ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2016)
78, *Part* 5, *pp.* 1079–1102

Tests for high dimensional generalized linear models

Bin Guo

Sichuan University, Chengdu, People's Republic of China

and Song Xi Chen

Peking University, Beijing, People's Republic of China, and Iowa State University, Ames, USA

[Received April 2014. Final revision October 2015]

Summary. We consider testing regression coefficients in high dimensional generalized linear models. By modifying the test statistic of Goeman and his colleagues for large but fixed dimensional settings, we propose a new test, based on an asymptotic analysis, that is applicable for diverging dimensions and is robust to accommodate a wide range of link functions. The power properties of the tests are evaluated asymptotically under two families of alternative hypotheses. In addition, a test in the presence of nuisance parameters is also proposed. The tests can provide *p*-values for testing significance of multiple gene sets, whose application is demonstrated in a case-study on lung cancer.

Keywords: Generalized linear model; Gene sets; High dimensional covariate; Nuisance parameter; *U*-statistics

1. Introduction

Owing to the surge of high dimensional data collection and analysis in bioinformatics and related fields, generalized linear models (McCullagh and Nelder, 1989) are widely used in high dimensional settings. High dimensionality can arise in at least two forms. The first form consists of multiple response variables but with low dimensional covariates. In this form, the responses represent readings for large numbers of genes and the covariates represent certain design and demographic variables. The second form consists of low dimensional response (an indicator for a disease) with high dimensional covariates, for instance gene expression levels. The current paper considers the second form of high dimensionality for generalized linear models.

The paper is focused on testing for significance of the regression coefficients in high dimensional generalized linear models, which has a range of applications including discovering significant gene sets. Statistical inference for generalized linear models under high dimensional settings has been the focus for a set of recent research references. van de Geer (2008) considered variable selection via a lasso approach. Fan and Song (2010) and Chang *et al.* (2013) proposed approaches via the sure independence screening of Fan and Lv (2008).

Biologically speaking, each gene does not function individually but tends to collaborate with other genes to achieve certain biological tasks. Gene sets are structured vocabularies in gene ontology (GO) systems that provide names of GO terms (sets of genes with shared annotation or functionality) (Barry *et al.*, 2008). The inferential context of gene set testing encounters

Address for correspondence: Song Xi Chen, Guanghua School of Management and Center for Statistical Science, Peking University, Beijing 100871, People's Republic of China.
E-mail: csx@gsm.pku.edu.cn

 1369–7412/16/781079

both high dimensionality and multiplicity, as the number of genes in a set can be much larger than the sample size and genes in different gene sets can overlap. These two features call for methods which can produce p-values for the significance of gene sets. For fixed dimensional data, the likelihood ratio test and the Wald test have been popular choices (McCullagh and Nelder, 1989). However, the high dimensionality renders these tests inapplicable. There is a set of references on testing the coefficients of high dimensional linear regression, which includes the tests of Goeman *et al.* (2006) for an empirical Bayes formulation, and Zhong and Chen (2011) that accommodates factorial designs. See also Lan *et al.* (2014). There are references on the inference of the regression coefficients associated with the lasso and other variable selection methods for linear models, for instance van de Geer *et al.* (2013), Lee *et al.* (2014), Taylor *et al.* (2014), Voorman *et al.* (2014) and Zhang and Zhang (2014).

In this paper, we consider testing high dimensional regression coefficients of generalized linear models without the sparsity assumption. In an important work, Goeman *et al.* (2011) proposed a test for high dimensional generalized linear models in the presence of nuisance parameters. The test provided a critical tool for performing multivariate tests when conventional likelihood ratio and Wald tests are not applicable. By allowing the dimension p to be larger than the sample size n, the test of Goeman *et al.* (2011) was formulated effectively for fixed p.

We propose tests for both the whole and partial regression coefficients for high dimensional generalized linear models with diverging p. We modify the test statistic of Goeman *et al.* (2011), which is designed to develop tests with accurate size and satisfactory power when p diverges along with the sample size. Our asymptotic analysis shows that this modification is critical for models with unbounded link functions, such as the log-link in Poisson or negative binomial regression. The tests proposed are studied by both asymptotic analysis and numerical simulations. The tests are applied to find significant gene sets in an empirical study on lung cancer. It is shown that the p-values from the tests, when used in conjunction with a proper control of the false discovery rate (FDR) (Storey *et al.*, 2004), can be used to select significant gene sets.

The paper is organized as follows. In Section 2, we review the inferential setting for generalized linear models. Section 3 studies the asymptotic properties of the test of Goeman *et al.* (2011) when p is diverging. Our proposed global test is outlined in Section 4, and the test in the presence of nuisance parameters in Section 5. Results from simulation studies are presented in Section 6. Section 7 details the case-study on lung cancer. All technical details and proofs are relegated to on-line supplementary material.

The data for the simulation studies and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Models and existing test

Let Y be a response variable to a p-dimensional covariate X. Generalized linear models (McCullagh and Nelder, 1989) provide a rich collection of specifications for the conditional mean of Y given X. Although they are connected to the exponential family of distributions, a more general view can be entertained via the semiparametric quasi-likelihood of Wedderburn (1974).

Conditioning on X, there is a monotone function $g(\cdot)$ and a non-negative function $V(\cdot)$ such that $E(Y|X) = \mu(\beta) = g(X^{\mathrm{T}}\beta)$ and $\mathrm{var}(Y|X) = V\{\mu(\beta); \phi\}$, where β is a p-dimensional vector of regression coefficients, $g^{-1}(\cdot)$ is the link function and ϕ is a dispersion parameter.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of (X, Y). The maximum quasi-likelihood estimator $\hat{\beta}_n$ of β can be obtained by solving the quasi-likelihood score equation

$$l_n(\beta) = \sum_{i=1}^{n} \frac{\{Y_i - g(X_i^{\mathrm{T}}\beta)\}\, g'(X_i^{\mathrm{T}}\beta) X_i}{V\{\mu_i(\beta); \hat{\phi}\}} = 0, \tag{2.1}$$

where $\mu_i(\beta) = g(X_i^{\mathrm{T}}\beta)$ and $\hat{\phi}$ is an estimator of ϕ, which can be obtained via the method of moments, such as given in Chen and Cui (2003). The consistency and asymptotic normality of $\hat{\beta}_n$ are established for fixed dimensional covariates (McCullagh and Nelder, 1989).

Let $\beta = (\beta^{(1)\mathrm{T}}, \beta^{(2)\mathrm{T}})^{\mathrm{T}}$ be a partition of the coefficient vector and $X_i = (X_i^{(1)\mathrm{T}}, X_i^{(2)\mathrm{T}})^{\mathrm{T}}$ be the corresponding partition of the covariates, where $\beta^{(1)}$ and $X_i^{(1)}$ are p_1 dimensional, $\beta^{(2)}$ and $X_i^{(2)}$ are p_2 dimensional and $p_1 + p_2 = p$. Suppose that we are interested in testing

$$H_0 : \beta^{(2)} = \beta_0^{(2)} \quad \textit{versus} \quad H_1 : \beta^{(2)} \neq \beta_0^{(2)}$$

on the effect of the covariate $X_i^{(2)}$, while treating $\beta^{(1)}$ as the nuisance parameter.

When the dimensions p_1 and p_2 are fixed, the modified Wald and the score tests based on the asymptotic χ^2-approximations (Fahrmeir and Tutz, 1994) can be performed to test the above hypothesis. However, the high dimensionality often ensures that $p_2 > n$ (Pan, 2009). When $p_2 > n$, the modified Wald and the likelihood ratio tests are inapplicable since the invertibility of the Fisher information matrix is not attainable and the maximum likelihood estimators for the parameters cannot be obtained.

Goeman *et al.* (2011) considered the following test formulation in the case of $p_2 > n$ with $g^{-1}(\cdot)$ being the canonical link. To ensure its general application, non-canonical links are considered by defining $\psi(X_i, \beta_0, \phi) = g'(X_i^{\mathrm{T}}\beta_0)/V\{\mu_i(\beta_0); \phi\}$ where g' denotes the derivative of g. For canonical links, $\psi(X_i, \beta_0, \phi)$ becomes constant.

Let $\hat{\beta}_0^{(1)}$ and $\hat{\phi}_0$ be the estimators of the nuisance parameters $\beta^{(1)}$ and ϕ under H_0, $\hat{\beta}_0 = (\hat{\beta}_0^{(1)\mathrm{T}}, \beta_0^{(2)\mathrm{T}})^{\mathrm{T}}$, $\hat{\mu}_{0i} = \mu_i(\hat{\beta}_0)$, $\hat{\mu}_0 = (\hat{\mu}_{01}, \ldots, \hat{\mu}_{0n})^{\mathrm{T}}$ and $\hat{\Psi}_0 = (\psi(X_1, \hat{\beta}_0, \hat{\phi}_0), \ldots, \psi(X_n, \hat{\beta}_0, \hat{\phi}_0))^{\mathrm{T}}$. Moreover, let $\mathbb{X}^{(2)} = (X_1^{(2)}, \ldots, X_n^{(2)})^{\mathrm{T}}$, $\mathbb{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and $\mathbb{D}$ be the $n \times n$ diagonal matrix that collects the diagonal elements of $\mathbb{X}^{(2)}\mathbb{X}^{(2)\mathrm{T}}$. The test statistic of Goeman *et al.* (2011) is

$$S_n = \frac{((\mathbb{Y} - \hat{\mu}_0) \circ \hat{\Psi}_0)^{\mathrm{T}} \mathbb{X}^{(2)} \mathbb{X}^{(2)\mathrm{T}} ((\mathbb{Y} - \hat{\mu}_0) \circ \hat{\Psi}_0)}{((\mathbb{Y} - \hat{\mu}_0) \circ \hat{\Psi}_0)^{\mathrm{T}} \mathbb{D} ((\mathbb{Y} - \hat{\mu}_0) \circ \hat{\Psi}_0)} \tag{2.2}$$

where $A \circ B = (a_{ij} b_{ij})$ for matrices $A = (a_{ij})$ and $B = (b_{ij})$. Under the null hypothesis H_0, the score function of $\beta^{(2)}$ is $l_2(\hat{\beta}_0^{(1)}, \beta_0^{(2)}) = \mathbb{X}^{(2)\mathrm{T}}((\mathbb{Y} - \hat{\mu}_0) \circ \hat{\Psi}_0)$. Hence, the numerator of S_n is a quadratic form of the score function, whereas the denominator is a plug-in estimator of the mean of the numerator for standardization.

3. Test of Goeman *et al.* (2011) when $p \to \infty$

The proposal of Goeman *et al.* (2011) was formulated for fixed dimension p while allowing $p > n$. In this section, we analyse its properties for diverging p as $n \to \infty$. It will be shown that the test of Goeman *et al.* (2011) remains powerful for diverging p when either g or g' is bounded, for instance, logistic or linear regression. However, it will be shown that there is a loss of power for the test for unbounded link functions such as Poisson or negative binomial regression with log-link. The analysis will provide useful insights for alternative formulations when p is diverging.

To make the discussion focused, we concentrate on testing the global hypothesis without the nuisance regression parameters, namely

$$H_0 : \beta = \beta_0 \quad \textit{versus} \quad H_1 : \beta \neq \beta_0.$$

Without loss of generality, we assume that $E(X)=0$ as otherwise X can be recentred by its mean. We denote $\Sigma_X=\mathrm{cov}(X)$, $\epsilon=Y-g(X^{\mathrm{T}}\beta)$ and $\epsilon_0=Y-g(X^{\mathrm{T}}\beta_0)$, we use $\|\cdot\|$ for the Euclidean norm and, for two sequences $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ means $a_n=O(b_n)$ and $b_n=O(a_n)$.

Our analysis makes the following assumptions.

Assumption 1. There is an m-variate random vector $Z_i=(z_{i1},\ldots,z_{im})^{\mathrm{T}}$ for some $m \geqslant p$ such that $X_i=\Gamma Z_i$, where Γ is a $p\times m$ constant matrix with $\Gamma\Gamma^{\mathrm{T}}=\Sigma_X$, $E(Z_i)=0$ and $\mathrm{var}(Z_i)=\mathbb{I}_m$, where $\mathbb{I}_m$ is the $m\times m$ identity matrix. Each z_{ij} has finite eighth moment and $E(z_{ij}^4)=3+\Delta$ for a constant $\Delta>-3$. For any integers $l_\nu \geqslant 0$ and distinct $j_1,\ldots,j_q$ with $\sum_{\nu=1}^q l_\nu \leqslant 8$,

$$E(z_{ij_1}^{l_1} z_{ij_2}^{l_2}\ldots z_{ij_q}^{l_q})=E(z_{ij_1}^{l_1})\,E(z_{ij_2}^{l_2})\ldots E(z_{ij_q}^{l_q}).$$

Assumption 2. As $n\to\infty$, $p\to\infty$, $\mathrm{tr}(\Sigma_X^2)\to\infty$ and $\mathrm{tr}(\Sigma_X^4)=o\{\mathrm{tr}^2(\Sigma_X^2)\}$.

Assumption 3. Let f_x be the probability density of X and $D(f_x)$ be its support. There are positive constants K_1 and K_2 such that $E(\epsilon^2|X=x)>K_1$ and $E(\epsilon^8|X=x)<K_2$ almost everywhere for $x\in D(f_x)$.

Assumption 4. There are positive constants c_1 and c_2 such that $c_1 \leqslant \psi^2(x,\beta_0,\phi) \leqslant c_2$ almost everywhere for $x\in D(f_x)$, and g is continuous differentiable, $V(\cdot)>0$.

Assumption 1 has been adopted by other prominent researchers, such as Bai and Saranadasa (1996) and Zhong and Chen (2011) to facilitate the analyses in ultrahigh dimensional tests for the means and linear regression. The model contains the Gaussian and other important multivariate distributions as special cases, which was also discussed in Chen *et al.* (2009). Assumption 2 is weaker than assuming explicitly the relative growth rates between p and n. It is noted that, when all the eigenvalues of Σ_X are bounded, $\mathrm{tr}(\Sigma_X^4)=o\{\mathrm{tr}^2(\Sigma_X^2)\}$ is true for any diverging p. Assumptions 3 and 4 are standard in the analysis of generalized linear models, for instance, assumption G in Fan and Song (2010). Assumption 4 is satisfied if Y is from the exponential family with canonical links.

To reduce the amount of notation, we assume that the dispersion parameter ϕ can be ignored in the inference for β. We shall reconsider ϕ in Section 5 when treating the nuisance parameters. To facilitate the analysis, we define three matrices:

$$\Delta_{\beta,\beta_0}=E[\{g(X^{\mathrm{T}}\beta)-g(X^{\mathrm{T}}\beta_0)\}\psi(X,\beta_0)X],$$
$$\Sigma_\beta(\beta_0)=E[V\{g(X^{\mathrm{T}}\beta)\}\psi^2(X,\beta_0)XX^{\mathrm{T}}]$$

and

$$\Xi_{\beta,\beta_0}=E[\{g(X^{\mathrm{T}}\beta)-g(X^{\mathrm{T}}\beta_0)\}^2\psi^2(X,\beta_0)XX^{\mathrm{T}}].$$

The test statistic S_n of Goeman *et al.* (2011) can be expressed as

$$S_n=1+U_n/A_n$$

where

$$U_n=\frac{1}{n}\sum_{i\neq j}^{n}\{(Y_i-\mu_{0i})(Y_j-\mu_{0j})\,\psi(X_i,\beta_0)\,\psi(X_j,\beta_0)X_i^{\mathrm{T}}X_j\}$$

and

$$A_n=\frac{1}{n}\sum_{i=1}^{n}\{(Y_i-\mu_{0i})^2\psi^2(X_i,\beta_0)X_i^{\mathrm{T}}X_i\}.$$

Insights on S_n can be gained via the means of U_n and A_n. Lemma A.1 in the on-line supplementary material shows that the means are respectively

$$\begin{aligned} \mu_{U_n} &= (n-1)\Delta_{\beta,\beta_0}^{\mathrm{T}}\Delta_{\beta,\beta_0}, \\ \mu_{A_n} &= \mathrm{tr}\{\Sigma_\beta(\beta_0) + \Xi_{\beta,\beta_0}\}. \end{aligned} \tag{3.1}$$

We note that for generalized linear models, as far as testing is concerned, the identification between β and β_0 is made via the closeness of $g(X^{\mathrm{T}}\beta)$ and $g(X^{\mathrm{T}}\beta_0)$, namely by Δ_{β,β_0}.

Lemma A.1 also gives the variances of A_n and U_n, which are respectively

$$\begin{aligned} \sigma_{A_n}^2 &= n^{-1}[E\{\epsilon_0^4\psi^4(X,\beta_0)(X^{\mathrm{T}}X)^2\} - E^2\{\epsilon_0^2\psi^2(X,\beta_0)(X^{\mathrm{T}}X)\}], \\ \sigma_{U_n}^2 &= 4(n-2)(1-n^{-1})\xi_1 + 2(1-n^{-1})\xi_2, \end{aligned} \tag{3.2}$$

where $\xi_1 = \Delta_{\beta,\beta_0}^{\mathrm{T}}\{\Sigma_\beta(\beta_0) + \Xi_{\beta,\beta_0}\}\Delta_{\beta,\beta_0} - (\Delta_{\beta,\beta_0}^{\mathrm{T}}\Delta_{\beta,\beta_0})^2$ and $\xi_2 = \mathrm{tr}\{\Sigma_\beta(\beta_0) + \Xi_{\beta,\beta_0}\}^2 - (\Delta_{\beta,\beta_0}^{\mathrm{T}}\Delta_{\beta,\beta_0})^2$.

By Taylor series expansion,

$$S_n = 1 + \mu_{A_n}^{-1}\mu_{U_n} - \mu_{A_n}^{-2}\mu_{U_n}(A_n - \mu_{A_n}) + \mu_{A_n}^{-1}(U_n - \mu_{U_n}) + \mu_{A_n}^{-3}\mu_{U_n}(A_n - \mu_{A_n})^2 + \ldots. \tag{3.3}$$

The high dimensionality makes a general identification of the leading order term of expansion (3.3) quite challenging. We consider two families of alternative hypotheses, which allow identification of leading order term. One is the so-called 'local' alternatives:

$$\begin{aligned} \mathscr{L}_\beta = \{&\beta_0 \in R^p | \Delta_{\beta,\beta_0}^{\mathrm{T}}\Sigma_X\Delta_{\beta,\beta_0} = o\{n^{-1}\,\mathrm{tr}(\Sigma_X^2)\} \text{ and either } \{g(X^{\mathrm{T}}\beta) - g(X^{\mathrm{T}}\beta_0)\}^2 = O(1) \\ &\text{almost surely or } (\beta-\beta_0)^{\mathrm{T}}\Sigma_X(\beta-\beta_0) = O(1) \text{ and } |g'(t)| \leqslant C_0 \text{ for any } t \in (-\infty,\infty)\} \end{aligned}$$

for a positive constant C_0. The other is the so-called 'fixed' alternatives:

$$\mathscr{L}_\beta^{\mathrm{F}} = \{\beta_0 \in R^p | \Delta_{\beta,\beta_0}^{\mathrm{T}}(\Xi_{\beta,\beta_0} + \Sigma_X)\Delta_{\beta,\beta_0} = o\{n^{-1}\,\mathrm{tr}(\Xi_{\beta,\beta_0}^2)\} \text{ and } \mathrm{tr}(\Sigma_X^2) = o\{\mathrm{tr}(\Xi_{\beta,\beta_0}^2)\}\}.$$

We use the term local because H_0 is part of $\mathscr{L}_\beta$. To gain insight on $\mathscr{L}_\beta$, we note that, for the linear model, the first condition in $\mathscr{L}_\beta$ becomes

$$\Delta_{\beta,\beta_0}^{\mathrm{T}}\Sigma_X\Delta_{\beta,\beta_0} = (\beta-\beta_0)^{\mathrm{T}}\Sigma_X^3(\beta-\beta_0) = o\{n^{-1}\mathrm{tr}(\Sigma_X^2)\};$$

and the second condition becomes, given that $g'(t) = 1$,

$$(\beta-\beta_0)^{\mathrm{T}}\Sigma_X(\beta-\beta_0) = O(1). \tag{3.4}$$

These two restrictions prescribe that $\|\beta-\beta_0\|$ is relatively small. For the logistic and probit models, as $g(t)$ are uniformly bounded, the first option of the second condition in $\mathscr{L}_\beta$ is satisfied. The first requirement of $\mathscr{L}_\beta$ implies that $\|\Delta_{\beta,\beta_0}\|$ is relatively small. Note that, for non-linear g, the 'localness' of β_0 to β is written via Δ_{β,β_0} rather than $\|\beta-\beta_0\|$ as in the linear case.

The fixed alternatives $\mathscr{L}_\beta^{\mathrm{F}}$ prescribe larger values for Δ_{β,β_0}. To appreciate this, we note that for the linear model

$$\Xi_{\beta,\beta_0} = E\{(X^{\mathrm{T}}\beta - X^{\mathrm{T}}\beta_0)^2 XX^{\mathrm{T}}\}.$$

Suppose that the largest and smallest eigenvalues of Ξ_{β,β_0} are of the same order, and let $\lambda_{\max}(\Sigma_X)$ be the largest eigenvalue of Σ_X. Then, sufficient conditions for the two requirements in $\mathscr{L}_\beta^{\mathrm{F}}$ are

$$\left.\begin{array}{c} n\,\lambda_{\max}(\Sigma_X) = o\{(\beta-\beta_0)^{\mathrm{T}}\Sigma_X^2(\beta-\beta_0)\}, \\ \mathrm{tr}^{1/2}(\Sigma_X^2) = o\{(\beta-\beta_0)^{\mathrm{T}}\Sigma_X^2(\beta-\beta_0)\}, \\ p/n \to \infty. \end{array}\right\} \tag{3.5}$$

For non-linear link functions, the sufficient conditions to ensure $\mathscr{L}_\beta^{\mathrm{F}}$ are

$$\left.\begin{array}{c} n\,\lambda_{\max}(\Sigma_X) = o(\|\Delta_{\beta,\beta_0}\|^2), \\ \mathrm{tr}^{1/2}(\Sigma_X^2) = o(\|\Delta_{\beta,\beta_0}\|^2), \\ p/n \to \infty \end{array}\right\} \tag{3.6}$$

provided that the largest and smallest eigenvalues of Ξ_{β,β_0} are of the same order. It is noted that conditions (3.6) are applicable to the Poisson or negative binomial model which has unbounded link function. An unbounded link function makes $\|\Delta_{\beta,\beta_0}\|^2$ more likely to be large.

The two families of the alternative hypotheses have different implications for the asymptotic behaviour of S_n. If $\beta_0 \in \mathscr{L}_\beta$, according to theorem 1,

$$S_n = 1 + \mu_{A_n}^{-1}\mu_{U_n} + \mu_{A_n}^{-1}(U_n - \mu_{U_n}) + o_p(\mu_{A_n}^{-1}\sigma_{U_n}), \tag{3.7}$$

namely S_n is primarily composed of the linear term $\mu_{A_n}^{-1}(U_n - \mu_{U_n})$. This is the same as the fixed dimensional case, which provides the foundation for the proposal of Goeman *et al.* (2011). However, if $\beta_0 \in \mathscr{L}_\beta^{\mathrm{F}}$, the analysis leading to theorem 2 shows that the other linear term $\mu_{A_n}^{-2}\mu_{U_n}(A_n - \mu_{A_n})$ can join $\mu_{A_n}^{-1}(U_n - \mu_{U_n})$ as the leading order terms of S_n. When this happens, the power of the test will be reduced as shown in theorem 2.

Let $\sigma_{S_n}^2 = 2\,\mathrm{tr}\{\Sigma_\beta(\beta_0) + \Xi_{\beta,\beta_0}\}^2\,\mathrm{tr}^{-2}\{\Sigma_\beta(\beta_0) + \Xi_{\beta,\beta_0}\}$. The following theorem establishes the asymptotic normality of S_n under the local alternatives.

Theorem 1. Suppose that assumptions 1–4 hold; then, for $\beta_0 \in \mathscr{L}_\beta$,

$$\sigma_{S_n}^{-1}(S_n - 1 - \mu_{A_n}^{-1}\mu_{U_n}) \xrightarrow{\mathrm{d}} N(0,1) \qquad \text{as } n \to \infty \text{ and } p \to \infty.$$

Under the null hypothesis H_0, $\mu_{A_n}^{-1}\mu_{U_n} = 0$ and $\sigma_{S_n}^2 = 2\,\mathrm{tr}\{\Sigma_{\beta_0}^2(\beta_0)\}\,\mathrm{tr}^{-2}\{\Sigma_{\beta_0}(\beta_0)\}$. To formulate a test procedure based on asymptotic normality, we need to estimate σ_{S_n} and hence $\mathrm{tr}\{\Sigma_{\beta_0}^2(\beta_0)\}$ and $\mathrm{tr}^2\{\Sigma_{\beta_0}(\beta_0)\}$. Let

$$\mathrm{tr}\{\widehat{\Sigma_{\beta_0}^2(\beta_0)}\} = \frac{1}{n(n-1)}\sum_{i\neq j}^{n}[\{Y_i - g(X_i^{\mathrm{T}}\beta_0)\}^2\{Y_j - g(X_j^{\mathrm{T}}\beta_0)\}^2\,\psi^2(X_i,\beta_0)\,\psi^2(X_j,\beta_0)(X_i^{\mathrm{T}}X_j)^2]$$

and

$$\begin{aligned}\mathrm{tr}^2\{\widehat{\Sigma_{\beta_0}(\beta_0)}\} = \frac{1}{n(n-1)}\sum_{i\neq j}^{n}[&\{Y_i - g(X_i^{\mathrm{T}}\beta_0)\}^2\{Y_j - g(X_j^{\mathrm{T}}\beta_0)\}^2\,\psi^2(X_i,\beta_0) \\ &\times \psi^2(X_j,\beta_0)(X_i^{\mathrm{T}}X_i)(X_j^{\mathrm{T}}X_j)].\end{aligned}$$

Under the null hypothesis, lemma A.3 in the on-line supplementary material shows that

$$\frac{\mathrm{tr}\{\widehat{\Sigma_{\beta_0}^2(\beta_0)}\}}{\mathrm{tr}\{\Sigma_{\beta_0}^2(\beta_0)\}} \xrightarrow{\mathrm{p}} 1 \quad \text{and} \quad \frac{\mathrm{tr}^2\{\widehat{\Sigma_{\beta_0}(\beta_0)}\}}{\mathrm{tr}^2\{\Sigma_{\beta_0}(\beta_0)\}} \xrightarrow{\mathrm{p}} 1 \qquad \text{as } n \to \infty \text{ and } p \to \infty.$$

Then, theorem 1 and Slutsky's lemma lead to an asymptotic α-level test that rejects H_0 if

$$S_n > 1 + z_\alpha[2\,\mathrm{tr}\{\widehat{\Sigma_{\beta_0}^2(\beta_0)}\}/\mathrm{tr}^2\{\widehat{\Sigma_{\beta_0}(\beta_0)}\}]^{1/2}, \tag{3.8}$$

where z_α is the upper α-quantile of $N(0,1)$.

Goeman *et al.* (2011) approximated the null distribution of S_n by a ratio of quadratic forms based on normally distributed variables, which involves numerical inversion of the characteristic function. The critical value that is obtained via the procedure of Goeman *et al.* (2011) is

asymptotically equivalent to the right-hand side of inequality (3.8) under hypothesis H_0 in the case of $p \to \infty$, which is confirmed by our simulation study. We shall use inequality (3.8) in the following power analysis.

Define the power of the test in inequality (3.8) under $\mathscr{L}_\beta$ as

$$\Omega(\beta, \beta_0) = P(S_n > 1 + z_\alpha [2\,\mathrm{tr}\{\widehat{\Sigma^2_{\beta_0}(\beta_0)}\}/\mathrm{tr}^2\{\widehat{\Sigma_{\beta_0}(\beta_0)}\}]^{1/2} | \beta_0 \in \mathscr{L}_\beta).$$

Corollary 1. Under assumptions 1–4 and for $\beta_0 \in \mathscr{L}_\beta$,

$$\Omega(\beta, \beta_0) = \Phi\left(-z_\alpha + \frac{n\|\Delta_{\beta,\beta_0}\|^2}{[2\,\mathrm{tr}\{\Sigma_\beta(\beta_0) + \Xi_{\beta,\beta_0}\}^2]^{1/2}} \right)\{1 + o(1)\} \qquad \text{as } n \to \infty \text{ and } p \to \infty.$$

Corollary 1 shows that the power of the test in inequality (3.8) is determined by

$$\mathrm{SNR}(\beta, \beta_0) = \frac{n\|\Delta_{\beta,\beta_0}\|^2}{[2\,\mathrm{tr}\{\Sigma_\beta(\beta_0) + \Xi_{\beta,\beta_0}\}^2]^{1/2}}.$$

We note that $\|\Delta_{\beta,\beta_0}\|^2 = \|E[\{g(X^\mathrm{T}\beta) - g(X^\mathrm{T}\beta_0)\}\psi(X, \beta_0)X]\|^2$ measures the difference between H_0 and H_1, and can be viewed as the signal of the test problem. At the same time, $[2\,\mathrm{tr}\{\Sigma_\beta(\beta_0) + \Xi_{\beta,\beta_0}\}^2]^{1/2}$ can be regarded as the noise due to its connection to the standard deviation of S_n. Hence, $\mathrm{SNR}(\beta, \beta_0)$ is the signal-to-noise ratio of the test.

Let us evaluate the power of the test under the fixed alternatives $\mathscr{L}^\mathrm{F}_\beta$, which is denoted

$$\Omega^\mathrm{F}(\beta, \beta_0) = P(S_n > 1 + z_\alpha [2\,\mathrm{tr}\{\widehat{\Sigma^2_{\beta_0}(\beta_0)}\}/\mathrm{tr}^2\{\widehat{\Sigma_{\beta_0}(\beta_0)}\}]^{1/2} | \beta_0 \in \mathscr{L}^\mathrm{F}_\beta).$$

As stated earlier, under the fixed alternatives, the leading order terms of S_n may involve $\mu^{-2}_{A_n}\mu_{U_n}(A_n - \mu_{A_n})$ and other terms in expansion (3.3). These terms are smaller order terms in the fixed dimensional case or the local alternatives considered above. However, they may no longer be smaller order terms under the fixed alternatives when p diverges. To appreciate this, rewrite expansion (3.3) as

$$S_n = 1 + \mu^{-1}_{A_n}\mu_{U_n} - \alpha_{n1}(A_n - \mu_{A_n})/\sigma_{A_n} + \alpha_{n2}(U_n - \mu_{U_n})/\sigma_{U_n} + \alpha_{n3}(A_n - \mu_{A_n})^2/\sigma^2_{A_n} + \ldots \quad (3.9)$$

where $\alpha_{n1} = \mu^{-2}_{A_n}\mu_{U_n}\sigma_{A_n}$, $\alpha_{n2} = \mu^{-1}_{A_n}\sigma_{U_n}$ and $\alpha_{n3} = \mu^{-3}_{A_n}\mu_{U_n}\sigma^2_{A_n}$. To control the quadratic term and beyond, we need to impose

$$\frac{\alpha^2_{n3}}{\alpha^2_{n1}} = \frac{\sigma^2_{A_n}}{\mu^2_{A_n}} = \frac{1}{n}\left[\frac{E\{\epsilon_0^4\psi_0^4(X^\mathrm{T}X)^2\}}{E^2\{\epsilon_0^2\psi_0^2(X^\mathrm{T}X)\}} - 1\right] = o(1), \quad (3.10)$$

where $\psi_0 = g'(X^\mathrm{T}\beta_0)/V\{\mu(\beta_0)\}$. Define

$$\kappa_n^2 = \alpha^2_{n1}/\alpha^2_{n2} = \mu^2_{U_n}\sigma^2_{A_n}/(\mu^2_{A_n}\sigma^2_{U_n}). \quad (3.11)$$

Under condition (3.10), and if κ_n^2 is bounded, the asymptotic properties of S_n will depend on κ_n^2 as shown in the following theorem, which requires the following assumption.

Assumption 5.

(a) As $n \to \infty$, $p \to \infty$, $\mathrm{tr}(\Xi^2_{\beta,\beta_0}) \to \infty$ and $\mathrm{tr}(\Xi^4_{\beta,\beta_0}) = o\{\mathrm{tr}^2(\Xi^2_{\beta,\beta_0})\}$;
(b) $E(\epsilon_0^4\psi_0^4[X^\mathrm{T}\{\Sigma_\beta(\beta_0) + \Xi_{\beta,\beta_0}\}X]^2) = o(n\,E^2[\epsilon_0^2\psi_0^2X^\mathrm{T}\{\Sigma_\beta(\beta_0) + \Xi_{\beta,\beta_0}\}X])$;
(c) $E\{(\epsilon_{01}\epsilon_{02}\psi_{01}\psi_{02}X_1^\mathrm{T}X_2)^4\} = o[n^2\,E^2\{(\epsilon_{01}\epsilon_{02}\psi_{01}\psi_{02}X_1^\mathrm{T}X_2)^2\}]$.

Assumption 5 is required to control properly the quadratic term and beyond in expansion (3.3) so that an asymptotic distribution can be established for S_n. Assumption 5, part (a), is

analogous to assumption 2 in the local case, which is trivially true if all the eigenvalues of Ξ_{β,β_0} are of the same order. Assumption 5, parts (b) and (c), are needed to control the higher order moments of U_n in the derivation of its asymptotic distribution by using the martingale central limit theorem.

Theorem 2. Under assumptions 1–5 and condition (3.10) and for $\beta_0 \in \mathscr{L}_\beta^{\mathrm{F}}$, and if $\kappa_n^2 = O(1)$,

$$\Omega^{\mathrm{F}}(\beta,\beta_0) = \Phi\left(\frac{1}{(1+\kappa_n^2)^{1/2}} \left[-z_\alpha + \frac{n\|\Delta_{\beta,\beta_0}\|^2}{\{2\,\mathrm{tr}(\Xi_{\beta,\beta_0}^2)\}^{1/2}} \right] \right) \{1+o(1)\} \qquad \text{as } n\to\infty \text{ and } p\to\infty.$$

The reason for obtaining the power expression $\Omega^{\mathrm{F}}(\beta,\beta_0)$ is that

$$S_n = 1 + \mu_{A_n}^{-1}\mu_{U_n} - \mu_{A_n}^{-2}\mu_{U_n}(A_n-\mu_{A_n}) + \mu_{A_n}^{-1}(U_n-\mu_{U_n}) + o_p(\mu_{A_n}^{-1}\sigma_{U_n}) \tag{3.12}$$

and the fact that κ_n^2 defines the ratio of the variances of the two linear terms. Theorem 2 implies that the power of the test based on S_n is reduced from that under the local alternatives because of the effect of $\mu_{A_n}^{-2}\mu_{U_n}(A_n-\mu_{A_n})$. The degree of the power reduction depends on κ_n^2. For larger κ_n^2, namely when the variance of $\mu_{A_n}^{-2}\mu_{U_n}(A_n-\mu_{A_n})$ is relatively larger than that of the linear term involving U_n, there will be more power reduction, and vice versa.

4. A new proposal

An important insight acquired in the previous section is that the A_n-term in the statistic

$$S_n = 1 + U_n/A_n$$

does not contribute to the signal of the test but rather can increase the variance and hence can adversely affect the power as revealed in theorem 2.

The role of A_n is to standardize so that the final test statistic is invariant under the scale transformation of Y. Indeed, in conventional inferential situations, the standardization is commonly used to produce invariance, which is known to be beneficial in fixed dimensional situations. However, for high dimensional data, the standardization developed under the fixed dimensionality may create some problems. One such case is the F-test for linear regression as shown in Zhong and Chen (2011). The statistic S_n under current study is another. We are not suggesting that there is no need for standardization. We actually standardize the proposed test statistic U_n by its standard deviation so that the final test statistic is asymptotically pivotal.

Our approach is to remove A_n from S_n and then to normalize at the end by taking into account the high dimensionality. Specifically, we use

$$U_n = \frac{1}{n}\sum_{i\neq j}^{n} \{(Y_i-\mu_{0i})(Y_j-\mu_{0j})\,\psi(X_i,\beta_0)\,\psi(X_j,\beta_0)X_i^{\mathrm{T}}X_j\}$$

as the test statistic. Compared with the expansion (3.3) of S_n involved, it is much simpler. Despite being simpler, U_n captures the signal of the test since $E(U_n) = (n-1)\|\Delta_{\beta,\beta_0}\|^2$, as shown in expression (3.1). We shall demonstrate that a test based on U_n attains better power than the test of Goeman *et al.* (2011) for diverging p under $\mathscr{L}_\beta^{\mathrm{F}}$, while maintaining the same asymptotic power under $\mathscr{L}_\beta$.

We consider testing the global hypothesis $H_0 : \beta = \beta_0$ in this section. A test in the presence of the nuisance parameters will be proposed in the next section. Let $\sigma_{U_n}^2 = 2\,\mathrm{tr}\{\Sigma_\beta(\beta_0)+\Xi_{\beta,\beta_0}\}^2\{1+o(1)\}$, which is the variance of U_n under $\mathscr{L}_\beta$, according to lemma A.2 in the on-line supplementary material.

Theorem 3. Under assumptions 1–4 and for $\beta_0 \in \mathscr{L}_\beta$,

$$\frac{U_n - n\|\Delta_{\beta,\beta_0}\|^2}{[2\,\mathrm{tr}\{\Sigma_\beta(\beta_0) + \Xi_{\beta,\beta_0}\}^2]^{1/2}} \xrightarrow{\mathrm{d}} N(0,1) \qquad \text{as } n \to \infty \text{ and } p \to \infty.$$

Theorem 3 implies that, under the null hypothesis H_0,

$$\frac{U_n}{[2\,\mathrm{tr}\{\Sigma^2_{\beta_0}(\beta_0)\}]^{1/2}} \xrightarrow{\mathrm{d}} N(0,1) \qquad \text{as } n \to \infty \text{ and } p \to \infty.$$

Using $\mathrm{tr}\{\widehat{\Sigma^2_{\beta_0}(\beta_0)}\}$ given in Section 3 to estimate $\mathrm{tr}\{\Sigma^2_{\beta_0}(\beta_0)\}$, the α-level test proposed rejects H_0 if

$$U_n > z_\alpha [2\,\mathrm{tr}\{\widehat{\Sigma^2_{\beta_0}(\beta_0)}\}]^{1/2}. \tag{4.1}$$

Let $\tilde{\Omega}(\beta,\beta_0) = P(U_n > z_\alpha[2\,\mathrm{tr}\{\widehat{\Sigma^2_{\beta_0}(\beta_0)}\}]^{1/2} \mid \beta_0 \in \mathscr{L}_\beta)$ be the power of the above test under the local alternatives $\mathscr{L}_\beta$.

Corollary 2. Under assumptions 1–4 and for $\beta_0 \in \mathscr{L}_\beta$,

$$\tilde{\Omega}(\beta,\beta_0) = \Phi\left(-z_\alpha + \frac{n\|\Delta_{\beta,\beta_0}\|^2}{[2\,\mathrm{tr}\{\Sigma_\beta(\beta_0) + \Xi_{\beta,\beta_0}\}^2]^{1/2}} \right)\{1 + o(1)\} \qquad \text{as } n \to \infty \text{ and } p \to \infty.$$

We note here that the power of the test proposed is asymptotically equivalent to $\Omega(\beta,\beta_0)$ of Goeman *et al.* (2011) given in corollary 1. This is expected since, in the case of $\mathscr{L}_\beta$,

$$1 + \mu_{A_n}^{-1}\mu_{U_n} + \mu_{A_n}^{-1}(U_n - \mu_{U_n}) \tag{4.2}$$

is the leading order term of S_n. Hence, the two tests are asymptotically equivalent.

From the proof of theorem 4 in the on-line supplementary material, the variance of U_n under the fixed alternatives $\mathscr{L}^{\mathrm{F}}_\beta$ is

$$\sigma^2_{U_n} = 2\,\mathrm{tr}(\Xi^2_{\beta,\beta_0})\{1 + o(1)\}.$$

Let $\tilde{\Omega}^{\mathrm{F}}(\beta,\beta_0) = P(U_n > z_\alpha[2\,\mathrm{tr}\{\widehat{\Sigma^2_{\beta_0}(\beta_0)}\}]^{1/2} | \beta_0 \in \mathscr{L}^{\mathrm{F}}_\beta)$ be the power under $\mathscr{L}^{\mathrm{F}}_\beta$.

Theorem 4. Under assumptions 1–5 and for $\beta_0 \in \mathscr{L}^{\mathrm{F}}_\beta$,

$$\tilde{\Omega}^{\mathrm{F}}(\beta,\beta_0) = \Phi\left[-z_\alpha + \frac{n\|\Delta_{\beta,\beta_0}\|^2}{\{2\,\mathrm{tr}(\Xi^2_{\beta,\beta_0})\}^{1/2}} \right]\{1 + o(1)\} \qquad \text{as } n \to \infty \text{ and } p \to \infty.$$

We note that the conditions of theorem 4 are much simpler than those in theorem 2, since the condition (3.10) and that on κ_n^2 are no longer needed. A gain of power of the proposed test is evident as $\tilde{\Omega}^{\mathrm{F}}(\beta,\beta_0) > \Omega^{\mathrm{F}}(\beta,\beta_0)$ asymptotically, since $\Omega^{\mathrm{F}}(\beta,\beta_0)$ involves κ_n^2.

To identify and estimate the parameters consistently as shown in Fan and Lv (2008) and Fan and Song (2010), the sparsity assumption on β is explicitly needed for estimation and variable selection of high dimensional regression parameters. However, for testing, identification is no longer a major issue, as long as $\|\Delta_{\beta,\beta_0}\|^2$ is not 0, which allows consistent test formulation. Sparsity is not explicitly needed in deriving a test as far as ensuring asymptotically correct size. However, the sparsity may affect the power of the test via its effect on the signal-to-noise ratio of the testing problem.

Goeman *et al.* (2006) investigated the locally optimal property of the score test under an empirical Bayesian framework by treating β as random such that

$$E(\beta) = \beta_0,$$
$$\mathrm{var}(\beta) = \tau^2 \Sigma,$$

where $\tau^2 \geqslant 0$ is a hyperparameter and Σ is a semipositive definite matrix. Under this framework, the original hypothesis $H_0 : \beta = \beta_0$ *versus* $H_1 : \beta \neq \beta_0$ is equivalent to

$$\tilde{H}_0 : \tau^2 = 0 \quad \textit{versus} \quad \tilde{H}_1 : \tau^2 > 0. \tag{4.3}$$

Using the notation of Goeman *et al.* (2006), let $f(\beta; \mathbb{Y})$ be the usual frequentist likelihood of the p-dimensional β given data $\mathbb{Y}$. The marginal likelihood of τ^2 is

$$\tilde{f}(\tau^2; \mathbb{Y}) = E\{f(\beta; \mathbb{Y}) | \tau^2\},$$

by taking conditional expectation on β given τ^2. Let $S = \mathrm{d} \log\{\tilde{f}(0; \mathbb{Y})\}/\mathrm{d}\tau^2$ be the score statistic; and, for a non-negative k, let

$$\tilde{w}(\beta) = P(S \geqslant k | \beta)$$

be the power function of the score test of size α, denoted as $\tilde{T}$, and $w(\beta)$ be the power function of any other test T for $H_0 : \beta = \beta_0$ such that the size of T is bounded from above by the size of the score test $\tilde{T}$, namely $w(\beta_0) \leqslant \tilde{w}(\beta_0)$ for any $k \geqslant 0$. Suppose that

$$\beta = \beta_0 + \tau \mathbf{b} \tag{4.4}$$

where $\mathbf{b}$ is uniformly distributed on the p-dimensional unit sphere. Rewrite the power functions of the tests T and $\tilde{T}$ as functions of τ under the local alternative model (4.4), i.e.

$$\begin{aligned} w_{\mathbf{b}}(\tau) &= w(\beta_0 + \tau \mathbf{b}), \\ \tilde{w}_{\mathbf{b}}(\tau) &= \tilde{w}(\beta_0 + \tau \mathbf{b}). \end{aligned} \tag{4.5}$$

Goeman *et al.* (2006) showed that, via the Neyman–Pearson argument,

$$E_{\mathbf{b}}\left\{ \frac{\mathrm{d}}{\mathrm{d}\tau^2} w_{\mathbf{b}}(0) \right\} \leqslant E_{\mathbf{b}}\left\{ \frac{\mathrm{d}}{\mathrm{d}\tau^2} \tilde{w}_{\mathbf{b}}(0) \right\}.$$

This means that the power of the score test has the largest expected gradient at $\tau = 0$ among all tests of size α; hence it is the locally most powerful test according to Cox and Hinkley (1974).

We shall show that the test proposed based on U_n is the locally most powerful test asymptotically under the framework specified by expressions (4.3)–(4.5). To achieve this, we restrict $\mathbb{Y}$ (defined in Section 2) to the given covariates X as from the exponential family such that either

$$V\{g(t), \phi\} = \phi g'(t) \tag{4.6a}$$

or

$$\mathrm{tr}^2(\Sigma_X) = o\{n \, \mathrm{tr}(\Sigma_X^2)\}. \tag{4.6b}$$

Condition (4.6a) is satisfied under the canonical link, which implies that $\psi(X, \beta)$ is a constant function. If condition (4.6a) is not satisfied, then the locally optimal test depends on $\partial \psi\{g(t)\}/\partial t|_{t = X^{\mathrm{T}}\beta}$. In this case, condition (4.6b) is needed to control the influence of $\psi(X, \beta)$.

Let $Q_n(\beta)$ be the likelihood function of β under the exponential family: the following corollary shows the local optimality of the test proposed.

Corollary 3. Under assumptions 1–4 and (4.4) and either condition (4.6a) or (4.6b), suppose that the first two derivatives of $\log\{Q_n(\beta)\}$ are bounded almost everywhere in a neighbourhood of $\beta=\beta_0$, and there is a positive constant c such that $[\partial\psi\{g(t)\}/\partial g(t)]^2|_{t=x^{\mathrm{T}}\beta_0}\leqslant c$ almost everywhere for $x\in D(f_x)$. Let

$$\tilde{\omega}_{\mathbf{b}}(\tau)=P_{y|\beta}(U_n>z_\alpha[2\,\mathrm{tr}\{\widehat{\Sigma^2_{\beta_0}(\beta_0)}\}]^{1/2})$$

be the power function of the proposed test with nominal size α and $\omega_{\mathbf{b}}(\tau)$ be the power function of any test of H_0 such that $\omega_{\mathbf{b}}(0)=\tilde{\omega}_{\mathbf{b}}(0)=\alpha$; then, as $n\to\infty$ and $p\to\infty$,

$$E_{\mathbf{b}}\left\{\frac{\mathrm{d}}{\mathrm{d}(\tau^2)}\omega_{\mathbf{b}}(0)\right\}\leqslant E_{\mathbf{b}}\left\{\frac{\mathrm{d}}{\mathrm{d}(\tau^2)}\tilde{\omega}_{\mathbf{b}}(0)\right\}.$$

Corollary 3 indicates that the power function of the test has an asymptotically optimal expected slope. Hence, it is asymptotically the locally most powerful test.

5. Test with nuisance parameter

We consider testing for parts of the regression coefficient vector. This is motivated by practical needs to consider the significance for a subset of covariates $X^{(2)}$, in the presence of other covariates $X^{(1)}$. For instance, we may have both gene expression levels and demographic variables collected in a study of a disease. The researcher may be interested only in the genetic effect, which means that the demographic coefficients together with the dispersion parameter ϕ may be treated as nuisance parameters.

Without loss of generality, we partition $\beta=(\beta^{(1)\mathrm{T}},\beta^{(2)\mathrm{T}})^{\mathrm{T}}$ and denote the nuisance parameters $\theta=(\beta^{(1)\mathrm{T}},\phi)^{\mathrm{T}}$. Suppose that the dimension of θ is p_1 and that of $\beta^{(2)}$ is p_2. It is of interest to test

$$H_{01}:\beta^{(2)}=\beta_0^{(2)}\quad\textit{versus}\quad H_{11}:\beta^{(2)}\neq\beta_0^{(2)}.$$

A test statistic along the lines of the global test statistic U_n will be proposed. The nuisance parameters $\beta^{(1)}$ and ϕ must be estimated first under H_{01}. The quasi-likelihood score of $\beta^{(1)}$ is

$$l_1(\beta^{(1)},\beta^{(2)},\phi)=\mathbb{X}^{(1)\mathrm{T}}((\mathbb{Y}-\mu)\circ\Psi),$$

where $\mathbb{X}^{(1)}$ is similarly defined as $\mathbb{X}^{(2)}$ in Section 2, $\Psi=(\psi(X_1,\beta,\phi),\ldots,\psi(X_n,\beta,\phi))^{\mathrm{T}}$ and $\mu=(\mu_1(\beta),\ldots,\mu_n(\beta))^{\mathrm{T}}$. The maximum quasi-likelihood estimator of $\beta^{(1)}$ under H_{01} solves

$$l_1(\beta^{(1)},\beta_0^{(2)},\hat{\phi}_0)=0,$$

which is denoted as $\hat{\beta}_0^{(1)}$, by plugging in $\hat{\phi}_0$, which can be either a maximum likelihood estimator or a moment estimator of ϕ as elaborated in McCullagh and Nelder (1989) or Chen and Cui (2003). Let $\hat{\beta}_0=(\hat{\beta}_0^{(1)\mathrm{T}},\beta_0^{(2)\mathrm{T}})^{\mathrm{T}}$, $\hat{\theta}_0=(\hat{\beta}_0^{(1)\mathrm{T}},\hat{\phi}_0)^{\mathrm{T}}$ and $\hat{\mu}_{0i}=\mu_i(\hat{\beta}_0)$.

We consider a statistic, which is similar to the global test statistic U_n,

$$\tilde{U}_n=\frac{1}{n}\sum_{i\neq j}^{n}\{(Y_i-\hat{\mu}_{0i})(Y_j-\hat{\mu}_{0j})\,\psi(X_i,\hat{\beta}_0,\hat{\phi}_0)\,\psi(X_j,\hat{\beta}_0,\hat{\phi}_0)X_i^{(2)\mathrm{T}}X_j^{(2)}\}. \tag{5.1}$$

Let $\Sigma_{X^{(i)}}=E(X^{(i)}X^{(i)\mathrm{T}})$ for $i=1$ and $i=2$. The following assumptions are needed.

Assumption 6. As $n\to\infty$, $p_2\to\infty$, $\mathrm{tr}(\Sigma^2_{X^{(2)}})\to\infty$ and $\mathrm{tr}(\Sigma^4_{X^{(2)}})=O\{n^{-1}\,\mathrm{tr}^2(\Sigma^2_{X^{(2)}})\}$.

Assumption 7. As $n\to\infty$, $p_1n^{-1/4}\to 0$ and there is a $\theta^*=(\beta^{*(1)\mathrm{T}},\phi^*)^{\mathrm{T}}\in R^{p_1}$ such that

$\|\hat{\theta}_0 - \theta^*\| = O_p(p_1 n^{-1/2})$ and, in particular under H_{01}, $\theta^* = \theta$, where $\theta = (\beta^{(1)\mathrm{T}}, \phi)^{\mathrm{T}}$ is the true value of the nuisance parameter.

Assumption 8. There is a positive constant λ_0 less than 1 such that $\lambda_0 \leqslant \lambda_{\min}(\Sigma_{X^{(1)}}) \leqslant \lambda_{\max}(\Sigma_{X^{(1)}}) \leqslant \lambda_0^{-1} < \infty$, where $\lambda_{\min}(\Sigma_{X^{(1)}})$ and $\lambda_{\max}(\Sigma_{X^{(1)}})$ are the smallest and largest eigenvalues of $\Sigma_{X^{(1)}}$ respectively.

Assumption 9. There are positive constants c_1 and c_2 such that, for $\beta_0^* = (\beta^{*(1)\mathrm{T}}, \beta_0^{(2)\mathrm{T}})^{\mathrm{T}}$ where $\beta^{*(1)}$ is defined in assumption 7, $c_1 \leqslant \psi^2(x, \beta_0^*, \phi^*) \leqslant c_2$ and $[\partial\psi\{g(t)\}/\partial g(t)]^2 \,|_{t=x^{\mathrm{T}}\beta_0^*} \leqslant c_2$ almost everywhere for $x \in D(f_x)$ and for t in a neighbourhood of $x^{\mathrm{T}}\beta_0^*$.

These assumptions are variations of assumptions 2–4 that were listed in Section 3. Assumption 6 is a counterpart of assumption 2 in the case of the nuisance parameter. The requirement that the growth rate of p_1 is slower than $n^{1/4}$ is to allow accurate estimation of the nuisance parameter under the high dimensionality. Assumption 7 maintains that the initial estimator $\hat{\theta}_0$ is consistent to a θ^* which may deviate from the true parameter θ, when the discrepancy between $\beta_0^{(2)}$ and $\beta^{(2)}$ is large. The θ^* minimizes the Kullback–Leibler divergence between the misspecified model under H_{01} and the model under H_{11}; see van der Vaart (2000) for discussion on inference under misspecified models. Assumption 8 is easier to be satisfied because $\Sigma_{X^{(1)}}$'s dimension is much more manageable than the case that was considered in Section 4. Assumption 9 is an updated version of assumption 4 to adjust the case for nuisance parameters.

To analyse the power, we introduce two matrices

$$\Delta^{(2)}_{\beta,\beta_0^*} = E[\{g(X^{\mathrm{T}}\beta) - g(X^{\mathrm{T}}\beta_0^*)\}\psi(X, \beta_0^*, \phi^*)X^{(2)}]$$

and

$$\Sigma^{(2)}_{\beta}(\beta_0^*) = E[V\{g(X^{\mathrm{T}}\beta)\}\psi^2(X, \beta_0^*, \phi^*)X^{(2)}X^{(2)\mathrm{T}}],$$

which are counterparts of Δ_{β,β_0} and $\Sigma_\beta(\beta_0)$ used in the study of the global test. There is no need to define a counterpart of Ξ_{β,β_0} since the local alternatives $\mathscr{L}_{\beta^{(2)}}$ defined below make it unnecessary.

The involvement of the estimated nuisance parameter $\hat{\theta}_0$ does complicate the power analysis. To expedite the study, our analysis is confined under the following family of the local alternatives:

$$\begin{aligned}\mathscr{L}_{\beta^{(2)}} = \{\beta_0^{(2)} \in R^{p_2} \mid \Delta^{(2)\mathrm{T}}_{\beta,\beta_0^*}\Sigma_{X^{(2)}}\Delta^{(2)}_{\beta,\beta_0^*} = o\{n^{-1}\operatorname{tr}(\Sigma^2_{X^{(2)}})\} \text{ and } E\{g(X^{\mathrm{T}}\beta) - g(X^{\mathrm{T}}\beta_0^*)\}^4 \\ = o(n^{-3/2})\}.\end{aligned}$$

We note here that the second component of $\mathscr{L}_{\beta^{(2)}}$ is stronger than that in $\mathscr{L}_\beta$ in Section 3, which simplifies the analysis.

Theorem 5. Under assumptions 1, 3 and 6–9, and for $\beta_0^{(2)} \in \mathscr{L}_{\beta^{(2)}}$,

$$\frac{\tilde{U}_n - n\|\Delta_{\beta,\beta_0^*}\|^2}{[2\operatorname{tr}\{\Sigma^{(2)}_\beta(\beta_0^*)\}^2]^{1/2}} \xrightarrow{\mathrm{d}} N(0,1) \qquad \text{as } n \to \infty \text{ and } p \to \infty.$$

To formulate a test procedure, we use

$$\hat{R}_n = \frac{1}{n(n-1)}\sum_{i\neq j}^{n}(Y_i - \hat{\mu}_{0i})^2(Y_j - \hat{\mu}_{0j})^2\psi^2(X_i, \hat{\beta}_0, \hat{\phi}_0)\psi^2(X_j, \hat{\beta}_0, \hat{\phi}_0)(X_i^{(2)\mathrm{T}}X_j^{(2)})^2$$

to estimate $\operatorname{tr}\{\Sigma^{(2)}_\beta(\beta_0^*)\}^2$. Under H_{01}, lemma A.7 in the on-line supplementary material shows that

$$\frac{\hat{R}_n}{\mathrm{tr}\{\Sigma_\beta^{(2)}(\beta_0^*)\}^2} \xrightarrow{\mathrm{p}} 1 \qquad \text{as } n \to \infty \text{ and } p \to \infty.$$

Hence, an asymptotic α-level test rejects H_{01} if $\tilde{U}_n > z_\alpha(2\hat{R}_n)^{1/2}$ and the proofs of theorem 5 and lemma A.7 show that the test procedure is invariant to the scale transformation of Y.

Define the power of the test under the local alternatives $\mathscr{L}_{\beta^{(2)}}$

$$\tilde{\Omega}^{(2)}(\beta, \beta_0^*) = P\{\tilde{U}_n > z_\alpha(2\hat{R}_n)^{1/2} | \beta_0^{(2)} \in \mathscr{L}_{\beta^{(2)}}\}.$$

Corollary 4. Under assumptions 1, 3 and 6–9 and the local alternatives $\mathscr{L}_{\beta^{(2)}}$,

$$\tilde{\Omega}^{(2)}(\beta, \beta_0^*) = \Phi\left(-z_\alpha + \frac{n\|\Delta^{(2)}_{\beta,\beta_0^*}\|^2}{[2\,\mathrm{tr}\{\Sigma_\beta^{(2)}(\beta_0^*)\}^2]^{1/2}}\right)\{1 + o(1)\} \qquad \text{as } n \to \infty \text{ and } p \to \infty.$$

The power $\tilde{\Omega}^{(2)}(\beta, \beta_0^*)$ has a similar form to that of $\tilde{\Omega}(\beta, \beta_0)$ in corollary 2. This is expected because of the close connection between the two test statistics. We note that the denominator inside Φ involves only $\Sigma_\beta^{(2)}(\beta_0^*)$ because of the assumption in the second part of $\mathscr{L}_{\beta^{(2)}}$.

We did not study the power under the fixed alternatives for the nuisance parameter setting, as we expect that the power performance would be largely similar to that reported in Section 4 for the test proposed. We also did not study the power property of the test of Goeman *et al.* (2011) with nuisance parameter, as the analysis would be quite involved because of the division of the A_n-term and the estimated nuisance parameter. However, we expect that the similar power properties that were revealed in the previous section would prevail. This is indeed confirmed by the simulation results that are reported in the next section.

6. Simulation studies

This section outlines results from simulation studies, which were designed to evaluate the performances of the proposed high dimensional test procedures for generalized linear models. Both the global test and the test in the presence of nuisance parameters were considered for both the tests proposed and the test of Goeman *et al.* (2011). The R package `globaltest` is used to carry out a version of the test of Goeman *et al.* (2011). We also carried out the test of Goeman *et al.* (2011) via expression (3.8) based on asymptotic normality. Both approaches produced similar empirical size and power, thus confirming that the two forms of the critical value lead to equivalent tests.

Throughout the simulation experiments, the covariates $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ were generated according to a moving average model

$$X_{ij} = \rho_1 Z_{ij} + \rho_2 Z_{i(j+1)} + \ldots + \rho_T Z_{i(j+T-1)}, \qquad j = 1, \ldots, p,$$

for some $T < p$, where $Z_i = (Z_{i1}, \ldots, Z_{i(p+T-1)})^{\mathrm{T}}$ were generated from the $(p+T-1)$-dimensional standard normal distribution $N(0, \mathbb{I}_{p+T-1})$. The coefficients $\{\rho_l\}_{l=1}^T$ were generated independently from the $U(0, 1)$ distribution and were kept fixed once generated. Here, T was used to prescribe different levels of dependence between the components of X_i. We experimented with $T = 5, 10, 20$, and we report the results for $T = 5$ since those for $T = 10$ and $T = 20$ were largely similar.

Four generalized linear models were considered in the simulation study: the logistic, linear, Poisson and negative binomial models. In the logistic regression model, the conditional mean of the response Y was given by

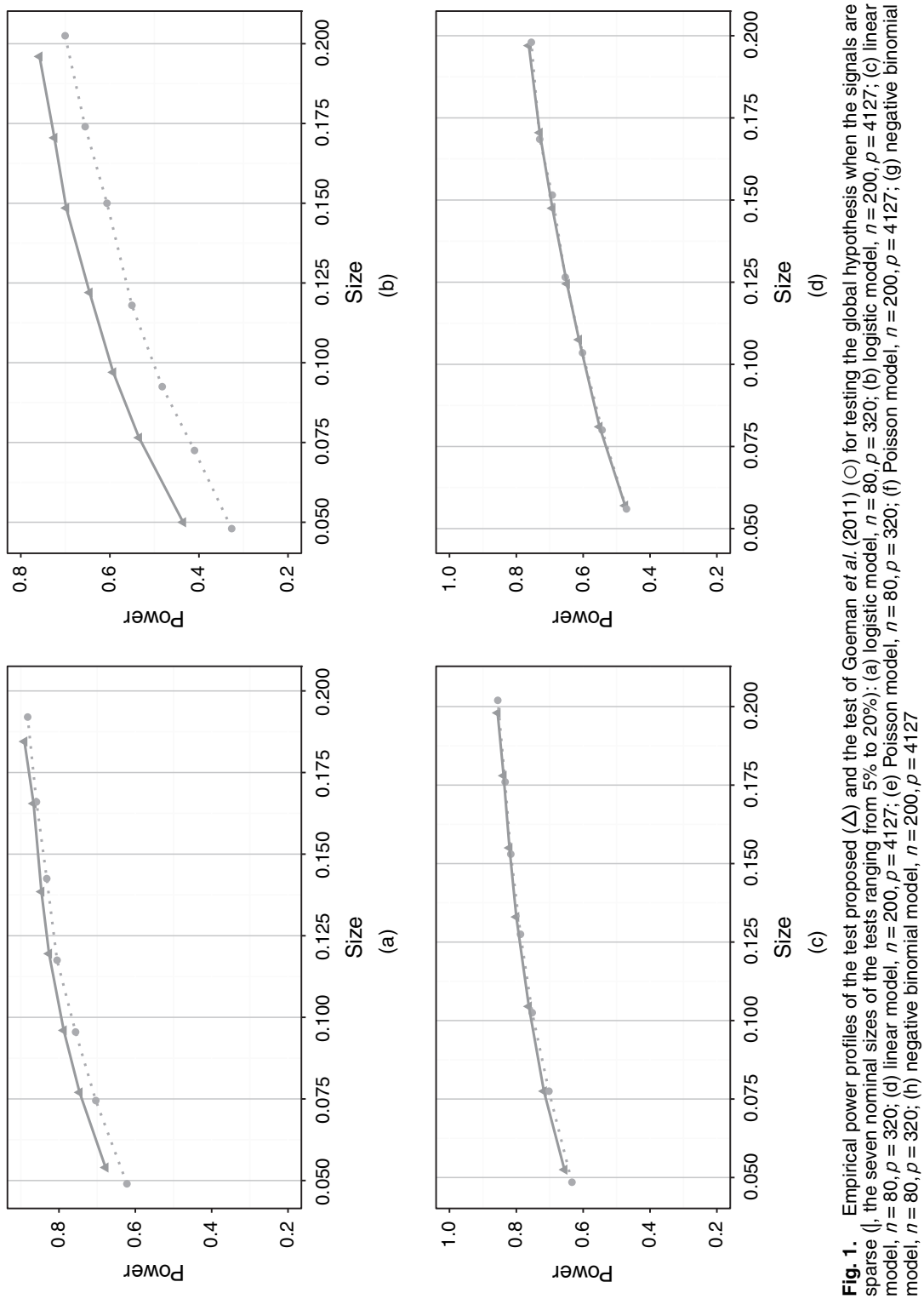

Fig. 1. Empirical power profiles of the test proposed (Δ) and the test of Goeman *et al.* (2011) ($\bigcirc$) for testing the global hypothesis when the signals are sparse ($|$, the seven nominal sizes of the tests ranging from 5% to 20%): (a) logistic model, $n = 80, p = 320$; (b) logistic model, $n = 200, p = 4127$; (c) linear model, $n = 80, p = 320$; (d) linear model, $n = 200, p = 4127$; (e) Poisson model, $n = 80, p = 320$; (f) Poisson model, $n = 200, p = 4127$; (g) negative binomial model, $n = 80, p = 320$; (h) negative binomial model, $n = 200, p = 4127$

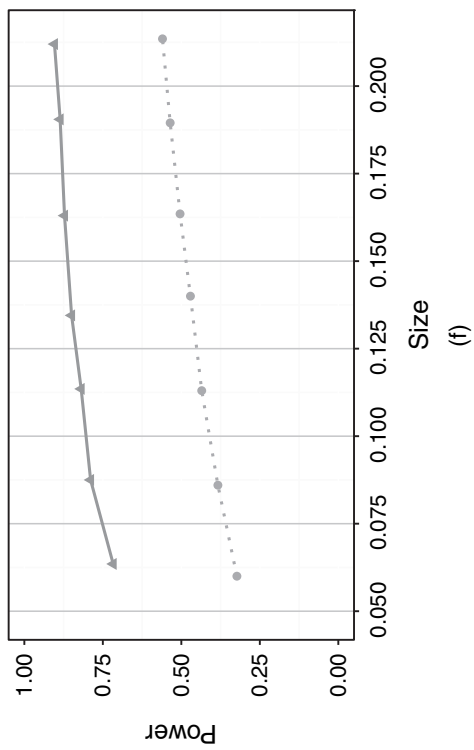

Fig. 1 (*continued*)

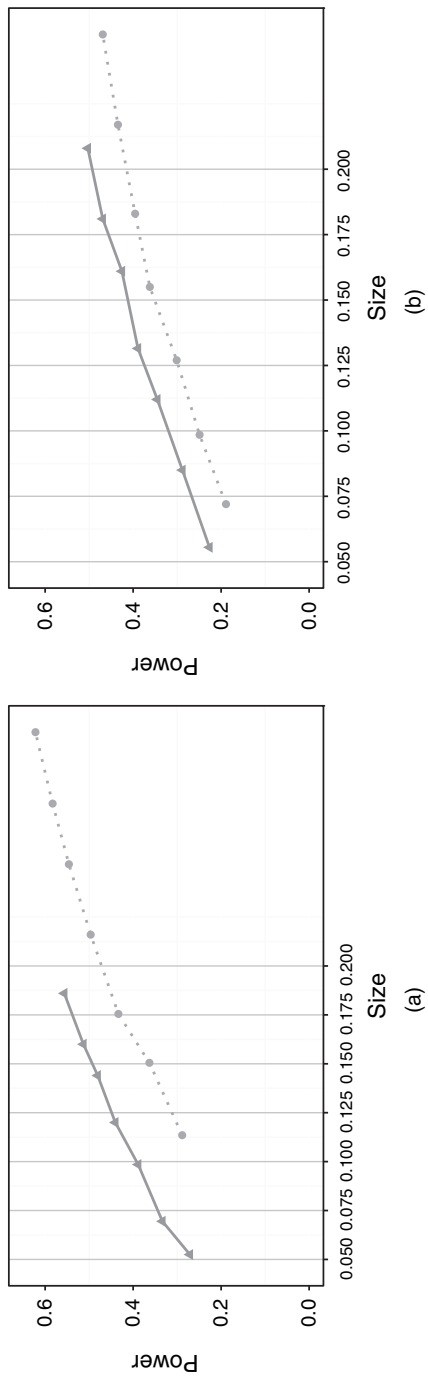

Fig. 2. Empirical power profiles of the test proposed (Δ) and the test of Goeman *et al.* (2011) (○) for testing the hypothesis with nuisance parameters when the signals are sparse (|, the seven nominal sizes of the tests ranging from 5% to 20%): (a) logistic model, $n = 80, p_1 = 10, p_2 = 320$; (b) logistic model, $n = 200, p_1 = 10, p_2 = 4127$; (c) linear model, $n = 80, p_1 = 10, p_2 = 320$; (d) linear model, $n = 200, p_1 = 10, p_2 = 4127$; (e) Poisson model, $n = 80, p_1 = 10, p_2 = 320$; (f) Poisson model, $n = 200, p_1 = 10, p_2 = 4127$; (g) negative binomial model, $n = 80, p_1 = 10, p_2 = 320$; (h) negative binomial model, $n = 200, p_1 = 10, p_2 = 4127$

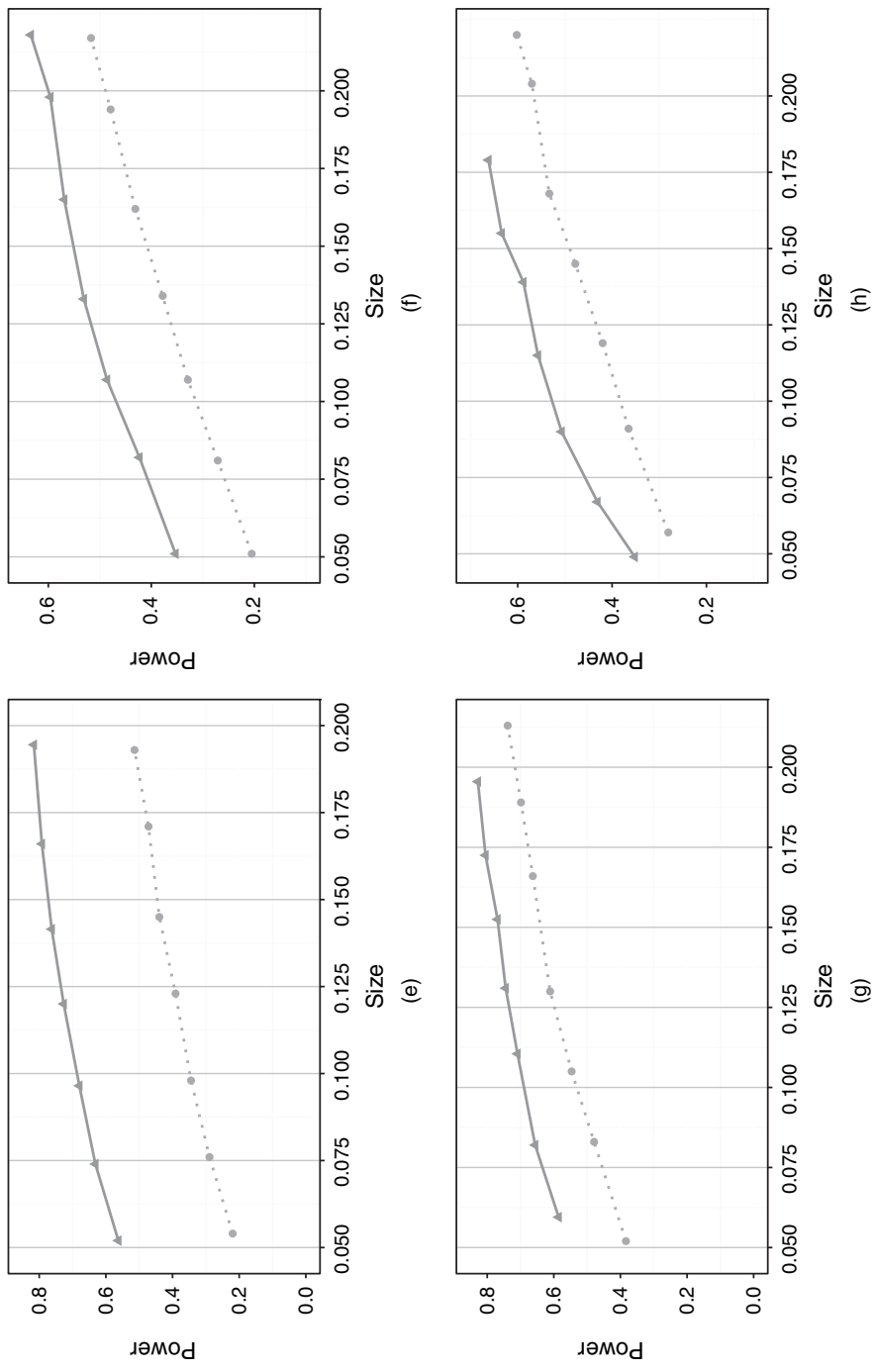

Fig. 2 (*continued*)

$$E(Y_i|X_i) = g(X_i^{\mathrm{T}}\beta) = \frac{\exp(X_i^{\mathrm{T}}\beta)}{1+\exp(X_i^{\mathrm{T}}\beta)},$$

and $Y_i|X_i \sim \text{Bernoulli}\{1, g(X_i^{\mathrm{T}}\beta)\}$. In the linear regression,

$$E(Y_i|X_i) = g(X_i^{\mathrm{T}}\beta) = X_i^{\mathrm{T}}\beta,$$

and $Y_i|X_i \sim N(X_i^{\mathrm{T}}\beta, 1)$. We note here that the test is invariant with respect to the nuisance dispersion parameter σ^2. Hence, we chose $\sigma = 1$. In the Poisson regression,

$$E(Y_i|X_i) = g(X_i^{\mathrm{T}}\beta) = \exp(X_i^{\mathrm{T}}\beta),$$

and $Y_i|X_i \sim \text{Poisson}\{g(X_i^{\mathrm{T}}\beta)\}$. In the negative binomial model,

$$Y|\lambda \sim \text{Poisson}(\lambda) \qquad \lambda \sim \text{gamma}\{\exp(X^{\mathrm{T}}\beta), 1\},$$

and $Y_i|X_i \sim \text{NB}\{\exp(X_i^{\mathrm{T}}\beta), \frac{1}{2}\}$, which prescribes overdispersion to the Poisson model.

To create regimes of high dimensionality, we chose (p, n) according to $p = \exp(n^{0.4})$ and specifically considered $(n, p) = (80, 320)$ and $(n, p) = (200, 4127)$ in the simulations. Both sparse and dense signals were experimented with; their details will be provided shortly. Seven nominal type I errors ranging from 0.05 to 0.2 were considered, and the corresponding empirical sizes and powers were evaluated from 2000 replications.

We first considered testing the global hypothesis

$$H_0 : \beta = \mathbf{0}_{p\times 1} \quad \textit{versus} \quad H_1 : \beta \neq \mathbf{0}_{p\times 1}. \tag{6.1}$$

In designing the alternative hypothesis for the linear model, we set $\|\beta\|^2 = 0.2$, the first five coefficients in β to be non-zero of equal magnitude and the rest of the coefficients to be 0. Hence, the non-zero coefficients were quite sparse. For the other three models, we impose the same sparsity on β as in the linear model but made $\|\beta\|^2 = 2$. To have a reasonable range for the response variable, as in Goeman *et al.* (2011), we restricted $E(Y_i|X_i)$ between $\exp(-4)/\{1+\exp(-4)\} = 0.02$ and $\exp(4)/\{1+\exp(4)\} = 0.98$ for the logistic model, between -1000 and 1000 for the linear model, and between $\exp(0) = 1$ and $\exp(4) = 55$ for the Poisson and negative binomial models respectively.

Fig. 1 displays the power–size profile (curves that link the empirical power and the corresponding empirical size). It can be seen that the global test proposed and the test of Goeman *et al.* (2011) had largely similar power profiles for the logistic and linear models as displayed by Figs 1(a)–1(d). We collected, under the linear and logistic models, the simulated averages of $\|\Delta_{\beta,\beta_0}\|^2$, which were 7.00 and 0.17 respectively. These averages may be used to argue that the simulation settings were in the regime of the local alternatives, which explained according to corollaries 1 and 2 that both tests had similar powers. Figs 1(a) and 1(b) show that the test proposed had a slightly higher power than that of Goeman *et al.* (2011) in the case of the logistic model. This may be understood as the effect of the A_n-term on the variance of S_n despite its being second order under the local alternatives.

Figs 1(e)–1(h) show a much larger discrepancy in the power profiles between the two tests for the Poisson and negative binomial models with the test proposed being significantly more powerful. For these two models, the averages $\|\Delta_{\beta,\beta_0}\|^2$ were respectively 310.27 and 302.40, which were much larger than those under the linear and the logistic models that were reported in the previous paragraph. The large $\|\Delta_{\beta,\beta_0}\|^2$ implied that the simulations were in the family of the fixed alternatives $\mathscr{L}_\beta^{\mathrm{F}}$. The simulated power profiles confirmed the findings in theorem 2 that, under $\mathscr{L}_\beta^{\mathrm{F}}$, the power of the test of Goeman *et al.* (2011) can be adversely affected by the high dimensionality, whereas the test proposed withstands the high dimensionality quite nicely.

We then conducted simulations for testing

$$H_0 : \beta^{(2)} = \mathbf{0}_{p_2 \times 1} \quad \textit{versus} \quad H_1 : \beta^{(2)} \neq \mathbf{0}_{p_2 \times 1} \tag{6.2}$$

in the presence of nuisance parameter $\beta^{(1)}$ for the same four generalized linear models as considered above. The nuisance parameter $\beta^{(1)}$ was 10 (p_1) dimensional, generated randomly from $U(0, 1)$. We chose $(n, p_2) = (80, 320)$ and $(n, p_2) = (200, 4127)$ via $p_2 = \exp(n^{0.4})$. To evaluate the power of the test, the first five elements of $\beta^{(2)}$ were set to be non-zero of equal magnitude with $\|\beta^{(2)}\|^2 = 0.5$ for the linear model and $\|\beta^{(2)}\|^2 = 2$ for the other three generalized linear models, whereas the rest of the $\beta^{(2)}$s were 0s.

The power profiles of the test proposed and the test of Goeman *et al.* (2011) are displayed in Fig. 2. It can be seen from Fig. 2(a) that, for the logistic model with $n = 80$ and $p_2 = 320$, the test of Goeman *et al.* (2011) had a very severe size distortion. When the sample size was increased to $n = 200$, Fig. 2(b) shows that the size distortion was reduced in comparison with that of $n = 80$. The size distortion with the test of Goeman *et al.* (2011) was largely absent for the test proposed. Fig. 2 shows that the test proposed had quite reasonable power with good control of the empirical size. For the Poisson and negative binomial models (Figs 2(e)–2(h)), it is observed that the test proposed had more advantageous power profiles than those of the test of Goeman *et al.* (2011). The large discrepancy in power profiles shown in Figs 2(e)–2(h) was similar to the global tests as demonstrated in Fig. 1.

As the simulation results conveyed in Figs 1 and 2 were based on designs where the signals (non-zero coefficients) under the alternatives were sparse, we experimented with dense signals settings. Specifically, we set all the components of β or $\beta^{(2)}$ to be non-zero under the alternatives in both the global test and the test with nuisance parameters, while replicating the other aspects of the simulation designs for Figs 1 and 2 reported above. The empirical power profiles for the global test with dense signals are shown in Fig. SM1, whereas those for the nuisance parameter are reported in Fig. SM2 in the on-line supplementary material. Both figures respectively conveyed very similar patterns to those in Figs 1 and 2 under sparse signals. It was fair to say that neither the test proposed nor the test of Goeman *et al.* (2011) was sensitive to the signals being sparse or dense. Whereas the test proposed and the test of Goeman *et al.* (2011) had similar performances for the logistic and linear models, their performances were sharply different for the Poisson and negative binomial models, where the test proposed outperformed the test of Goeman *et al.* (2011) with much higher power.

7. Empirical study

We analyse a data set that contained microarray expressions for 22 283 genes in a study on the large airway epithelial cells for smokers, which is available from `http://www.ncbi.nlm.nih.gov/geo/` with access label GDS2771. The study consisted of 187 smokers, of whom the treatment group consisted of 97 smokers with lung cancer, and the control group consisted of 90 non-cancer smokers. The data set included information on participants' age and gender, and a binary indicator on whether a participant had quit smoking for less than 10 years or not. The data set has been analysed in Spira *et al.* (2007) and Danaher *et al.* (2014), with different inferential purposes.

Each gene tends to work collaboratively with other genes to perform certain biological tasks and biologists have defined gene sets under the GO system. The gene sets, which are also called GO terms, have been classified into three broad functional categories: biological processes, cellular components and molecular functions.

The main purpose of the case-study is to demonstrate the test procedures proposed, and in

particular their ability to provide p-values for the significance of gene sets in the context of high dimensional generalized linear models. Without p-values, it is difficult to argue for the significance of any gene sets under both high dimensionality and multiplicity.

After preliminary gene filtering by using the algorithm proposed in Gentleman *et al.* (2005), there were 2249 unique gene sets in biological processes, 320 in cellular components and 411 in molecular functions, which involved 10 114 unique genes in total. The study's aim was to identify gene sets which were significantly expressed between the lung cancer and the control group for each functional category. We formulated the analysis as a binary regression problem with the response Y_i being 1 if the participant was in the lung cancer group and 0 if in the control. The covariate $X_{ig} = (X_i^{(1)\mathrm{T}}, X_{ig}^{(2)\mathrm{T}})^{\mathrm{T}}$ for the ith participant's profile corresponding to a gene set g, where $X_i^{(1)}$ contains a gender, age and a binary indicator for quitting smoking for less than 10 years, and $X_{ig}^{(2)}$ is the vector of gene expressions of the gth gene set.

We considered the logistic regression model

$$E(Y_i|X_i^{(1)}, X_{ig}^{(2)}) = \frac{\exp(X_i^{(1)\mathrm{T}}\beta_g^{(1)} + X_{ig}^{(2)\mathrm{T}}\beta_g^{(2)})}{1+\exp(X_i^{(1)\mathrm{T}}\beta_g^{(1)} + X_{ig}^{(2)\mathrm{T}}\beta_g^{(2)})}. \tag{7.1}$$

The focus of the study was to find significant gene sets, while considering the effects of the three non-genetic covariates in $X^{(1)}$. The study was carried out by treating $\beta_g^{(1)}$ as the nuisance parameter and testing the hypothesis

$$H_0 : \beta_g^{(2)} = 0 \quad \textit{versus} \quad H_1 : \beta_g^{(2)} \neq 0.$$

The FDR (Benjamini and Hochberg, 1995) is widely used in multiple-hypothesis testing to control the expected proportion of false positive results among all rejected null hypotheses. Storey *et al.* (2004) proposed a procedure which can incorporate dependence between p-values and is shown to be less conservative than the original FDR control method of Benjamini and Hochberg (1995). By controlling the FDR by using the procedure of Storey *et al.* (2004) at 1%, 1108 gene sets in the functional category of biological processes, 183 in cellular components and 186 in molecular functions were found significant.

Fig. 3 displays the p-values and the standardized test statistics. Among those significant gene sets in biological processes, there was one with identifier GO 0007049. The GO term has a close relationship with the phosphatidylinositol 3-kinase pathway. The activation of the pathway is significantly increased in the cytologically normal bronchial airway of smokers with lung cancer (Gustafson *et al.*, 2010).

We also carried out the global test for the significance of the entire regression coefficient vector $\beta_g = (\beta_g^{(1)\mathrm{T}}, \beta_g^{(2)\mathrm{T}})^{\mathrm{T}}$ by performing tests on

$$H_0 : \beta_g = 0 \quad \textit{versus} \quad H_1 : \beta_g \neq 0.$$

Fig. 4 displays the histograms of p-values and the standardized global test statistics. It can be seen that the bulk of the test statistics (Figs 4(d)–4(f)) took extremely large values in the scale of the standard normal distribution, implying that most of the p-values would be very small around zero and the significance of many sets of genes. Comparing Fig. 3 with Fig. 4, it is found that the bodies of the histograms were much less extreme in Fig. 3 than those in Fig. 4. By controlling the FDR at 1%, 1608 gene sets in the functional category of biological processes, 245 in cellular components and 281 in molecular functions were found significant. Comparing these with the test results for $H_0 : \beta_g^{(2)} = 0$ by treating $\beta^{(1)}$ as the nuisance parameter, we see a substantial increase in the number of significant gene sets. The increase was largely driven by the difference in the three non-genetic covariates, rather than the genetic aspects. This highlights that filtering

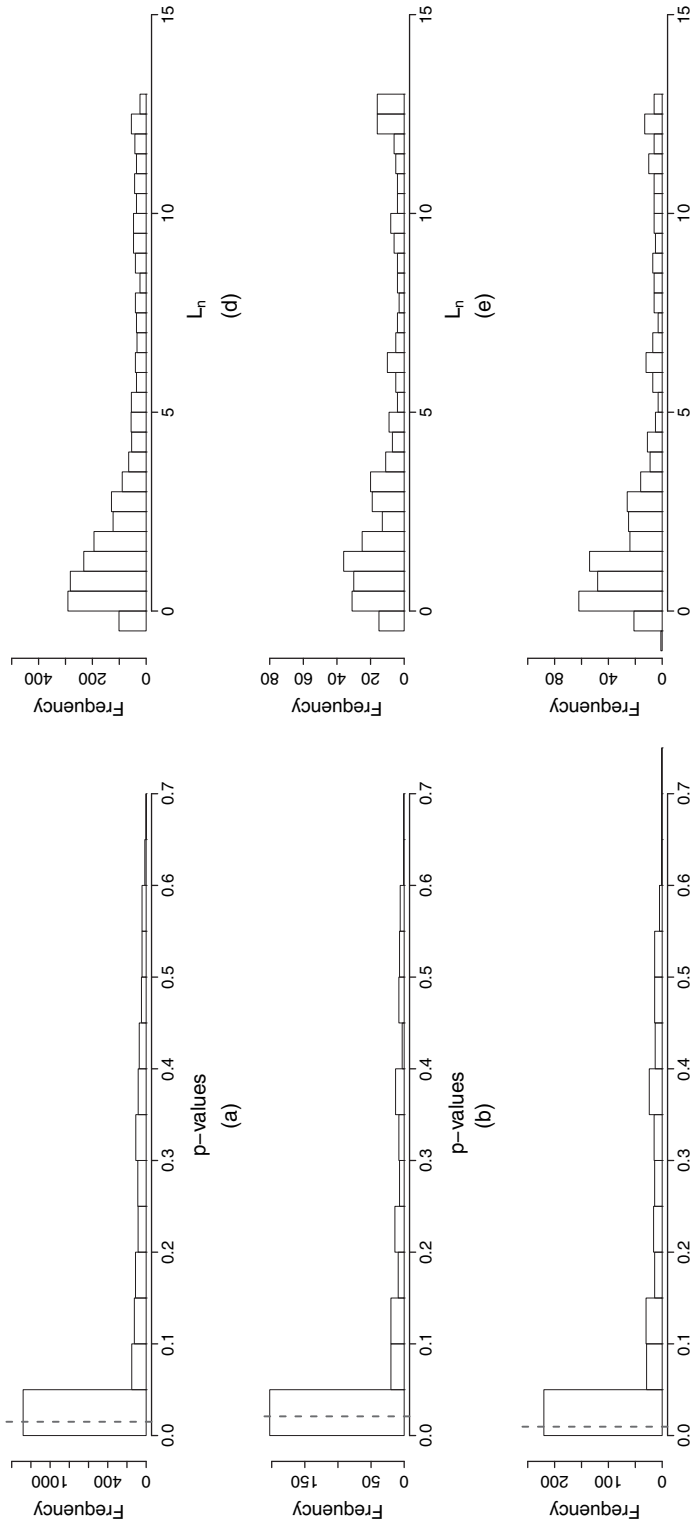

Fig. 3. (a)–(c) Histograms of p-values and (d)–(f) the standardized test statistics in the presence of the nuisance parameter $\dot{i}$, critical p-values corresponding to the 1% FDR): (a), (d) biological processes; (b), (e) cellular components; (c), (f) molecular functions

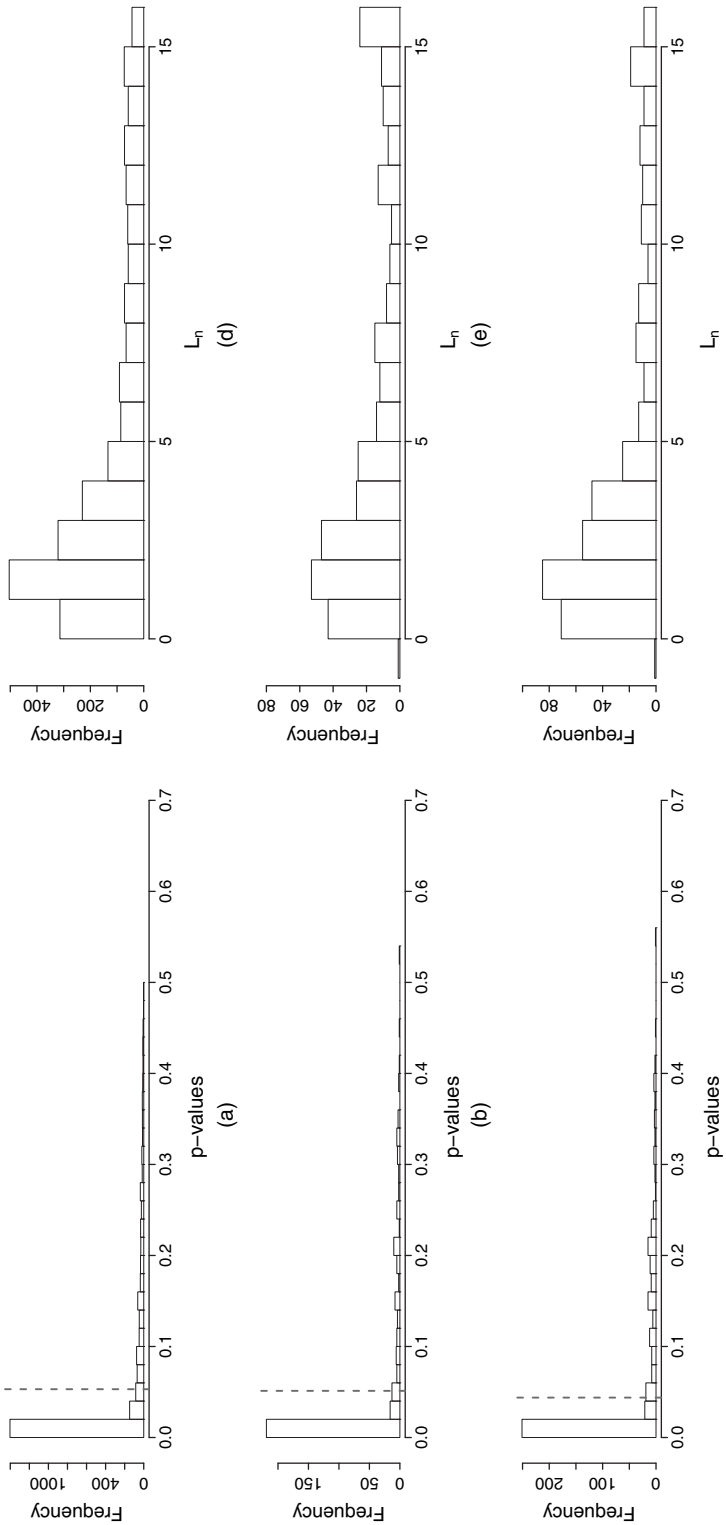

Fig. 4. (a)–(c) Histograms of *p*-values and (d)–(f) the standardized test statistics for the global hypothesis $\dot{i}$, critical *p*-values corresponding to the 1% FDR): (a), (d) biological processes; (b), (e) cellular components; (c), (f) molecular functions

out the effect of the three non-genetic covariates was necessary when considering the effect of the gene sets between the lung cancer group and the controls. The approach proposed offers a way to do this.

8. Discussion

As generalized linear models are widely used tools in analysing genetic data, the tests proposed, as they are more adaptive to the high dimensionality, are useful additions to the existing test procedures for the significance of regression coefficients for generalized linear models. As shown in the case-study, testing for the significance of gene sets requires high dimensional multivariate test procedures, which can produce p-values under both high dimensionality and multiplicity. The tests proposed and the tests of Goeman *et al.* (2011) may be used for gene sets testing, in conjunction with an FDR control procedure, when testing a large number of hypotheses simultaneously.

The test of Goeman *et al.* (2011) was proposed for fixed dimension p. We have found that it is quite resilient to diverging p in both theoretical analysis and numerical analysis, as long as either the inverse of the link function or its derivative is bounded. The latter encompasses the logistic and linear regression models. The tests proposed are designed to improve the performance of the test of Goeman *et al.* (2011) for diverging p. This is especially so when the difference between $g(X^{\mathrm{T}}\beta)$ and $g(X^{\mathrm{T}}\beta_0)$ is large, since the high dimensionality can exert an adverse influence on the test of Goeman *et al.* (2011). The test statistics proposed, owing to their simpler formulations, can avoid some of the high dimensional effects, and hence lead to better performance with more accurate size approximation and are more powerful. When the dimension p is much smaller than the sample size n, the test that was given in McCullagh and Nelder (1989) and the test of Goeman *et al.* (2011) may be used as the test proposed may not be operational under the smaller dimensionality.

There have been studies on the inference of the regression coefficients associated with the lasso and other variable selection methods for linear models. Most of these methods are based on the sparsity assumption such that the non-zero regression coefficients are sparsely populated; see van de Geer *et al.* (2013), Voorman *et al.* (2014) and Zhang and Zhang (2014). However, the sparsity assumption is quite difficult to validate from data. Our proposed tests are valid without the sparsity assumption and may be used first when the sparsity level of a testing problem is unknown. More research is needed on how to combine the two strains of inference methods in the setting of high dimensional generalized linear models.

Acknowledgements

The authors thank two reviewers for constructive comments and suggestions which improved the presentation of the paper. Part of this research is based on material contained in the Peking University doctoral thesis of Bin Guo. The authors acknowledge support from National Science Foundation grants DSM-1309210, China's National Natural Science Foundation grants 11131002 and 71371016, National Key Basic Research Program grant 2015CB856000 and the Center for Statistical Science and Laboratory of Mathematical Economics and Quantity Finance at Peking University.

References

Bai, Z. D. and Saranadasa, H. (1996) Effect of high dimension: by an example of two sample problem. *Statist. Sin.*, **6**, 311–329.

Barry, W. T., Nobel, A. B. and Wright, F. A. (2008) A statistical framework for testing functional categories in microarray data. *Ann. Appl. Statist.*, **2**, 286–315.
Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.
Chang, J., Tang, C. Y. and Wu, Y. (2013) Marginal empirical likelihood and sure independence feature screening. *Ann. Statist.*, **41**, 2123–2148.
Chen, S. X. and Cui, H. J. (2003) An extended empirical likelihood for generalized linear models. *Statist. Sin.*, **13**, 69–81.
Chen, S. X., Peng, L. and Qin, Y. L. (2009) Effects of data dimension on empirical likelihood. *Biometrika*, **96**, 711–722.
Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. New York: Chapman and Hall.
Danaher, P., Wang, P. and Witten, D. M. (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Statist. Soc.* B, **76**, 373–397.
Fahrmeir, L. and Tutz, G. (1994) *Multivariate Statistical Modelling based on Generalized Linear Models*, 2nd edn. New York: Springer.
Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc.* B, **70**, 849–911.
Fan, J. and Song, R. (2010) Sure independent screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, **38**, 3567–3604.
van de Geer, S. A. (2008) High-dimensional generalized linear models and the lasso. *Ann. Statist.*, **36**, 614–645.
van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2013) On asymptotically optimal confidence regions and tests for high-dimensional models. *Preprint arXiv:1303.0518*.
Gentleman, R., Irizarry, R. A., Carey, V. J., Dudoit, S. and Huber, W. (2005) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. New York: Springer.
Goeman, J. J., van de Geer, S. A. and van Houwelingen, H. C. (2006) Testing against a high dimensional alternative. *J. R. Statist. Soc.* B, **68**, 477–493.
Goeman, J. J., Van Houwelingen, H. C. and Finos, L. (2011) Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika*, **98**, 381–390.
Gustafson, A. M., Soldi, R., Anderlind, C., Scholand, M. B., Qian, J., Zhang, X., Cooper, K., Walker, D., McWilliams, A., Liu, G., Szabo, E., Brody, J., Massion, P. P., Lenburg, M. E., Lam, S., Bild, A. H. and Spira, A. (2010) Airway PI3K pathway activation is an early and reversible event in lung cancer development. *Sci. Transltnl Med.*, **2**, 26ra25.
Lan, W., Wang, H. and Tsai, C. L. (2014) Testing covariates in high-dimensional regression. *Ann. Inst. Statist. Math.*, **66**, 279–301.
Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2014) Exact inference after model selection via the Lasso. *Preprint arXiv:1311.6238*. Stanford University, Stanford.
Lockhart, R., Taylor, J., Tibshirani, J. R. and Tibshirani, R. (2014) A significance test for the lasso (with discussion). *Ann. Statist.*, **42**, 413–468.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Pan, W. (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidem.*, **33**, 497–507.
Spira, A., Beane, J., Shah, V., Steiling, K., Liu, G., Schembri, F., Gilman, S., Dumas, Y., Calner, P., Sebastiani, P., Sridhar, S., Beamis, J., Lamb, C., Anderson, T., Gerry, N., Keane, J., Lenburg, M. and Brody, J. (2007) Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat. Med.*, **13**, 361–366.
Storey, J. D., Taylor, J. E. and Siegmund, D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc.* B, **66**, 187–205.
Taylor, J., Lockhart, R., Tibshirani, J. R. and Tibshirani, R. (2014) Post-selection adaptive inference for Least Angle Regression and the Lasso. *Preprint arXiv:1401.3889*. Stanford University, Stanford.
van der Vaart, A. W. (2000) *Asymptotic Statistics*. Cambridge: Cambridge University Press.
Voorman, A., Shojaie, A. and Witten, D. (2014) Inference in high dimensions with the penalized score test. *Preprint arXiv:1401.2678*. University of Washington, Seattle.
Wedderburn, R. W. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.
Zhang, C.-H. and Zhang, S. S. (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc.* B, **76**, 217–242.
Zhong, P. S. and Chen, S. X. (2011) Tests for high dimensional regression coefficients with factorial designs. *J. Am. Statist. Ass.*, **106**, 260–274.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplemental material: Tests for high dimensional generalized linear models'.