the misspecified model. This pseudo parameter is not the parameter in the true model, but the estimator must still converge to this pseudo parameter. Therefore, an additional assumption on the convergence of the estimator to the pseudo parameter is needed; see Assumption 6. This assumption is empirically difficult to verify for an arbitrary generalized linear model in practical applications. Second, in order for the proposed test to have non-trivial power asymptotically, the dimension of the testing parameter is only allowed to grow polynomially with the sample size. For more details, please see the discussion after Theorem 3.

The following notations will be used in this paper. X denotes a random variable or a vector of random variables, X is the design matrix, and x is a observation of X. A p dimensional random vector X is sub-Gaussian with variance proxy σ^2, if for any $v \in \mathbb{R}^p$ such that $\|v\|_2 = 1$, $P(|v^T X| > t) \leq 2\exp(-t^2/2\sigma^2)$ for every $t \geq 0$. The smallest and largest eigenvalue of matrix A are $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ respectively. For $\beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{p_\beta}$, we define $||\beta||_d = (\sum_{i=1}^{p_\beta} |\beta_i|^d)^{1/d}$ for $1 \leq d < \infty$. $||\beta||_0$ represents the number of nonzero coordinates of β. For a matrix $A = [a_{ij}]$, its maximum norm is $\|A\|_\infty = \max_{ij}|a_{ij}|$. We denote $a_n \ll b_n$ if $a_n = o(b_n)$; $a_n \gg b_n$ if $b_n = o(a_n)$; $a_n = O(b_n)$, if $a_n \leq Cb_n$ for some $C > 0$; and $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. For a given likelihood function $l(\theta)$ of $\theta = (\gamma^T, \beta^T)^T$, we define the gradient $\bigtriangledown_\theta l(\theta) = \partial l(\theta)/\partial\theta$ with partitions $\bigtriangledown_\gamma l(\theta)$ and $\bigtriangledown_\beta l(\theta)$, and the Hessian matrix $I_\theta = E\{\partial^2 l(\theta)/\partial\theta^2\}$, where $I_{\gamma\beta}$ and $I_{\beta\beta}$ are the corresponding partitions of I_θ.

2. METHODS

Assume the random scalar Y is from an exponential family with the probability density function taking the canonical form $f_Y(y;\eta) = \exp\{y\eta - b(\eta) + c(y)\}$ for known functions $b(\cdot)$ and $c(\cdot)$, and canonical parameter η. The dispersion parameter is not considered because we only need to model the mean regression and estimate regression coefficients. However, the results in this paper are still valid if the dispersion parameter is considered (McCullagh & Nelder, 1989). We are interested in the class of generalized linear models:

$$g(\mu) = z^T\gamma + w^T\beta, \tag{1}$$

where $\mu = E(Y \mid Z = z, W = w) = b'\{\eta(z,w)\}$, $g(\cdot)$ is the link function, Z and W are vectors of random variable, and $\gamma \in \mathbb{R}^{p_\gamma}$ and $\beta \in \mathbb{R}^{p_\beta}$ are corresponding vectors of coefficients respectively.

We aim to have simultaneous inference on parameter β, and treat γ as a nuisance parameter. Without the loss of generality, the hypothesis to be tested is:

$$H_0 : \beta = 0 \text{ versus } H_a : \beta \neq 0. \tag{2}$$

It is straightforward to extend the method to test $H_0 : \beta = \beta_\emptyset$ for other specific hypothesized values $\beta_\emptyset$. In this article, the subscript $\emptyset$ stands for the model under the null hypothesis. As discussed in Introduction

Section, when the dimension of γ grows with the sample size, there lacks statistical methods to test a high-dimensional β. Consequently, we concentrate on that both γ and β in model (1) are high-dimensional in this paper.

Suppose $\{(z_i, w_i, y_i), i = 1, \ldots, n\}$ are independent and identically distributed copies of (Z, W, Y). The classical score test is not applicable for testing hypothesis (2) in the high-dimensional setting because the limiting distribution of the Hessian matrix is intractable. To address this issue, Geoman et al. (2011) proposed a standardized score function in quadratic form to test high-dimensional β with low-dimensional γ. Under model (1), the specific form of the test statistics of Geoman et al. (2011) is $[1/n\sum_{i=1}^n\sum_{j=1}^n\{(y_i-\mu_{\emptyset i})(y_j-\mu_{\emptyset j})w_i^Tw_j\}]/[1/n\sum_{i=1}^n\{(y_i-\mu_{\emptyset i})^2w_i^Tw_i\}]$, where $\mu_{\emptyset i}$ is the conditional expectation $E(Y \mid Z = z_i)$ under the null hypothesis. Later, this test statistic was modified by Guo & Chen (2016) to be $1/n\sum_{i\neq j}^n\{(y_i-\mu_{\emptyset i})(y_j-\mu_{\emptyset j})w_i^Tw_j\}$, to improve the performance of the test for diverging p_β. It also allows for a wide range of link functions.

The methods by Geoman et al. (2011) and Guo & Chen (2016) were designed for testing hypothesis (2) for a low-dimensional γ. We extend their works for the case that γ is high-dimensional. We propose an asymptotic α-level test that rejects H_0 if $|\hat{U}_n|/\sqrt{2\hat{R}_n} > z_{1-\alpha/2}$ where

$$\hat{U}_n = \frac{1}{n}\sum_{i\neq j}^n\{(y_i-\hat{\mu}_{\emptyset i})(y_j-\hat{\mu}_{\emptyset j})w_i^Tw_j\},\ \ \hat{R}_n = \frac{1}{n(n-1)}\sum_{i\neq j}^n\{(y_i-\hat{\mu}_{\emptyset i})^2(y_j-\hat{\mu}_{\emptyset j})^2(w_i^Tw_j)^2\},$$

$\hat{\mu}_{\emptyset i} = g^{-1}(z_i^T\hat{\gamma}_\emptyset)$, and $\hat{\gamma}_\emptyset$ is the estimate from the model under the null hypothesis $g(\mu) = z^T\gamma$.

Compared with some existing works, the proposed approach has certain advantages. Using a similar spirit to the profile score function in a low-dimensional setting, Ning & Liu (2017) developed a decorrelated score function to test hypothesis (2) by projecting the score function with respect to β to the linear span of the score function with respect to γ, which has the form $S(\beta) = \bigtriangledown_\beta l(\gamma,\beta) - \{I_{\beta\gamma}I_{\gamma\gamma}^{-1}\}\bigtriangledown_\gamma l(\gamma,\beta)$, where $l(\gamma,\beta)$ is a general loss function. In their work, they need a consistent estimator for (γ,β) with fast convergence rate. Thus, their method needs to impose a sparsity condition on β. Moreover, the methods by Ning & Liu (2017) and Sun & Zhang (2020) also need to estimate $I_{\beta\gamma}I_{\gamma\gamma}^{-1}$ to construct the test statistics. Similarly, the works involving desparsified Lasso approach (Zhang & Zhang, 2014; van de Geer et al., 2014; Zhang & Cheng, 2017) require estimating $I^{-1}_{(\gamma,\beta)}$. They proposed to solve a column-wise penalized optimization problem to obtain an estimator for the Hessian matrix, which could be computationally demanding in high-dimensional setting. For example, when $p_\gamma = p_\beta = 1000$, a total of 1000 minimization problems has to be solved, each of which estimates one column of $I_{\beta\gamma}I_{\gamma\gamma}^{-1}$. For each optimization problem, it also needs to search for the optimal tuning parameter. In addition, to test the high-dimensional parameter, the method by Ning & Liu (2017) and Zhang & Cheng (2017) requires a multiplier bootstrap to approximate the limiting distribution of the test statistic. In contrast, our method does not have any constraint on β since we only need to estimate γ from the model under the null hypothesis. Moreover, our approach is computationally more efficient, as it only needs a single fitting for the model including nuisance parameter only, and the limiting distribution of the test statistics has a closed-form.

In comparison with Geoman et al. (2011) and Guo & Chen (2016), a crucial difference is that they only allow the nuisance parameter γ to be low-dimensional, but we allow its dimension to diverge. Thus, different from their approaches, we need to estimate γ in high dimension to construct the test statistics. We obtain the estimate $\hat{\gamma}_\emptyset$ by solving a penalized likelihood problem

$$\hat{\gamma}_\emptyset = \underset{\gamma}{\operatorname{argmin}}\left(-\frac{1}{n}\sum_{i=1}^n\left[y_i\eta(z_i^T\gamma) - b\{\eta(z_i^T\gamma)\}\right] + \zeta||\gamma||_1\right), \tag{3}$$

where we impose an L_1-penalty on γ in order to have a sparse solution, and ζ is the tuning parameter. The optimization (3) can be achieved by standard coordinate gradient descent algorithms, which have been shown to be numerically stable and efficient (Friedman et al., 2010). Additionally, there are several other approaches can be used to obtain $\hat{\gamma}_\emptyset$, including the Dantzig selector (Candes & Tao, 2012), square-root lasso (Belloni et al., 2011), or scaled lasso (Sun & Zhang, 2012). For all these methods, an important job in

practice is to select the optimal value for the tuning parameter, which impacts the convergence rate of the estimates. There are various methods for tuning parameter selection, such as K-fold cross validation and Bayesian information criteria, among many others. We use K-fold cross validation to choose the optimal value for ζ.

The high dimensionality in the nuisance parameter poses great challenges in understanding the theoretical properties of the test. The theories established in Geoman et al. (2011) and Guo & Chen (2016) are not directly applicable. We will elucidate this in the following section.

3. Asymptotic Properties

In this section, we will first build the asymptotic type I error rate of the proposed test, explore the consequence of model misspecification in high-dimensional setting, and then examine the power of the test. Denote $\theta = (\gamma^T, \beta^T)^T$ as the true model parameters, such that $||\gamma||_0 = s_\gamma$ and $||\beta||_0 = s_\beta$. Let $X = (Z^T, W^T)^T$ and $\omega_\gamma = \partial g^{-1}(z^T\gamma)/\partial(z^T\gamma)$. To build the asymptotic properties of the test, we need the following assumptions.

Assumption 1. Assume that Z and W are sub-Gaussian vectors, and $\varepsilon = Y - Z^T\gamma - W^T\beta$ is a sub-Gaussian variable.

Assumption 2. The inverse link function $g^{-1}(\cdot)$ is continuously differentiable.

Assumption 3. Under H_0, there exist positive constants c_1 and C_1 such that $0 < c_1 \leq \inf_{\gamma^* \in B(\gamma,r)} \lambda_{\min}\{E(\omega_{\gamma^*} XX^T)\} \leq \sup_{\gamma^* \in B(\gamma,r)} \lambda_{\max}\{E(\omega_{\gamma^*} XX^T)\} \leq C_1 < \infty$, where $B(\gamma, r)$ is a ball centered at γ with radius $r = C_2 s_\gamma \log p_\gamma / n$ for a constant C_2.

Assumption 4. Under H_0, it holds that $||\hat{\gamma}_\varnothing - \gamma||_2 = O_P\{(s_\gamma \log p_\gamma/n)^{1/2}\}$ and $s_\gamma \log p_\gamma / n = o(1)$.

Assumption 5. Under H_0, with probability close to 1, $\|\breve{\gamma}^{J_0^c}\|_1 \leq C_3 \|\breve{\gamma}^{J_0}\|_1$, for a positive constant C_3, where $\breve{\gamma} = \hat{\gamma}_\varnothing - \gamma$, J_0 is the set of nonzero coefficients of γ, and $\breve{\gamma}^{J_0}$ and $\breve{\gamma}^{J_0^c}$ are sub-vectors of $\breve{\gamma}$ with indices in J_0 and J_0^c respectively.

The sub-Gaussian condition in Assumption 1 is widely used in the literature (Fan et al., 2017; Lugosi & Mendelison, 2019). Assumptions 2 is an ordinary regularity condition for generalized linear model. In particular, Assumption 3 requires that the eigenvalues of the expected Hessian matrix to be bounded away from zero and infinity when γ^* varies within a small neighborhood of γ. When $s_\gamma \log p_\gamma / n \to 0$, this assumption is mild since it is often assumed that $c < \lambda_{\min}\{E(\omega_\gamma XX^T)\} \leq \lambda_{\max}\{E(\omega_\gamma XX^T)\} < C$ for some positive constants c and C (van de Geer et al., 2014; Ning & Liu, 2017; Zhu & Bradic, 2018). Assumption 4 requires the estimation error of $\hat{\gamma}_\varnothing$ in the order of $O_P\{(s_\gamma \log p_\gamma/n)^{1/2}\}$. With a suitable choice of the tuning parameter, Bickel et al. (2009) and Buhlmann & Van de Geer (2011) have proved that the proposed estimate in (3) satisfies Assumption 4. As shown in the literature, such a convergence rate can also be achieved by Dantzig selector (Candes & Tao, 2012; Castro, 2013), square-root lasso (Belloni et al., 2011), or scaled lasso (Sun & Zhang, 2012; Lederer et al., 2019). Thus, any of these methods can be applied to obtain $\hat{\gamma}_\varnothing$. The inequality shown in Assumption 5 is general and leads to restricted eigenvalue condition and compatibility condition in high-dimensional literature (Bickel et al., 2009; Buhlmann & Van de Geer, 2011). Under these conditions, we are able to prove that the size of the proposed test is asymptotically at the nominal level. Define $\Lambda_W^\varepsilon = \mathrm{tr}[E\{\mathrm{var}(\varepsilon)WW^T\}^2]$, we have the following Theorem for the asymptotic size of the proposed test.

Theorem 1. *Under H_0, if Assumptions 1–5 hold, $s_\gamma \log p_\gamma / \sqrt{2\Lambda_W^\varepsilon} = o(1)$, and $s_\gamma \log p_\beta / n = o(1)$, then*

$$\lim_{n\to\infty} \sup_{\|\gamma\|_2 = O(1)} P\left(|\hat{U}_n| / \sqrt{2\hat{R}_n} > z_{1-\alpha/2}\right) = \alpha.$$

The result of Theorem 1 is impacted by the dimension p_β through the two conditions: $s_\gamma \log p_\gamma / \sqrt{2\Lambda_W^\varepsilon} = o(1)$ and $s_\gamma \log p_\beta / n = o(1)$. Let's assume a simple condition that $\Lambda_W^\varepsilon =$

$C_4 p_\beta^a$ with $a \in [1, 2)$ for a positive constant C_4, to understand the role of p_β in the condition $s_\gamma \log p_\gamma / \sqrt{2\Lambda_W^\varepsilon} = o(1)$. The constant a is related to the dependence structure among W, e.g., assume $a = 1$ if all components of W are uncorrelated with each other. A larger value of a is needed if correlations among W are stronger. It can be seen that the dimension p_β has to diverge to meet the condition $s_\gamma \log p_\gamma / \sqrt{2\Lambda_W^\varepsilon} = o(1)$. The sparsity of the nuisance parameter is constrained by three conditions: $s_\gamma \log p_\gamma / n = o(1)$, $s_\gamma \log p_\gamma / \sqrt{2\Lambda_W^\varepsilon} = o(1)$, and $s_\gamma \log p_\beta / n = o(1)$. The correct asymptotic type I error for the testing method by Zhang & Cheng (2017) requires ultra-sparsity assumption: $s \ll \sqrt{n}/\log p$. Instead, our proposed test is general and still has the correct type I error when $\sqrt{n}/\log p_\gamma \ll s_\gamma \ll n/\log p_\gamma$ and $p_\beta \geq C_5 n^{2/a}$ for a positive constant C_5. But it may not be the case for the Zhang and Cheng's method. Under the ultra-sparsity, $s_\gamma \log p_\gamma / \sqrt{n} = o(1)$, we have correct asymptotic type I error when $p_\beta \geq C_5 n^{1/a}$. The condition $s_\gamma \log p_\beta / n = o(1)$ define the relationship between the sparsity of nuisance parameter and the dimension of testing parameter: smaller s_γ allows larger p_β. For example, $p_\beta = O\{\exp(\sqrt{n})\}$ is allowed under the ultra-sparsity, $s_\gamma \log p_\gamma / \sqrt{n} = o(1)$. In summary, the condition $s_\gamma \log p_\gamma / \sqrt{2\Lambda_W^\varepsilon} = o(1)$ implies a lower bound, and the condition $s_\gamma \log p_\beta / n = o(1)$ implies a upper bound for p_β.

Theorem 1 implies that the validity of the proposed test does not beg for selection consistency or a "beta-min" condition on the minimal signal strength. Although the focus of this paper is the high-dimensional γ, Theorem 1 still applies when the dimension of γ is fixed. The classical approaches for fixed dimensional generalized linear model can be applied to obtain $\hat{\gamma}_\emptyset$, and it is natural to assume $||\hat{\gamma}_\emptyset - \gamma||_2^2 = O_P(1/n)$. In this special case, the problem is similar to the one considered by Guo & Chen (2016). However, the related asympotic behavior of the test statistics in this article does not automatically reduce to that by Guo & Chen (2016), because there are differences in model specifications and assumptions between this article and Guo & Chen (2016).

The rest of this section examines the asymptotic power of the proposed test. We first introduce some notations and assumptions. Let $\gamma_\emptyset$ be the minimizer of the Kullback-Leibler divergence defined as $E\{L_0 - L_\emptyset(\gamma)\}$, where L_0 is the log-likelihood function for the true model under the alternative hypothesis and $L_\emptyset(\gamma)$ is the log-likelihood function for a working model that only uses Z to predict Y. Under Assumption 2, Theorem 5 in Lv & Liu (2014) shows that $\gamma_\emptyset$ also minimizes $E\{-L_\emptyset(\gamma)\}$. We propose to use the M-estimator $\hat{\gamma}_\emptyset$ defined in (3) to estimate $\gamma_\emptyset$. Denote $\theta_\emptyset = (\gamma_\emptyset^T, 0^T)^T$, $\tilde{\theta} = \theta - \theta_\emptyset = (\gamma^T - \gamma_\emptyset^T, \beta^T)^T$, $\omega_\theta = \partial g^{-1}(x^T\theta)/\partial(x^T\theta)$, and $\omega_\theta^* = \{g^{-1}(x^T\theta) - g^{-1}(x^T\theta_\emptyset)\}/x^T\tilde{\theta}$. We consider the following assumptions.

Assumption 6. Under H_a, the estimate $\hat{\gamma}_\emptyset$ satisfies $||\hat{\gamma}_\emptyset - \gamma_\emptyset||_2 = O_P\{(s_\gamma^\emptyset \log p_\gamma / n)^{1/2}\}$ and $s_\gamma^\emptyset \log p_\gamma / n = o(1)$, where $s_\gamma^\emptyset = ||\gamma_\emptyset||_0$.

Assumption 7. $||\mathsf{X}\theta_0||_\infty = O(1)$ and $||\mathsf{Z}\boldsymbol{\gamma}_\emptyset||_\infty = O(1)$, where X and Z are the $n \times p$ and $n \times p_\gamma$ design matrices.

Assumption 8. Under H_a, there exist constants c_6 and C_6 such that $0 < c_6 \leq \lambda_{\min}\{E(\omega_\theta X X^T)\} \leq \lambda_{\max}\{E(\omega_\theta X X^T)\} \leq C_6 < \infty$, $0 < c_6 \leq \lambda_{\min}\{E(\omega_{\gamma_\emptyset} X X^T)\} \leq \lambda_{\max}\{E(\omega_{\gamma_\emptyset} X X^T)\} \leq C_6 < \infty$, and $0 < c_6 \leq \lambda_{\min}\{E(\omega_\theta^* X X^T)\}$.

Assumption 9. Under H_a, with probability close to 1, $\|\ddot{\gamma}^{J_\emptyset^{0c}}\|_1 \leq C_7 \|\ddot{\gamma}^{J_\emptyset^0}\|_1$, for a positive constant C_7, where $\ddot{\gamma} = \hat{\gamma}_\emptyset - \gamma_\emptyset$, and $J_\emptyset^0$ is the set of nonzero elements of $\gamma_\emptyset$.

Assumption 6 is an important assumption for deriving the power of the proposed test. It requires $\hat{\gamma}_\emptyset$ to converge to the pseudo parameter $\gamma_\emptyset$. In Assumption 6, we assume $\gamma_\emptyset$ is exactly sparse. It can be further relaxed to a weak sparsity assumption as to be discussed below Theorem 2. We acknowledge that this assumption is unusual, since most sparsity assumptions are imposed on the true parameter γ. For that reason, we quantify their difference $\tilde{\gamma} = \gamma - \gamma_\emptyset$ in Theorem 2. Assumptions 7 and 8 are some regularity conditions. Assumption 9 is analogous to Assumption 5 to ensure $\ddot{\gamma}$ belongs to a cone when the alternative hypothesis holds.

THEOREM 2. *For generalized linear model (*1*), suppose* $E\{L_0 - L_\emptyset(\gamma)\} = O(1)$ *under* H_a*,*

(a) If Assumption 2 holds and $E(\omega_\theta^* ZZ^T)$ *is invertible, then* $\tilde{\gamma} = \gamma - \gamma_\emptyset = \{E(\omega_\theta^* ZZ^T)\}^{-1} E(\omega_\theta^* ZW^T)\beta$*;*

(b) If Assumptions 2 and 8 hold, then $||\tilde{\gamma}||_2 = O(||\beta||_2)$.

Despite the recent advances in high-dimensional statistics, few works (Buhlmann & Van de Geer , 2015) have studied the consequence of model missepecification for high-dimensional models. Theorem 2 can be viewed as an exploration in this direction. Theorem 2(a) shows the functional relationship between $\tilde{\gamma}$ and β. It requires a mild condition that $E(\omega_\theta^* ZZ^T)$ is invertible. It provides insights about Assumption 6. Theorem 2(b) shows that $||\tilde{\gamma}||_2$ is bounded by $O(||\beta||_2)$. Thus, when $||\beta||_2$ is small, $\gamma_\emptyset$ is still close to the true γ under H_a. It also implies that $\gamma_\emptyset$ could be weakly sparse, if not exactly sparse, in the sense that many of its elements are small. Based on Negahban et al. (2012), in Supplementary Materials, we show that under the weak sparsity assumption, we have $||\hat{\gamma}_\emptyset - \gamma_\emptyset||_2 = O_P\left\{\sqrt{R_q}(\log p_\gamma/n)^{1/2-q/4}\right\} = o_P(1)$ for some $q \in [0,1]$, where R_q is a weak sparsity measurement. Even though this convergence rate is different from the one in Assumption 6, we are still able to establish the power of our test as long as $||\hat{\gamma}_\emptyset - \gamma_\emptyset||_2 = o_P(1)$; see Lemma S2 in Supplementary Materials. Next, we give the asymptotic power of the proposed test.

THEOREM 3. *Under* H_a*, if Assumptions 1, 2, 6, 7, 8, 9 hold, and* $s_\gamma^\emptyset \log p_\beta/n = o(1)$*,*

$$\lim_{n\to\infty} \inf_{||\theta||_2 = O(1)} P\left(\frac{|\hat{U}_n|}{\sqrt{2\hat{R}_n}} > z_{1-\alpha/2}\right) = \begin{cases} \alpha \text{ if } \frac{n\|\beta\|_2^2}{\sqrt{2\Lambda_W^\varepsilon}} = o(1), \|\beta\|_2 = o(1), \text{ and } \frac{s_\gamma^\emptyset \log p_\gamma}{\sqrt{2\Lambda_W^\varepsilon}} = o(1); \\ 1 \text{ if } \frac{n\|\beta\|_2^2}{\sqrt{2\Lambda_W^\varepsilon}} \to \infty \text{ and } \frac{\sqrt{n}\|\beta\|_2^2}{\sqrt{2\Lambda_W^\varepsilon}} = O(1). \end{cases}$$

Theorem 3 indicates the power of the test is determined by the magnitude of three quantities: $n\|\beta\|_2^2/\sqrt{2\Lambda_W^\varepsilon}$, $\sqrt{n}\|\beta\|_2^2/\sqrt{2\Lambda_W^\varepsilon}$, and $s_\gamma^\emptyset \log p_\gamma/\sqrt{2\Lambda_W^\varepsilon}$. In principal, the power goes to 1 if $n\|\beta\|_2^2/\sqrt{2\Lambda_W^\varepsilon}$ diverges and $\sqrt{n}\|\beta\|_2^2/\sqrt{2\Lambda_W^\varepsilon}$ is bounded. The power is equal to α if $\|\beta\|_2$ decays in a rate such that $n\|\beta\|_2^2/\sqrt{2\Lambda_W^\varepsilon}$ vanishes when $s_\gamma^\emptyset \log p_\gamma/\sqrt{2\Lambda_W^\varepsilon} = o(1)$. To simplify the discussion, let us assume $\Lambda_W^\varepsilon \asymp p_\beta^a$ with $a \in [1,2)$. Under a reasonable assumption that $||\beta||_2 \asymp 1$, Theorem 3 implies that our method has non-trivial power when $p_\beta \asymp n^b$, where $1/a \le b < 2/a$. However, when β is exactly sparse, our proposed test requires stronger conditions than the maximum type tests (Chernozhukov et al., 2013; Zhang & Cheng, 2017) to have nontrivial power.

The exact sparsity in Assumption 6 provides theoretical advantages to examine the power behavior of the proposed test. Moreover, if the pseudo parameter $\gamma_\emptyset$ is not exactly sparse, the weak sparsity assumption would still enable us to build the asymptotic power as long as the associated convergence rate vanishes as the sample size grows, see Lemma S.2 in Supplementary Materials for the details.

It is known that a confidence set for regression parameter can be constructed by inverting the acceptance regions of hypothesis tests. Thus our testing procedure can also be used to construct confidence set for a high-dimensional parameter in generalized linear model. The desparsified lasso approach (Zhang & Zhang, 2014; van de Geer et al., 2014) provided confidence interval for an individual element of regression parameter. However, for a simultaneous inference on the multi-dimensional regression parameter, the width of confidence interval by desparsified lasso approach tends to be wide due to adjusting for multiple comparison. To achieve the asymptotical efficiency, the methods by Zhang & Zhang (2014), van de Geer et al. (2014), Zhang & Cheng (2017), and Ning & Liu (2017), require ultra-sparsity on all model parameters. If the regression parameter is dense, obtaining an unbiased estimator through these methods increase significantly the variance. Thus, an important implication of Theorems 1 and 3 is that: our proposed test is particularly useful for the case that the parameter β is dense and γ is in high dimension.

4. SIMULATIONS

We assess the performance of our method using simulations from a linear model, $Y = Z^T\gamma + W^T\beta + \varepsilon$ where $\varepsilon \sim N(0,1)$, and a logistic regression, $\text{logit}\{\text{p}(Y=1)\} = Z^T\gamma + W^T\beta$. For each model, we generate two scenarios for γ: sparse γ in Scenario 1, and dense γ in Scenario 2. Because the proposed

method only requires sparsity assumption on γ, Scenario 1 is used to assess the type I error and the local power of the proposed method when the sparsity assumption is satisfied, and Scenario 2 is designed to evaluate the robustness of our method when the sparsity assumption is violated.

In Scenario 1, we choose $p_\gamma = p_\beta$, and generate $X = (Z^T, W^T)^T$ from a multivariate Gaussian distribution $N(0, \Sigma)$, where the covariance matrix follows the Toeplitz design $\Sigma_{jk} = 0.6^{|j-k|}$. We let the first 5% elements of γ to be 0.5 and 1 for linear and logistic models respectively, and the rest be zeros. There are three choices for the values of β: Setting 1 (assessing type I error): all values of β are zeros; Setting 2 (assessing power with spare alternative): the first 5% elements of β has value 0.5 and 1 for linear and logistic models respectively, and the rest are zeros; Setting 3 (assessing power with dense alternative): all elements of β are equal such that $||\beta||_2 = ||\gamma||_2$. In each setting, we generate datasets for $n = 200$ and $p \in \{400, 4000, 8000\}$, where $p = p_\gamma + p_\beta$. For each combination of n and p, 500 datasets are generated. We fit the data simulated from Scenario 1 using two approaches: (1). the proposed method; (2). the multiplier bootstrap extension of the decorrelated score test by Ning & Liu (2017). We choose $\gamma = 0.05$ in all tests. The empirical type I error and power are shown in Figures 1 and 2.

In the second scenario, the nuisance parameter is dense with a small proportion of covariates have relatively strong signal while the rest are weak. Moreover, there are sparse interactions. Specifically, γ is dense with 10% of elements equal to 1, the rest equal to 0.1, and $Z_i \sim N(0, 1)$ for $i \in \{1, \ldots, p_\gamma\}$. The vector W is $p_\gamma(p_\gamma - 1)/2$ dimensional including all pairwise interaction terms derived from components in Z. To assessing the power of the test, we randomly assign 2% elements in β to be 1, and the rest to be 0. That is, the data structure does not follow the strong heredity condition in interactions. The goal is to apply the proposed method to test the existence of interactions with null hypothesis $H_0 : \beta = 0$. The sample size is 200. The number of covariates are $p_\gamma \in \{50, 100, 200\}$, which results in a very high dimension in interaction terms to be tested compared to the sample size, i.e., $p_\beta \in \{1225, 4950, 19900\}$ respectively. We generate 500 datasets for each value of p_γ, and only fit data simulated from Scenario 2 using the proposed method. The empirical type I error and power are summarized in Figure 3.

It can be seen from Figures 1 and 2 that the proposed method performs reasonably well in terms of the empirical type I error rate even the dimension of β is very high. It agrees with Theorem 1. The empirical power of the test shown in Figures 1 and 2 agrees with Theorem 3 in the following points: first, the power decreases when the dimension of p_β increases; 2. there is no significant difference in power between sparse and dense alternatives as long as $||\beta||_2$ stay the same. It can be seen that our method performs better than the method by Ning & Liu (2017), which is not surprising. An explanation is that the decorrelated test was originally designed for testing low-dimensional parameter with high-dimensional nuisance parameter by Ning & Liu (2017). It can be extended to simultaneous inference on high-dimensional parameter with high-dimensional nuisance utilizing the multiplier bootstrap procedure (Chernozhukov et al., 2013). However, this extension is prone to type I error inflation and loss of power. The power of method by Ning & Liu (2017) deteriorates sharply if the testing parameter is dense, while our method works adequately for testing dense parameter. In addition, as shown in Figure 3, the proposed method is robust to the violation of the sparsity assumption for parameter under null space, in terms of both the type I error rates and the power.

5. CHINESE FAMINE SAMPLE DATA: GENE-ENVIRONMENT INTERACTION

Schizophrenia is a severe psychiatric disorder with a global life-time risk around 1% and a typical onset in late adolescence and early adulthood. In collaboration with the University of Changchun in China, Boks et al. (2018) included schizophrenia patients and healthy controls that had been exposed to famine within the first 3 months of gestation based on a birth date between January 1960 and September 1961. A total of 74 schizophrenia patients and 79 healthy controls were assessed. Boks et al. (2018) focused on examining the role of changes in DNA methylation in the increased risk to develop schizophrenia after in utero exposure to famine. Data from this study have been deposited in the Gene Expression Omnibus repository under the accession number GSE116379. We are interested in utilizing this data to demonstrate the application of our method to identify gene-environment interactions.

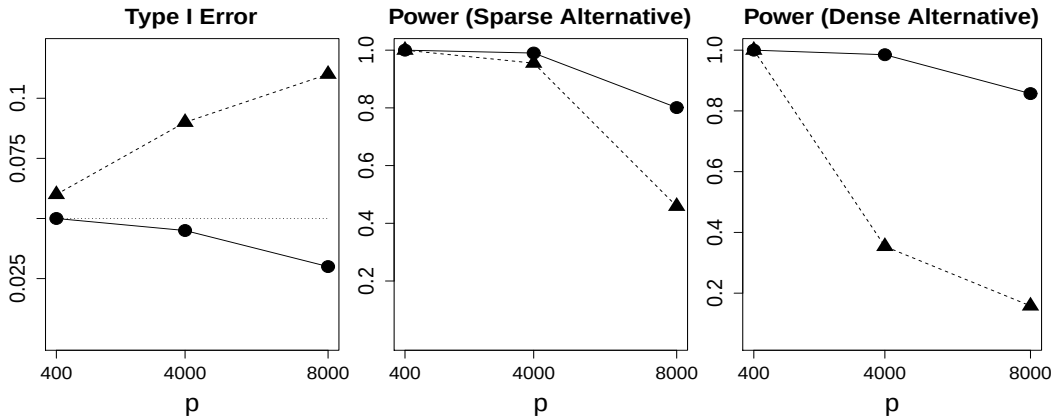

Fig. 1: Testing Results for Data Simulated from Linear Regression in Scenario 1. Solid Circle: results for the proposed method; Solid Triangle: results for the method by Ning and Liu (2017).

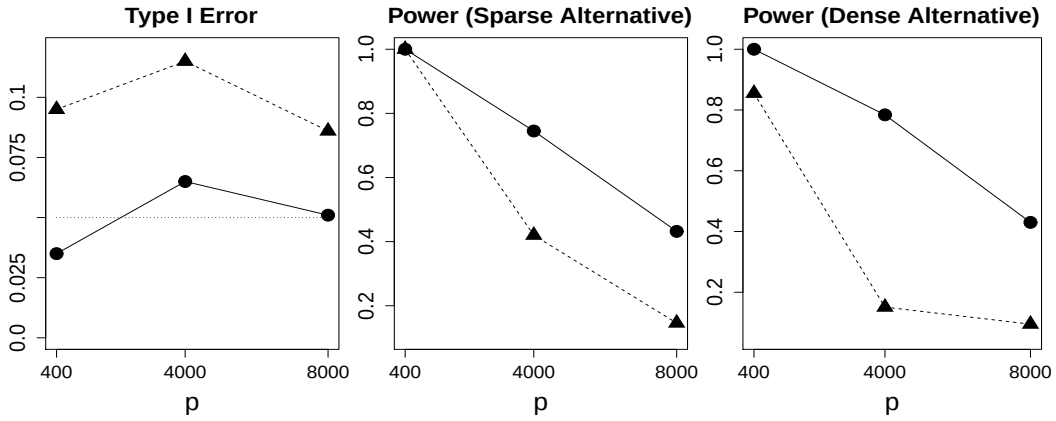

Fig. 2: Testing Results for Data Simulated from Logistic Regression in Scenario 1. Solid Circle: results for the proposed method propose; Solid Triangle: results for the method by Ning and Liu (2017).

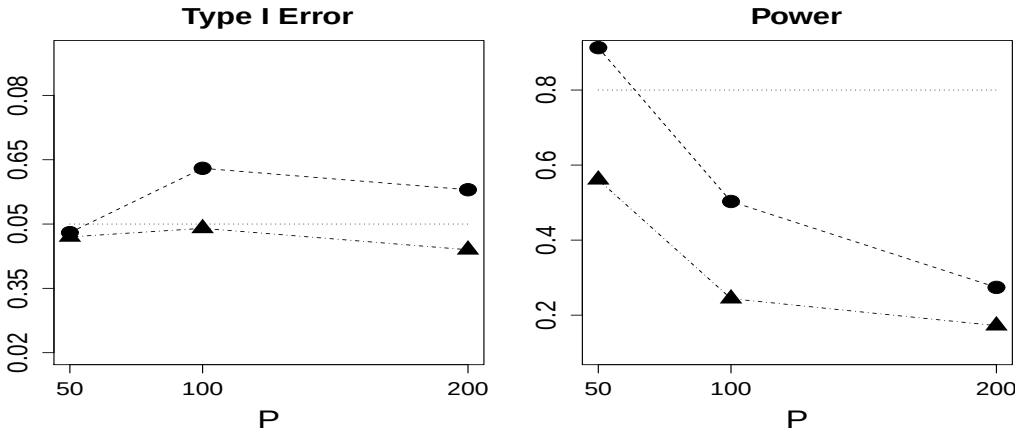

Fig. 3: Testing Results for Simulated Data in Scenario 2. P in x-axis is p_γ. Solid Circle: results for linear regression; Solid Triangle: results for logistic regression.

Table 1: The Number of Significant Gene-environment Interactions using Chinese Famine Sample Data Set. The significance is under Bonferroni adjustment: p-value$< 2.34 \times 10^{-5}$

Interaction	Total Number of Significant Interactions
DNA methylation sets $\times$ famine status	8
DNA methylation sets $\times$ gender	414
DNA methylation sets $\times$ city	12
DNA methylation sets $\times$ age	220

The phenotype variable is a binary indicator of schizophrenia. For DNA methylation, beta values (the ratio between methylated and unmethylated probe intensities as a measure of methylation percentage) were used in analysis. There are four environmental variables: famine (yes: 48; no: 105), gender (male: 76; female: 77), city (city: 119; rural: 34), and age (44-51). We evenly divide the total of 427291 methylations into 2137 sets, where each of 2136 sets has 200 methylations, and one set has 91 methylations. For each set, we are interested in testing if it has interaction with an environmental variable. Specifically, the model for the kth methylation set has the form:

$$\text{logit}\{\text{prob}(\text{schizophrenia})\}) = Z_e\gamma_{e_k} + Z_{g_k}^T\gamma_{g_k} + Z_e \times {Z_{g_k}}^T\beta_k,$$

where Z_e is the environmental variable, Z_{g_k} is the vector for the kth set, $Z_e \times Z_{g_k}$ denotes the interaction between the kth set and the environment variable, $(\gamma_{e_k}, \gamma_{g_k}^T, \beta_k^T)^T$ is the vector of coefficients, and $k \in \{1, \ldots, 2137\}$. The null hypothesis of interests is $H_0 : \beta_k = 0$. Taking the gene-famine interactions as an example, we conduct a total of 2137 tests of interactions between DNA methylation sets and famine status. This process is repeated for other three environmental variables.

We apply the proposed method to test the interactions between each DNA methylation set and each of the four environmental variables. Figure 3 shows the distributions of the p-values for each environmental variable. As shown in Table 1, among the 2137 tests, we identified that the 8, 414, 12, and 220 DNA methylation sets have significant interactions with famine status, gender, city/rural status, and age respectively after Bonferroni adjustment (i.e., p-value$< 2.34 \times 10^{-5}$).

ACKNOWLEDGEMENT

We are grateful for Dr. Yang Ning who kindly shared the R code for their paper (Ning & Liu, 2017). We thank for Dr. Cun-Hui Zhang for his helpful comments on an early version of this manuscript. We thank the editor, the associate editor and two referees for their constructive comments, which have led to a substantial improvement of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material available online includes the proofs of Theorem 3, Lemma S.1, and Lemma S.2.

APPENDIX

Proof of Theorem 1

We decompose $\hat{U}_n$ as

$$\hat{U}_n = \underbrace{\frac{1}{n}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(y_j-\mu_j)w_i^T w_j\}}_{I_{\hat{U}_n}} + \underbrace{\frac{1}{n}\sum_{i\neq j}^{n}\{(\mu_i-\hat{\mu}_{\emptyset i})(\mu_{0j}-\hat{\mu}_{\emptyset j})w_i^T w_j\}}_{II_{\hat{U}_n}} + \underbrace{2\frac{1}{n}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(\mu_j-\hat{\mu}_{\emptyset j})w_i^T w_j\}}_{III_{\hat{U}_n}}.$$

Let's first examine the term $II_{\hat{U}_n}$:

$$\frac{II_{\hat{U}_n}}{n} = \left[\frac{1}{n}\sum_{i=1}^{n}\{(\mu_i-\hat{\mu}_{\emptyset i})w_i^T\}\right]\left[\frac{1}{n}\sum_{i=1}^{n}\{w_i(\mu_i-\hat{\mu}_{\emptyset i})\}\right] - \frac{1}{n^2}\sum_{i=1}^{n}\{(\mu_i-\hat{\mu}_{\emptyset i})^2 w_i^T w_i\}$$

$$= \underbrace{(\hat{\gamma}_{\emptyset}-\gamma)^T\left[\frac{1}{n}\sum_{i=1}^{n}\{z_i\omega_{\gamma^*{}_i}w_i^T\}\right]\left[\frac{1}{n}\sum_{i=1}^{n}\{w_i\omega_{\gamma^*{}_i}z_i^T\}\right](\hat{\gamma}_{\emptyset}-\gamma)}_{II_1} - \underbrace{\frac{1}{n^2}\sum_{i=1}^{n}\{(\mu_i-\hat{\mu}_{\emptyset i})^2 w_i^T w_i\}}_{II_2}.$$

Let $\hat{\Sigma}_{z\omega w} = 1/n\sum_{i=1}^{n}\{z_i\omega_{\gamma^*{}_i}w_i^T\}$ and $\Sigma_{z\omega w} = E(Z\omega_{\gamma^*}W^T)$. For each $k\in\{1,\ldots,p\}$, X_k is sub-gaussian by Assumption 1. ω_{γ^*} is bounded by Assumption 2. It indicates that $\omega_{\gamma^*}X_k$ is also sub-gaussian. It is not difficulty to see that $X_k\omega_{\gamma^*}X_j$ is sub-exponential for $k,j\in\{1,\ldots,p\}$ because a product of two sub-gaussian variables is sub-exponential. From here, following the example 14.1 (page 491) and problem 14.3 (page 535) in Buhlmann & Van de Geer (2011), we can derive

$$\|\hat{\Sigma}_{z\omega w}-\Sigma_{z\omega w}\|_\infty = \tau = O_P\left[\sqrt{\{\log\max(p_\gamma,p_\beta)\}/n}\right]. \tag{A1}$$

Let b be a vector of size p_β not orthogonal to $\breve{\gamma}^T\hat{\Sigma}_{z\omega w}$ satisfying $d_1\leq\|b\|_1\leq D_1$ and $d_1\leq\|b\|_\infty\leq D_1$ for constants d_1 and D_1, we have

$$\begin{aligned}&\frac{\breve{\gamma}^T\hat{\Sigma}_{z\omega w}b}{\|\breve{\gamma}\|_2\|b\|_2} = \frac{\breve{\gamma}^T\Sigma_{z\omega w}b}{\|\breve{\gamma}\|_2\|b\|_2} + \frac{\breve{\gamma}^T(\hat{\Sigma}_{z\omega w}-\Sigma_{z\omega w})b}{\|\breve{\gamma}\|_2\|b\|_2} \leq \frac{\breve{\gamma}^T\Sigma_{z\omega w}b}{\|\breve{\gamma}\|_2\|b\|_2} + \tau\frac{\|\breve{\gamma}\|_1\|b\|_1}{\|\breve{\gamma}\|_2\|b\|_2} \leq D_2 + \tau\frac{\|\breve{\gamma}\|_1\|b\|_1}{\|\breve{\gamma}\|_2\|b\|_\infty}\\ &\leq D_2 + D_3\tau\frac{(1+C_3)\|\breve{\gamma}^{J_0}\|_1}{\|\breve{\gamma}^{J_0}\|_2} \leq D_2 + D_4\sqrt{\frac{s_\gamma\{\log\max(p_\gamma,p_\beta)\}}{n}} = D_2 + o(1),\end{aligned} \tag{A2}$$

for a positive constant D_2 by Assumptions 3 and 5 and bound in (A1). Then we can derive

$$D_2^2 \geq \frac{\breve{\gamma}^T\hat{\Sigma}_{z\omega w}bb^T\hat{\Sigma}_{z\omega w}^T\breve{\gamma}}{\|\breve{\gamma}\|_2^2\|b\|_2^2} = \frac{\lambda_{\max}(bb^T)\breve{\gamma}^T\hat{\Sigma}_{z\omega w}\hat{\Sigma}_{z\omega w}^T\breve{\gamma}}{\|\breve{\gamma}\|_2^2\|b\|_2^2} = \frac{\breve{\gamma}^T\hat{\Sigma}_{z\omega w}\hat{\Sigma}_{z\omega w}^T\breve{\gamma}}{\|\breve{\gamma}\|_2^2},$$

which implies $|II_1| = O_P(\|\breve{\gamma}\|_2^2) = O_P(s_\gamma\log p_\gamma/n)$.

Let $\underline{\mu} = (\mu_1,\ldots,\mu_n)$ and $\underline{\hat{\mu}_{\emptyset}} = (\hat{\mu}_{\emptyset 1},\ldots,\hat{\mu}_{\emptyset n})$. We have

$$nII_2 \leq \|\underline{\mu}-\underline{\hat{\mu}_{\emptyset}}\|_\infty^2\frac{1}{n}\sum_{i=1}^{n}(w_i^T w_i).$$

Next, we will prove that $\|\underline{\mu} - \underline{\hat{\mu}}_{\emptyset}\|_\infty = o_P(1)$. Let $\xi_n = D_5 s_\gamma \log p_\gamma / n$ for a positive constant D_5, for $t > 0$, we can derive

$$
\begin{aligned}
&P(\|\underline{\mu} - \underline{\hat{\mu}}_{\emptyset}\|_\infty > t) = P(\|\underline{\mu} - \underline{\hat{\mu}}_{\emptyset}\|_\infty > t, \|\breve{\gamma}\|_2^2 \le \xi_n) + P(\|\underline{\mu} - \underline{\hat{\mu}}_{\emptyset}\|_\infty > t, \|\breve{\gamma}\|_2^2 > \xi_n) \\
&= P(\|\underline{\mu} - \underline{\hat{\mu}}_{\emptyset}\|_\infty > t \mid \|\breve{\gamma}\|_2^2 \le \xi_n) P(\|\breve{\gamma}\|_2^2 \le \xi_n) + P(\|\underline{\mu} - \underline{\hat{\mu}}_{\emptyset}\|_\infty > t \mid \|\breve{\gamma}\|_2^2 > \xi_n) P(\|\breve{\gamma}\|_2^2 > \xi_n) \\
&\le P(\|\underline{\mu} - \underline{\hat{\mu}}_{\emptyset}\|_\infty > t \mid \|\breve{\gamma}\|_2^2 \le \xi_n) + P(\|\breve{\gamma}\|_2^2 > \xi_n) \\
&\le \sum_{i=1}^{n} P(|\mu_i - \hat{\mu}_{\emptyset i}| > t \mid \|\breve{\gamma}\|_2^2 \le \xi_n) + P(\|\breve{\gamma}\|_2^2 > \xi_n) \\
&= nP\{|g^{-1}(Z^T\gamma) - g^{-1}(Z^T\hat{\gamma}_{\emptyset})| > t \mid \|\breve{\gamma}\|_2^2 \le \xi_n\} + P(\|\breve{\gamma}\|_2^2 > \xi_n) \\
&\le nP\left(|\breve{\gamma}^T Z| > D_6^{-1} t \mid \|\breve{\gamma}\|_2^2 \le \xi_n\right) + P\left(\|\breve{\gamma}\|_2^2 > \xi_n\right) \\
&\le 2n \exp\left(-\frac{t^2}{2D_6^2 \sigma^2 \|\breve{\gamma}\|_2^2}\right) + P\left(\|\breve{\gamma}\|_2^2 > \xi_n\right) \\
&\le 2n \exp\left(-\frac{t^2}{2D_6^2 \sigma^2 \xi_n}\right) + P\left(\|\breve{\gamma}\|_2^2 > \xi_n\right) \\
&= 2n \exp\left(-\frac{nt^2}{2D_6^2 D_5 \sigma^2 s_\gamma \log p_\gamma}\right) + P\left(\|\breve{\gamma}\|_2^2 > \xi_n\right) \to 0,
\end{aligned}
$$

because $n/(s_\gamma \log p_\gamma) \to \infty$ and $P(\|\breve{\gamma}\|_2^2 > \xi_n) = o(1)$ by Assumption 4. Then, we conclude $\|\underline{\mu} - \underline{\hat{\mu}}_{\emptyset}\|_\infty = o_P(1)$. Since $n^{-1}\sum_{i=1}^n (w_i^T w_i) = O_P(\sqrt{2\Lambda_W^{\varepsilon}})$, it follows that $nII_2 = o_P(\sqrt{2\Lambda_W^{\varepsilon}})$. Combining the bounds of II_1 and II_2, we conclude that $II_{\hat{U}_n} = o_P\left(\sqrt{2\Lambda_W^{\varepsilon}}\right)$ when $s_\gamma \log p_\gamma / \sqrt{2\Lambda_W^{\varepsilon}} = o(1)$.

Then, we examine the term $III_{\hat{U}_n}$:

$$
\begin{aligned}
\frac{III_{\hat{U}_n}}{n} &= \left[\frac{1}{n}\sum_{i=1}^{n}\{(\mu_i - \hat{\mu}_{\emptyset i}) w_i^T\}\right]\left[\frac{1}{n}\sum_{i=1}^{n}\{w_i(y_i - \mu_i)\}\right] - \frac{1}{n^2}\sum_{i=1}^{n}\{(y_i - \mu_i)(\mu_i - \hat{\mu}_{\emptyset i}) w_i^T w_i\} \\
&= \underbrace{(\hat{\gamma}_{\emptyset} - \gamma)^T\left[\frac{1}{n}\sum_{i=1}^{n}\{z_i \omega_{\gamma^* i} w_i^T\}\right]\left[\frac{1}{n}\sum_{i=1}^{n}\{w_i(y_i - \mu_i)\}\right]}_{III_1} \underbrace{- \frac{1}{n^2}\sum_{i=1}^{n}\{(y_i - \mu_i)(\mu_i - \hat{\mu}_{\emptyset i}) w_i^T w_i\}}_{III_2}.
\end{aligned}
$$

Denote $\varsigma = \frac{1}{n}\sum_{i=1}^n \{w_i(y_i - \mu_i)\}$, it can be seen $\|\varsigma\|_2^2 = O_P(n^{-1}\sqrt{2\Lambda_W^{\varepsilon}})$ because $\|\varsigma\|_2^2 = n^{-1}[I_{\hat{U}_n} + 1/n\sum_{i=1}^n \{(y_i - \mu_i)^2 w_i^T w_i\}]$, $I_{\hat{U}_n} = O_P(\sqrt{2\Lambda_W^{\varepsilon}})$, and $1/n\sum_{i=1}^n \{(y_i - \mu_i)^2 w_i^T w_i\} = O_P(\sqrt{2\Lambda_W^{\varepsilon}})$. Following the same argument as (A2) and applying Cauchy-Schwarz inequality, we have

$$
|III_1| = |\breve{\gamma}^T \hat{\Sigma}_{z\omega w}\varsigma| \le \|\breve{\gamma}^T \hat{\Sigma}_{z\omega w}\|_2 \|\varsigma\|_2 \le D_7 \|\breve{\gamma}\|_2 \|\varsigma\|_2
$$

for a positive constant D_7, which implies $III_1 = O_P(n^{-1}\sqrt{s_\gamma \log p_\gamma}\sqrt[4]{2\Lambda_W^{\varepsilon}})$. For III_2, let $\underline{y} = (y_i, \cdots, y_n)$, we have

$$
nIII_2 \le \frac{1}{n}\sum_{i=1}^{n}\{|(y_i - \mu_i)(\mu_i - \hat{\mu}_{\emptyset i})| w_i^T w_i\} \le \|\underline{y} - \underline{\mu}\|_\infty \|\underline{\mu} - \underline{\hat{\mu}}_{\emptyset}\|_\infty \frac{1}{n}\sum_{i=1}^{n}(w_i^T w_i) = o_P(\sqrt{2\Lambda_W^{\varepsilon}}),
$$

since $\|\underline{y} - \underline{\mu}\|_\infty = O_P(1)$ and $\|\underline{\mu} - \underline{\hat{\mu}}_{\emptyset}\|_\infty = o_P(1)$. In summary, combining the bounds of III_1 and III_2, we have $III_{\hat{U}_n} = o_P(\sqrt{2\Lambda_W^{\varepsilon}})$ when $s_\gamma \log p_\gamma / \sqrt{2\Lambda_W^{\varepsilon}} = o(1)$.

Following proof of Theorem 3 in Guo and Chen (2016), it can be seen

$$
I_{\hat{U}_n} / \sqrt{2\Lambda_W^{\varepsilon}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty.
$$

More detailed proof of this result can be found in Lemma S.1 in the Supplementary Materials. Then, it follows that:

$$P\left(\frac{|\hat{U}_n|}{\sqrt{2\Lambda_W^{\varepsilon}}} > z_{1-\alpha/2}\right) = P\left[\left\{\frac{|I_{\hat{U}_n}|}{\sqrt{2\Lambda_W^{\varepsilon}}} + o_p(1)\right\} > z_{1-\alpha/2}\right] \to \alpha$$

For $\hat{R}_n$, it can be seen

$$\begin{aligned}
\hat{R}_n &= \frac{1}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i+\mu_i-\hat{\mu}_{\varnothing i})^2(y_j-\mu_j+\mu_j-\hat{\mu}_{\varnothing j})^2(w_i^Tw_j)^2\} \\
&= \underbrace{\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)^2(y_j-\mu_j)^2(w_i^Tw_j)^2\}}_{I_{\hat{R}_n}} + 4\underbrace{\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)^2(y_j-\mu_j)(\mu_j-\hat{\mu}_{\varnothing j})(w_i^Tw_j)^2\}}_{II_{\hat{R}_n}} \\
&+ 2\underbrace{\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)^2(\mu_j-\hat{\mu}_{\varnothing j})^2(w_i^Tw_j)^2\}}_{III_{\hat{R}_n}} + 4\underbrace{\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(\mu_i-\hat{\mu}_{\varnothing i})(\mu_j-\hat{\mu}_{\varnothing j})^2(w_i^Tw_j)^2\}}_{IV_{\hat{R}_n}} \\
&+ \underbrace{4\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\{(y_i-\mu_i)(\mu_i-\hat{\mu}_{\varnothing i})(y_j-\mu_j)(\mu_j-\hat{\mu}_{\varnothing j})(w_i^Tw_j)^2\}}_{V_{\hat{R}_n}} + \underbrace{\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\{(\mu_i-\hat{\mu}_{\varnothing i})^2(\mu_j-\hat{\mu}_{\varnothing j})^2(w_i^Tw_j)^2\}}_{VI_{\hat{R}_n}}.
\end{aligned}$$

We first examine the term $V_{\hat{R}_n}$ and derive

$$\begin{aligned}
|V_{\hat{R}_n}| &\leq 4\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\{|(y_i-\mu_i)(\mu_i-\hat{\mu}_{\varnothing i})(y_j-\mu_j)(\mu_j-\hat{\mu}_{\varnothing j})|(w_i^Tw_j)^2\} \\
&\leq 4\|\underline{y}-\underline{\mu}\|_{\infty}^2\|\underline{\mu}-\underline{\hat{\mu}}_{\varnothing}\|_{\infty}^2\underbrace{\frac{1}{n(n-1)}\sum_{i\neq j}^{n}\{(w_i^Tw_j)^2\}}_{\tilde{V}_{\hat{R}_n}},
\end{aligned}$$

which implies $|V_{\hat{R}_n}| = o_P(|\tilde{V}_{\hat{R}_n}|)$ because $\|\underline{y}-\underline{\mu}\|_{\infty} = O_P(1)$ and $\|\underline{\mu}-\underline{\hat{\mu}}_{\varnothing}\|_{\infty} = o_P(1)$.

The term $\tilde{V}_{\hat{R}_n}$ is a U-statistic. Define $H_{\tilde{V}_{\hat{R}_n}}(W_1,W_2) = (w_1^Tw_2)^2$. It is not difficult to see that $E(\tilde{V}_{\hat{R}_n}) = O(\Lambda_W^{\varepsilon})$. Following Hoeffding decomposition in the variance evaluation of U-statistic (Hoeffding, 1948), we have

$$H1_{\tilde{V}_{\hat{R}_n}} = E\{H_{\tilde{V}_{\hat{R}_n}}(W_1,W_2)|W_1\} = \mathrm{tr}\{w_1w_1^TE(WW^T)\},$$

and

$$\mathrm{Var}(\tilde{V}_{\hat{R}_n}) = \frac{4(n-2)}{n(n-1)}\mathrm{Var}(H1_{\tilde{V}_{\hat{R}_n}}) + \frac{2}{n(n-1)}\mathrm{Var}\{H_{\tilde{V}_{\hat{R}_n}}(W_1,W_2)\} = O[n^{-1}\mathrm{tr}^2\{E(WW^T)^2\}].$$

It implies that $\tilde{V}_{\hat{R}_n} = O_P(\Lambda_W^{\varepsilon})$, and we concludes $V_{\hat{R}_n}/\Lambda_W^{\varepsilon} \xrightarrow{p} 0$ when $n \to \infty$.

Similarly, we can found $II_{\hat{R}_n}/\Lambda_W^{\varepsilon} \xrightarrow{p} 0$, $III_{\hat{R}_n}/\Lambda_W^{\varepsilon} \xrightarrow{p} 0$, $IV_{\hat{R}_n}/\Lambda_W^{\varepsilon} \xrightarrow{p} 0$, $VI_{\hat{R}_n}/\Lambda_W^{\varepsilon} \xrightarrow{p} 0$ when $n \to \infty$.

In summary, we conclude that: $\hat{R}_n/\Lambda_W^{\varepsilon} \xrightarrow{p} 1$ as $n \to \infty$.

By Slutsky's Theorem, we have $\hat{U}_n/\sqrt{2\hat{R}_n}$ converge in distribution to $\hat{U}_n/\sqrt{2\Lambda_W^{\varepsilon}}$. It implies, as $n \to \infty$, we have

$$P\left(|\hat{U}_n|/\sqrt{2\hat{R}_n} > z_{1-\alpha/2}\right) \to \alpha.$$

The proof is valid for any point $\gamma \in H_0$ as long as $\|\gamma\|_2^2 = O(1)$, the uniform convergence of type I error rate follows.

Proof of Theorem 2

For generalized linear model with canonical link, the population score equations lead to $E\{Zg^{-1}(X^T\theta_0)\} = 0$ and $E\{Zg^{-1}(X^T\theta_\varnothing^0)\} = 0$. That is $E\{Zg^{-1}(X^T\theta_0)\} - E\{Zg^{-1}(X^T\theta_\varnothing^0)\} = 0$. By Assumption 2, the function $f(\theta) = E\{Zg^{-1}(X^T\theta)\} : \theta \to \mathbb{R}^{p_\gamma}$ is continuously differentiable. By the mean value theorem on vector valued functions (Rudin, 1976), we derive

$$f(\theta_0) - f(\theta_\varnothing^0) = \left[\int_0^1 \left\{Df(\theta) \mid_{\theta=\theta_\varnothing^0+t\tilde{\theta}}\right\} dt\right] \tilde{\theta} = \left\{\int_0^1 Df(\theta_\varnothing^0 + t\tilde{\theta})dt\right\} \tilde{\theta} = 0$$

where Df denote the Jacobin matrix of f that the integral is componentwise. The component in the kth row and lth column of Df is

$$Df_{kl}(\theta_\varnothing^0 + t\tilde{\theta}) = E\left[Z_k X_l \frac{dg^{-1}\{X^T(\theta_\varnothing^0 + t\tilde{\theta})\}}{d\{X^T(\theta_\varnothing^0 + t\tilde{\theta})\}}\right].$$

We have

$$\int_0^1 Df_{kl}(\theta_\varnothing^0 + t\tilde{\theta})dt = E\left[Z_k X_l \frac{g^{-1}\{X^T(\theta_\varnothing^0 + t\tilde{\theta})\}}{X^T\tilde{\theta}}\right]\Bigg|_0^1 = E\left\{Z_k X_l \frac{g^{-1}(X^T\theta_0) - g^{-1}(X^T\theta_\varnothing^0)}{X^T\tilde{\theta}}\right\}.$$

That is $E\left\{ZX^T\omega_\theta^*\right\}\tilde{\theta} = 0$, which implies

$$E(\omega_\theta^* ZZ^T)\tilde{\gamma} = E(\omega_\theta^* ZW^T)\beta_0. \tag{A3}$$

The singular matrix decomposition of a $l \times m$ matrix A is the factorization of A into the product form $A = UDV^T$. The $l \times l$ matrix U and $m \times m$ matrix V have orthonormal columns, and absolute value 1 eigenvalues. The matrix D is diagonal with positive real entries. The diagonal entries of D are known as singular values of A. By Assumption 8, the eigenvalues of $E_0(\omega_\theta^* ZZ^T)$ and $E_0(\omega_\theta^* WW^T)$ are bounded because $E_0(\omega_\theta^* ZZ^T)$ and $E_0(w_{\theta^*} WW^T)$ are principal sub-matrices of $E_0(\omega_\theta^* XX^T)$. Applying Hölder's inequality, we can derive that the largest singular value for $E(\omega_\theta^* ZW^T)$ is also bounded. We conclude $||\tilde{\gamma}||_2^2 = O(||\beta_0||_2^2)$ for generalized linear model based on equation (A3).

REFERENCES

BARNETT, I., MUKHERJEE, R., LIN, X. (2017). The generalized higher criticism for testing SNP-set effects in genetic association studies. *Journal of the American Statistical Association* **112**, 64–76.

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

BICKEL, P. J., RITOV, Y., TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Annals of Statistics* **37**, 1705–1732.

BIEN, J., TAYLOR, J., TIBSHIRANI, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics* **41**, 1111–1141.

BOKS, M. P., HOUTEPEN, L. C., XU, Z., HE, Y., ET AL. (2018). Genetic vulnerability to DUSP22 promoter hypermethylation is involved in the relation between in utero famine exposure and schizophrenia. *NPJ Schizophr* **4**, DOI: 10.1038/s41537-018-0058-4.

BUHLMANN, P., VAN DE GEER, S. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics* **9**, 1449–1473.

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

CANDES, E., TAO, T. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Annals of Statistics* **40**, 2389–2420.

CASTRO, Y. (2013). A remark on the lasso and Dantzig selector. *Statistics and Probability Letters* **83**, 304–314.

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

FAN, J., LI, Q., WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society, Series B* **79**, 247–265.

FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

GEOMAN, J. J., VAN HOUWELINGEN, H. C., FINOS, L. (2011). Tests againt a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika* **98**, 381–390.

GUO, B. & CHEN, S. X. (2016). Tests for high dimensional generalized linear models. *J. R. Statist. Soc.* B **78**, 1079–1102.

HAMADA, M. & WU, C.F. J. (1992). Analysis of designed experiments with complex aliasing *Journal of Quality Technology.* **24**, 130–137.

HAO, N., FENG, Y., ZHANG, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association* **113**, 615-625.

HAO, N., & ZHANG, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **109**, 1285-1301.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* **19**, 293–325.

HUANG, H., LIN, Y., CHEN, P., WANG C., HUANG, S., TZENG, J. (2016). Detection of genegene interactions using multistage sparse and low-rank regression. *Biometrics* **72**, 85-94.

LEDERER, J., YU, L., GAYNANOVA, I (2019). Oracle inequalities for high-dimensional prediction. *Bernoulli* **25**, 1225-1255.

LIAN, H., LIANG, H., RUPPERT, D. (2015). Seperation of covariates into nonparametric and parametric parts in high-dimensional partially linear additive models. *Statistica Sinica* **25**, 591-607.

LUGOSI, G., & MENDELISON, S. (2019). Sub-Gaussian estimators of the mean of a random vector. *Annals of Statistics* **47**, 783–794.

LV, J., & LIU, S. J. (2014). Model selection principle in missepecified models. *Journal of the Royal Statistical Society, Series B* **76**, 141–167.

MA, R., CAI, T., LI, H. (2019). Global and simultaneous hypothesis testing for high-dimensional logistic regression models.. *Journal of the American Statistical Association* https://doi.org/10.1080/01621459.2019.1699421

MAIDMAN, A., & WANG, L. (2018). New semiparametric models for predicting high-costs patients. *Biometrics* **74**, 1104-1111.

MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models.* New York: Chapman and Hall.

MEINSHAUSEN, N. & YU, B. (2009). Lasso-type recovery of sparse representation for high-dimensional data. *Annals of Statistics* **37**, 246–270.

NEGAHBAN, S., RAVIKUMAR, P., WAINWRIGHT, M., YU, B. (2012). A unified framework for high-dimentional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27**, 538–557.

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

RUDIN, W. (1976). *Principles of Mathematical Analysis.* New York: McGraw-Hill.

SHI, C., SONG, R., CHEN, Z., LI, R. (2019). Linear hypothesis testing for high dimensional generalized linear models. *Annals of Statistics* **47**, 2671–2703.

SUN, Q., & ZHANG, H. (2020). Targeted inference involving high-dimensional data using nuisance penalized regression. *Journal of the American Statistical Association* DOI: 10.1080/01621459.2020.1737079

SUN, T., & ZHANG, C. H. (2012). Scaled sparse linear regression. *Biometrika* **98**, 879–898.

SUR, P., & CANDES, E.J. (2019). A modern maximum-likelihood theory for highdimensional logistic regression. *Proceedings of the National Academy of Sciences* **116**, 14516–14525.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

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

WU, T. T., CHEN, Y. F, HASTIE, T., SOBEL, E., LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714–721.

WU, C., XU, G., SHEN, X., CHENG, G. (2020). A regularization-based adaptive test for high-dimensional generalized linear models. *Journal of Machine Learning Research* **21**, 1–67.

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

ZHANG, C. & ZHANG, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc.* B **76**, 217–242.

ZHU, Y. & BRADIC, J. (2018). Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association* **113**, 1583–1600.