Submitted to the Annals of Statistics

DETECTING RARE AND FAINT SIGNALS VIA THRESHOLDING MAXIMUM LIKELIHOOD ESTIMATORS

By Yumou Qiu[†], Song Xi Chen[‡] and Dan Nettleton[§]

University of Nebraska Lincoln[†], Peking University[‡], and Iowa State University[§]

Motivated by the analysis of RNA sequencing (RNA-seq) data for genes differentially expressed across multiple conditions, we consider detecting rare and faint signals in high-dimensional response variables. We address the signal detection problem under a general framework, which includes generalized linear models for count-valued responses as special cases. We propose a test statistic that carries out a multi-level thresholding on maximum likelihood estimators (MLEs) of the signals, based on a new Cramér type moderate deviation result for multi-dimensional MLEs. Based on the multi-level thresholding test, a multiple testing procedure is proposed for signal identification. Numerical simulations and a case study on maize RNA-seq data are conducted to demonstrate the effectiveness of the proposed approaches on signal detection and identification.

1. Introduction. With the advance of technology, high-dimensional data are becoming increasingly common in scientific studies ranging from bioinformatics, signal processing and astrophysics. An important task in these studies is to detect rare and faint signals leading to scientific discovery. The goal of signal detection is to determine the existence of signals in the parameters of interest based on noisy data. If any signal is detected, identifying the subset of parameters carrying the signal becomes important. Suppose each observation consists of a high-dimensional response vector and a low-dimensional vector of explanatory variables which can represent treatment regimes and covariate information. In this paper, we intend to detect and identify signals defined, in general, by an association between the explanatory vector and the high-dimensional response.

A primary motivation for our work is analyzing data from Next Generation Sequencing of RNA (RNA-seq), which provides information about transcript abundance for each gene. When there are two or more treatments, we are interested in detecting whether any of the genes are differentially expressed across treatments. Unlike continuous microarray data, RNA-seq data

[‡]Corresponding author

Keywords and phrases: Detection boundary, False discovery proportion, Generalized linear model, Moderate deviation, Multiple testing procedure, RNA-seq data.

are usually collected in the form of counts that are associated with expression levels of genetic features. Generalized linear models (GLMs) and their extensions are often used to model such data. See, for example, Robinson and Smyth (2007, 2008), Anders and Huber (2010) and Lund et al. (2012).

In this paper, we first consider testing sparse and faint covariate effects among all the responses variables, which includes testing regression coefficients in GLMs as a special case. This amounts to testing whether certain linear combinations of parameters are nonzero when the deviations of the linear combinations from zero (signals) are small in magnitude (faintness of signal) and occur in few dimensions (rareness of signals). We consider the high-dimensional paradigm where the number of the response variables is much larger than the number of replications. Due to high dimensionality and rareness and faintness of signals, we carry out thresholding on maximum likelihood estimates (MLEs) of the signals in each dimension to remove non-signal bearing dimensions. A thresholding test statistic is constructed by summing the thresholded signal MLEs over all the response dimensions based on a newly established Cramér type moderate deviation result for the MLEs. We propose a multi-level thresholding test statistic constructed by maximizing the standardized thresholding statistic over a set of thresholds. This provides a data-driven strategy for automated threshold selection that produces an attractive detection boundary for testing rare and faint signals. Built on the promise of the multi-level thresholding tests, we propose a procedure for identifying the signal-bearing dimensions.

For testing rare and faint signals in means, Donoho and Jin (2004) showed that the Higher Criticism (HC) test can attain the optimal detection boundary (Ingster, 1997) for uncorrelated Gaussian random vectors; see Delaigle, Hall and Jin (2011), Hall and Jin (2008, 2010) and Zhong, Chen and Xu (2014) for further studies and extensions. Testing for high dimensional means is advantageous because estimators for the means (i.e., sample means) are readily available, as well as the large deviation results for the sample means needed for the analysis of HC statistics. For testing the regression coefficients in GLMs, the study becomes more challenging. Although MLEs for the regression coefficients can be obtained, we need moderate deviation results for MLEs to uncover the performance of the proposed thresholding statistics. The new moderate deviation result with specific error rates allows us to analyze the properties of the proposed test. There is work for GLMs with univariate response but high dimensional covariates, which includes Fan and Song (2010) for Sure Independent Screening of covariates, and Zhong and Chen (2011), Goeman et al. (2011) and Guo and Chen (2016) for testing high dimensional regression coefficients. Statistical inference for data with sparse

and faint signals has been also considered in the areas of high-dimensional linear regression and classification; see, for example, Arias-Castro, Candès and Plan (2011), Ji and Jin (2012) and Fan, Jin and Yao (2013). We address a different problem where covariates are low dimensional but responses are high dimensional, reflecting the situation of RNA-seq data.

The paper is organized as follows. The models and hypothesis of interest are introduced in Section 2. Section 3 presents thresholding for MLEs together with the moderate deviation result. The multi-level thresholding test is proposed in Section 4, where the powers of both the single- and multiple-level thresholding tests are investigated. Signal identification is discussed in Section 5. Simulation results and an analysis of maize RNA-seq data are presented in Sections 6 and 7, respectively. Section 8 provides extensions of the proposed methods. Technical details are given in both the Appendix and the supplementary material.

2. Models and Hypotheses. Suppose p response variables are measured for n experimental units. Let y_{ij} be the measurement of response j for experimental unit i ($i = 1, \cdots, n$ and $j = 1, \cdots, p$). Let $z_i = (z_{i1}, \cdots, z_{im})'$ be a vector of fixed and known explanatory variable values for experimental unit i, and let $\beta_j = (\beta_{j1}, \cdots, \beta_{jm})'$ be parameters representing explanatory variable effects and ϕ_j be an ancillary parameter for response j. Let $f_j(y; z_i, \theta_j)$ be the density function of y_{ij}, where $\theta_j = (\beta_j', \phi_j)'$.

We are interested in testing, for a known matrix $D_{d\times m}$,

$$H_0 : D\beta_j = 0 \text{ for all } j \quad \text{vs.} \quad H_a : \ D\beta_j \neq 0 \text{ for some } j. \tag{2.1}$$

The matrix D is determined by the context of an application that would lead to a specific model parameterization and the hypothesis of interest in terms of model parameters. Each row of the matrix D contains coefficients of a linear combination of the parameters. Under the null hypothesis, the d linear combinations of regression coefficients determined by D are all zero for all j. Hypothesis (2.1) is a general setup, which includes testing for main effects of factors and interactions among factors as special cases. For example, in the case of testing the equivalence of two sample means where $\beta_j = (\beta_{j1}, \beta_{j2})'$ stands for the population means of the two groups, D may be chosen as $(1, -1)$. As another example, consider the test for interaction in a 2-by-2 factorial design, where $\beta_j = (\beta_{j,11}, \beta_{j,12}, \beta_{j,21}, \beta_{j,22})'$ is the vector of treatment means corresponding to the four combinations of factor levels. Then D may be $(1, -1, -1, 1)$ to specify the null hypothesis of no interaction.

We consider a setting of (2.1) which facilitates the study of test performance for the challenging case of sparse and weak signals. Let $\beta_{j,0}$ be the

value of β_j under the null hypothesis. Suppose that under H_a, β_j takes the value $\beta_{j,0}$ with probability $1-\epsilon$ and takes the value $\beta_{j,0}+\beta_{j,a}$ with probability ϵ for an $\epsilon > 0$. This means only ϵ proportion of the responses are expected to carry signal under the alternative with signal strength $\beta_{j,a}$. This leads to a specific form of the hypotheses in (2.1):

$$\begin{aligned} &H_0 : \beta_j = \beta_{j,0} \ \text{ such that } D\beta_{j,0} = 0 \text{ for all } j \ \text{ vs.} \\ &H_a : \beta_j \overset{ind}{\sim} (1-\epsilon)\nu_{\beta_{j,0}} + \epsilon\nu_{\beta_{j,0}+\beta_{j,a}} \ \text{ for all } j, \end{aligned} \tag{2.2}$$

where ν_β stands for the point mass distribution at β, $\epsilon = p^{-\kappa}$ for $\kappa \in (0,1)$, and $\beta_{j,a} = r_j\sqrt{2(\log p)/n}$ for an m-dimensional vector $r_j \in (0,1)^m$. Here, κ and $\{r_j\}$ specify the sparsity and the signal strengths, respectively. The signal strengths $\{r_j\}$ are assumed to be independently drawn from a super population compactly supported on a set $\mathcal{G}$ where $P(Dr_j \neq 0) = 1$.

Note that (2.2) is a specialized version of (2.1), which has been used to evaluate high-dimensional test procedures in the literature, for instance Donoho and Jin (2004) and Hall and Jin (2010). Although the method we develop is for testing the hypotheses in (2.1), it is important to understand its performance for testing (2.2), which offers the most challenging setting for signal detection in high dimension. The challenge is reflected in two aspects: the sparsity and faintness of signals under H_a in (2.2). Good performance for testing (2.2) implies good performance for testing (2.1) and is important in a variety of applications, including the empirical study of RNA-seq data in Section 7, where important biological signals could be both rare and faint. Thus, we construct a test that is powerful under (2.2) and also directly applicable in the general settings of (2.1).

A relevant special case of (2.2) is when only one component of β_j is of interest, say β_{jk} for some $k \in \{1, \cdots, m\}$. Let $\beta_{jk,a}$ be the kth element of $\beta_{j,a}$. The corresponding hypotheses under consideration are

$$H_0 : \beta_{jk} = 0 \text{ for all } j \ \text{ vs. } \ H_a\text{: } \beta_{jk} \sim (1-\epsilon)\nu_0 + \epsilon\nu_{\beta_{jk,a}}. \tag{2.3}$$

Testing for treatment effects under GLMs is a special case of the above framework, where the distribution of y_{ij} is within the exponential family with mean μ_{ij} and dispersion parameter ϕ_j. The relationship between μ_{ij} and β_j is modeled as $g(\mu_{ij}) = z_i'\beta_j$ via a link function $g(\cdot)$. The following are some specific models.

Linear Regression. The mean μ_{ij} of the response is linearly related to the covariates z_i as $y_{ij} = z_i'\beta_j + \varepsilon_{ij}$ for i.i.d. error $\varepsilon_{ij} \sim N(0, \sigma_j^2)$ ($\phi_j = \sigma_j^2$).

Binomial Regression. Suppose y_{ij} follows a binomial distribution with parameters n_{ij} and p_{ij}, where $p_{ij} = \mathrm{E}(y_{ij}/n_{ij})$ is the expected success propor-

tion. With the logistic link, the relationship between p_{ij} and z_i is prescribed as $p_{ij} = \exp(z_i'\beta_j)/\{\exp(z_i'\beta_j) + 1\}$, and $\phi_j = 1$.

Poisson Regression. The dependence of $\mu_{ij} = \mathrm{E}(y_{ij})$ on the covariates z_i is usually assumed to be $\log(\mu_{ij}) = z_i'\beta_j$. Under the Poisson model, the dispersion parameter is a constant (i.e., $\phi_j = 1$).

Negative Binomial Regression. Real data, such as RNA-seq data, often show evidence of over-dispersion, where the variance of the response is larger than its mean. The negative binomial distribution provides a way to account for over-dispersion because its variance increases quadratically as the mean increases: $\mathrm{Var}(y_{ij}) = \mu_{ij} + \mu_{ij}^2/\phi_j$. We consider the log link $\log(\mu_{ij}) = z_i'\beta_j$, and assume that the dispersion parameter ϕ_j is unknown and may change from one response variable to another.

3. Thresholding MLEs. In this section, we construct a thresholding test for the hypothesis (2.1). We have already explained that z_i are fixed and known. For the jth response variable, define

$$I_{\theta,j}(\theta_j) = -\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left\{\frac{\partial^2}{\partial\theta_j\partial\theta_j'}\log f_j(y_{ij}; z_i, \theta_j)\right\}$$

to be the average Fisher information of θ_j, where $\theta_j = (\beta_j', \phi_j)'$. Let $\hat{\theta}_j = (\hat{\beta}_j', \hat{\phi}_j)'$ and $\theta_j^0 = (\beta_j^{0\prime}, \phi_j^0)'$ be the MLE and the true value of θ_j, respectively, where $\hat{\beta}_j = (\hat{\beta}_{j1}, \cdots, \hat{\beta}_{jm})'$. As each θ_j^0 is low dimensional, $\hat{\theta}_j$ can be readily obtained for each dimension. Let $\hat{I}_{\theta,j} = I_{\theta,j}(\hat{\theta}_j)$ be the estimated average Fisher information matrix of θ_j. Let $I_j^{-1}(\theta_j)$ and $\hat{I}_j^{-1} = I_j^{-1}(\hat{\theta}_j)$ be the true and estimated inverse average Fisher information matrix corresponding to β_j, which are the upper-left $m \times m$ blocks of $I_{\theta,j}^{-1}(\theta_j)$ and $\hat{I}_{\theta,j}^{-1}$, respectively. Because there are no treatment effects on most of the responses under a sparse alternative, we apply thresholding on the estimated treatment effects for each response. Let $|\cdot|$, $\|\cdot\|$ and $\mathbf{I}(\cdot)$ be the Euclidean norm for vectors, the Frobenius norm for matrices and the indicator function, respectively.

To formulate the thresholding procedure, we need to first establish a moderate deviation result for MLEs of non-identically distributed data, which requires the following two assumptions.

A1. Suppose Θ is a compact subset of $\mathbf{R}^{m+1}$, and $\theta_j^0 \in \mathrm{int}\,\Theta$. There exist non-negative measurable functions $H_{ij}(\cdot, z_i)$ and $G_{ij}(\cdot, z_i)$, such that for any y in the support of y_{ij},

(i) $|\log f_j(y; z_i, \theta_1) - \log f_j(y; z_i, \theta_2)| \le H_{ij}(y, z_i)|\theta_1 - \theta_2|$ for any $\theta_1, \theta_2 \in \Theta$ and $\limsup n^{-1}\sum_{i=1}^n \mathrm{E}H_{ij}(y_{ij}, z_i) \le H_j < \infty$ for $H_j > 0$;

(ii) there exists a constant $\delta_0 > 0$ such that for $\theta_1 \in \Theta$ and $|\theta_1 - \theta_j^0| < \delta_0$,

$$\left\| \frac{\partial^2}{\partial\theta\partial\theta'} \log f_j(y; z_i, \theta_1) - \frac{\partial^2}{\partial\theta\partial\theta'} \log f_j(y; z_i, \theta_j^0) \right\| \le G_{ij}(y, z_i)|\theta_1 - \theta_j^0|$$

and $\limsup n^{-1} \sum_{i=1}^n \mathrm{E} G_{ij}(y_{ij}, z_i) \le G_j < \infty$ for $G_j > 0$.

A2. There exists a constant $\delta > 0$ such that $\mathrm{E}\big[\exp\{\delta|\frac{\partial}{\partial\theta} \log f_j(y_{ij}; z_i, \theta_j^0)|\}\big]$, $\mathrm{E}\big[\exp\{\delta\|\frac{\partial^2}{\partial\theta\partial\theta'} \log f_j(y_{ij}; z_i, \theta_j^0)\|\}\big]$ and $\mathrm{E}\big[\exp\{\delta G_{ij}(y_{ij}, z_i)\}\big]$ are finite.

Assumption A1 prescribes the Lipschitz condition, which is commonly assumed for likelihood inference (Jensen and Wood, 1998; van der Vaart, 2000). The existence of the moment generating function in A2 is a necessary condition for the Cramér type moderate deviation results (Petrov, 1976; Saulis and Statulevičius, 1991). We verify in the supplementary material that these conditions are satisfied for the models discussed in Section 2.

Lemma 1. *Suppose Assumptions A1 and A2 are satisfied for all* $i = 1, \cdots, n$ *and* $j = 1, \cdots, p$. *Then,*

(i) for $w_n = o(n^{1/6})$ *and* $w_n > \sqrt{(2C_0)^{-1} \log n}$,

$$P\big(|\hat{I}_{\theta,j}^{1/2}(\hat{\theta}_j - \theta_j^0)| \ge w_n/\sqrt{n}\big) = P\big(|\mathcal{N}_{m+1}| \ge w_n\big)\{1 + O(w_n^3/\sqrt{n})\},$$

where $\mathcal{N}_{m+1} \sim N(0, I_{m+1})$ *and* $C_0 > 1$ *is a large positive constant;*

(ii) for $w_n = O(\sqrt{n})$ *and some positive constants* C *and* M,

$$P\big(|\hat{I}_{\theta,j}^{1/2}(\hat{\theta}_j - \theta_j^0)| \ge w_n/\sqrt{n}\big) \le C \exp(-w_n^2/M).$$

Lemma 1 (i) provides the Cramér type moderate deviation result for MLEs from independent but not identically distributed data with estimated Fisher information matrix. It shows that the tail of standardized MLEs can be well approximated by the tail of standard normal distribution. Lemma 1 (ii) provides an exponential bound for the tail probability of MLEs. These results suggest that the threshold level for standardized MLEs is $\sqrt{2s \log p}$ for $s \in (0, 1)$.

Lemma 1 holds for i.i.d. data with more concise conditions. For the i.i.d. case, Inglot and Kallenberg (2003) obtained results for the moderate deviation of an MLE $\hat{\theta}$ under model mis-specification, where the amount of mis-specification converges to 0 as $n \to \infty$. They showed that

$$\lim_{n\to\infty} w_n^{-2} \log\big\{P\big(\sqrt{n}|I_\theta^{1/2}(\hat{\theta} - \theta^0)| \ge w_n\big)\big\} = -1/2$$

for $w_n = o(n^{1/2})$, where I_θ is the Fisher information of θ. However, such a result is not enough for the analysis of the thresholding approach. The error rate $w_n^3/\sqrt{n}$ in Lemma 1 is needed to facilitate the analysis for this paper.

We are now ready to define, for an $s \in (0,1)$, a thresholding test statistic for hypothesis (2.1) as

$$T_n(s) = \sum_{j=1}^{p} n(D\hat{\beta}_j)'\hat{V}_j^{-1}(D\hat{\beta}_j)\mathbf{I}\big(|\hat{V}_j^{-1/2}D\hat{\beta}_j| > \sqrt{(2s\log p)/n}\big), \tag{3.1}$$

where $\hat{V}_j/n = D\hat{I}_j^{-1}D'/n$ is the estimated variance of the estimated signals, $D\hat{\beta}_j$. Let e_k be the m-dimensional unit vector with the kth element being 1 and all others being 0. The thresholding test statistics for hypothesis (2.3) can be obtained from (3.1) by setting $D = e_k'$, which leads to

$$T_{n,k}(s) = \sum_{j=1}^{p} n\hat{J}_{j,kk}^{-1}\hat{\beta}_{jk}^2\ \mathbf{I}\big(|\hat{J}_{j,kk}^{-1/2}\hat{\beta}_{jk}| \geq \sqrt{(2s\log p)/n}\big),$$

where $\hat{J}_{j,kk}$ is the kth diagonal element of $\hat{J}_j = \hat{I}_j^{-1}$.

Thresholding approaches have been applied on sample means in the HC test (Donoho and Jin, 2004) for testing high dimensional means. The properties of thresholding on general MLEs are more challenging due to the diverse form of the parameters and less knowledge of moderate deviation results.

To derive the variance of $T_n(s)$, we need to introduce the notion of ρ-mixing. Let $Y_i = (y_{i1}, \cdots, y_{ip})'$ for $i = 1, \cdots, n$, and $\mathcal{F}_a^b(Y_i) = \sigma\{y_{ij} : a \leq j \leq b\}$ be the σ-field generated by Y_i for $-\infty \leq a \leq b \leq \infty$. Define the ρ-mixing coefficients (Bradley, 2005) of the sequence $\{y_{ij}\}_{j=1}^p$ as $\rho_i(k) = \sup_{m\in\mathbb{Z}} \rho\{\mathcal{F}_{-\infty}^m(Y_i), \mathcal{F}_{m+k}^{\infty}(Y_i)\}$, where for two σ-algebras $\mathcal{A}$ and $\mathcal{B}$

$$\rho(\mathcal{A},\mathcal{B}) = \sup\{|\operatorname{Corr}(f,g)| :\ f \in \mathcal{L}^2(\mathcal{A}), g \in \mathcal{L}^2(\mathcal{B})\},$$

where $\operatorname{Corr}(\cdot,\cdot)$ denotes the correlation operator and $\mathcal{L}^2(\mathcal{A})$ is the collection of random variables on $\mathcal{A}$ with finite second moment. The following assumption prescribes the dependence among $\{y_{ij}\}_{j=1}^p$.

A3. The sequences of response variables $\{y_{ij}\}_{j=1}^p$ are ρ-mixing, and the mixing coefficients satisfy $\rho_i(k) \leq C\alpha^k$ for a constant $\alpha \in (0,1)$, any positive integer k and $i = 1, \cdots, n$.

Because the thresholding statistic $T_n(s)$ in (3.1) involves summation over p response variables, we only require that A3 holds for some permutation of the response variables, but we do not need to know the permutation.

Let $\lambda_p(s) = 2s\log p$, and $\bar{F}_d(\cdot)$ and $f_d(\cdot)$ be the survival and the density functions of a chi-square random variable with d degrees of freedom, respectively. Define

$$\begin{aligned} &\mu_0(s) = pd\bar{F}_{d+2}(\lambda_p(s)) \ \text{ and} \\ &\sigma_0^2(s) = pd(d+2)\bar{F}_{d+4}(\lambda_p(s)) - pd^2\bar{F}_{d+2}^2(\lambda_p(s)). \end{aligned} \tag{3.2}$$

Based on Lemma 1 and the ρ-mixing condition, we have the following theorem giving the mean, variance and the limiting distribution of $T_n(s)$.

Theorem 1. *Under H_0, A1, A2, A3 and $\log p = o(n^{1/3})$,*

$$\mathrm{E}\{T_n(s)|H_0\} = \mu_0(s)\{1+O(\lambda_p(s)^{3/2}/\sqrt{n})\},\ \mathrm{Var}\{T_n(s)|H_0\} = \sigma_0^2(s)\{1+o(1)\}$$

for any $s \in (0,1)$ and

$$\frac{T_n(s) - \mathrm{E}\{T_n(s)|H_0\}}{\sqrt{\mathrm{Var}\{T_n(s)|H_0\}}} \xrightarrow{d} N(0,1) \quad \text{as } n, p \to \infty. \tag{3.3}$$

To formulate a testing procedure, $\mathrm{E}\{T_n(s)|H_0\}$ and $\mathrm{Var}\{T_n(s)|H_0\}$ can be estimated by their main orders $\mu_0(s)$ and $\sigma_0^2(s)$, respectively. By Slutsky's theorem, $\{T_n(s) - \mu_0(s)\}/\sigma_0(s)$ converges in distribution to $N(0,1)$ if $\mathrm{E}\{T_n(s)|H_0\} - \mu_0(s) = o\{\sigma_0(s)\}$. The latter is satisfied if $n \sim p^\xi$ for a $\xi \in (0,1)$ and $s > 1-\xi$ as stated in the following corollary.

Corollary 1. *Under H_0, A1, A2, A3 and $n \sim p^\xi$ for a $\xi \in (0,1)$, $\{T_n(s) - \mu_0(s)\}/\sigma_0(s) \xrightarrow{d} N(0,1)$ for $s > 1-\xi$ as $n, p \to \infty$.*

As both $\mu_0(s)$ and $\sigma_0(s)$ are known, a single-level thresholding test rejects H_0 in (2.1) at significance level α if $T_n(s) - \mu_0(s) > z_\alpha \sigma_0(s)$, where z_α is the upper α quantile of $N(0,1)$. The restrictions $n \sim p^\xi$ and $s > 1-\xi$ can be removed if we employ an estimator $\hat{\mu}(s)$ that satisfies

$$\sigma_0^{-1}(s)\big[\mathrm{E}\{T_n(s)|H_0\} - \hat{\mu}(s)\big] = o(1). \tag{3.4}$$

Such an estimator may be constructed by utilizing the specific distributional information of the GLM in conjunction with bias correction. It can be shown that under the linear model with Gaussian errors, $\mu_0(s) = \mathrm{E}\{T_n(s)|H_0\}$ which satisfies the condition (3.4). Implications of using different forms of $\hat{\mu}(s)$ on the proposed multi-level thresholding test will be discussed after Theorem 2 in the next section.

4. Multi-level Thresholding Test. Single-level thresholding is known (Donoho and Jin, 2004) to be incapable in testing sparse and faint signals in (2.2) and (2.3) with unknown signal strength and sparsity. To adapt to the unknown signal strength and sparsity, we propose a multi-level thresholding procedure that considers multiple thresholding levels $s \in (0,1)$. This avoids the issue of threshold selection encountered in the single-level thresholding case. To simplify our exposition, the main results in this section are presented under (3.4); extensions without (3.4) are also discussed.

Donoho and Jin (2004) considered testing the high-dimensional mean of a standard normally distributed random vector. They studied the setting

$$H_0 : \mu_j = 0 \text{ for } j = 1, \cdots, p \quad \text{vs.} \quad H_a : \mu_1, \cdots, \mu_p \overset{i.i.d.}{\sim} (1-\epsilon)\nu_0 + \epsilon\nu_{\mu_a} \tag{4.1}$$

for $\epsilon = p^{-\kappa}$, $\mu_a = r\sqrt{2(\log p)/n}$, $\kappa \in (0,1)$ and $r \in (0,1)$. Ingster (1997) showed that

$$\mathrm{DB}(\kappa) = \begin{cases} \max\{0, \kappa - 1/2\} & \text{if } 0 < \kappa \le 3/4, \\ (1-\sqrt{1-\kappa})^2 & \text{if } 3/4 < \kappa < 1 \end{cases} \tag{4.2}$$

is the optimal detection boundary for testing (4.1) for standard normally distributed data. This means that for any test of hypothesis (4.1),

$$\text{Type I Error} + \text{Type II Error} \to 1 \quad \text{if} \quad r^2 < \mathrm{DB}(\kappa) \tag{4.3}$$

as $n, p \to \infty$. And, there exists an optimal test such that,

$$\text{Type I Error} + \text{Type II Error} \to 0 \quad \text{if} \quad r^2 > \mathrm{DB}(\kappa) \tag{4.4}$$

as $n, p \to \infty$. Donoho and Jin (2004) showed that their HC test attains the optimal detection boundary for independent normal data with unit variance.

We need knowledge about the power of the single-level thresholding test before presenting the multi-level thresholding test. Let

$$\Delta_n(s) = \frac{\mathrm{E}(T_n(s)|H_a) - \mathrm{E}(T_n(s)|H_0)}{\sqrt{\mathrm{Var}(T_n(s)|H_a)}} \tag{4.5}$$

be the signal to noise ratio. Given a nominal level α and H_a in (2.2), the power of the single-level thresholding test is

$$\mathrm{Power}_n(s;\alpha) = P\left(\frac{T_n(s) - \mathrm{E}(T_n(s)|H_a)}{\sqrt{\mathrm{Var}(T_n(s)|H_a)}} > z_\alpha \sqrt{\frac{\mathrm{Var}(T_n(s)|H_0)}{\mathrm{Var}(T_n(s)|H_a)}} - \Delta_n(s) \middle| H_a \right).$$

It can be shown that $\mathrm{Var}(T_n(s)|H_0)/\mathrm{Var}(T_n(s)|H_a)$ is between 0 and 1, and $\big[T_n(s) - \mathrm{E}(T_n(s)|H_a)\big]/\sqrt{\mathrm{Var}(T_n(s)|H_a)}$ is stochastically bounded. To ensure the power converges to 1, $\Delta_n(s)$ has to diverge to ∞ as $n \to \infty$. Hence, $\Delta_n(s)$ is a key power determinant, which depends on the sparsity κ and the signal strengths in $\{r_j\}$.

To make the test adaptive to the unknown sparsity and the signal strength, we consider a test based on multiple threshold levels in the spirit of the HC test of Donoho and Jin (2004) and its L_2 version proposed by Zhong, Chen and Xu (2014). Let

$$\hat{T}_n(s) = \frac{T_n(s) - \hat{\mu}(s)}{\sigma_0(s)},$$

where $\hat{\mu}(s)$, as conveyed in Section 3, is an estimate of $\mathrm{E}\{T_n(s)|H_0\}$ that satisfies (3.4). The strategy is to maximize $\hat{T}_n(s)$ over multiple threshold levels. Let

$$\mathcal{S}_n(\omega) = \{s_j : s_j = n(D\hat{\beta}_j)'\hat{V}_j^{-1}(D\hat{\beta}_j)/(2\log p) \text{ and } s_j \le 1-\omega,\ 1 \le j \le p\}$$

for a small positive constant ω. The multi-level thresholding statistic is

$$\mathcal{T}_n = \max_{s\in\mathcal{S}_n(\omega)} \hat{T}_n(s). \tag{4.6}$$

The following theorem states the asymptotic null distribution of $\mathcal{T}_n$.

Theorem 2. *Under H_0 in (2.1), A1, A2, A3, (3.4) and $\log p = o(n^{1/3})$, for any $x \in \mathbb{R}$ and $\omega \in (0,1)$,*

$$P\big(a_p\mathcal{T}_n - b_p(\omega) \le x\big) \to \exp\{-\exp(-x)\} \quad \text{as } n \to \infty,$$

where $a_p = \{2\log(\log(p))\}^{1/2}$ and $b_p(\omega) = 2\log(\log(p)) + 2^{-1}\log(\log(\log(p))) + \log(1-\omega) - 2^{-1}\log(4\pi)$.

Based on Theorem 2, a level α multi-level thresholding test rejects H_0 in (2.1) if $\mathcal{T}_n > a_p^{-1}(g_\alpha + b_p(\omega))$, where g_α is the upper α quantile of the Gumbel distribution $\exp\{-\exp(-x)\}$. We choose ω small to obtain good power for the proposed test.

In practice, we may choose $\hat{\mu}(s) = \mu_0(s)$ in the formulation of $\mathcal{T}_n$. In the case that $\mu_0(s)$ does not satisfy (3.4), we have to (i) restrict the relationship between n and p such that $n \sim p^\xi$ for a $\xi \in (0,1)$; (ii) modify the multi-level thresholding statistic by restricting $\mathcal{S}_n(\omega)$ such that $s_j > 1-\xi$ for all j and choosing ω small enough such that $\omega < \xi$. It can be shown that Theorem 2 is still valid with $b_p(\omega)$ replaced by $b_p(1+\omega-\xi)$. In this case, the multi-level thresholding test rejects H_0 in (2.1) if $\mathcal{T}_n > a_p^{-1}(g_\alpha + b_p(1+\omega-\xi))$.

To study the power of the proposed test against the sparse and weak hypothesis in (2.2), we consider a general setting which allows the response distributions, parameters and signal strength to vary across dimensions. In the following we use c_0 to denote a small positive constant.

B1. There are H (a positive integer) possible families of distributions for the responses. Specifically, a $\tau_h \ge c_0$ proportion out of the total p responses are distributed according to a distribution family with density $f_{(h)}(y_{ij}; z_i, \theta_j)$ that satisfies Assumptions A1 and A2 for $h = 1, \cdots, H$.

B2. The parameters under H_0, $\{(\beta'_{j,0}, \phi_j)'\}_{j=1}^p$, are i.i.d. copies from an $(m+1)$-dimensional super population with a density function q_1 which is compactly supported on a set $\mathcal{K} \subset \mathbf{R}^{m+1}$ such that $q_1(\theta) \ge c_0$ and $D\beta = 0$ for any $\theta = (\beta', \phi)' \in \mathcal{K}$;

B3. The signal strengths in $\{r_j\}$ are independently drawn from a super population with a density function q_2 which is compactly supported on a set $\mathcal{G}$, where $Dr \neq 0$ and $q_2(r) \geq c_0$ for any $r \in \mathcal{G}$.

Let $I_{\theta,h,\infty}(\cdot) = -\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{E} \frac{\partial^2}{\partial\theta\partial\theta'} \log f_{(h)}(y_{ij}; z_i, \cdot)$. Let $I^{-1}_{h,\infty}(\theta)$ be the upper $m \times m$ block of $I^{-1}_{\theta,h,\infty}(\theta)$. Suppose the jth response is from the h_jth family of distributions. Define the standardized signal strength

$$\tilde{r}_{h_j}(r_j, \theta_j) = r_j' D' V^{-1}_{h_j,\infty}(\theta_j) D r_j, \tag{4.7}$$

where $V_{h_j,\infty}(\theta_j) = D I^{-1}_{h_j,\infty}(\theta_j) D'$, and for $\mathcal{H} = \{1, \cdots, H\}$, let

$$\tilde{r} = \max_{h\in\mathcal{H}, r\in\mathcal{G}, \theta\in\mathcal{K}} \tilde{r}_h(r, \theta) \tag{4.8}$$

be the maximal standardized signal strength. Unlike the setting of the means in (4.1), where the signal strength is solely determined by r, both $\tilde{r}_{h_j}(r_j, \theta_j)$ and $\tilde{r}$ depend on both r_j and θ_j.

As we will demonstrate shortly, the power of the multi-level thresholding test is critically determined by $\tilde{r}$. Although it may seem strange for an L_2-type test's power to depend on the maximal signal strength, this connection with $\tilde{r}$ is due to the thresholding step we have augmented to the L_2 formulation to make the procedure adaptive to weak and faint signals. It should be noted that this maximal signal is not "isolated". Indeed, under the alternative hypothesis of (2.2), due to the compact support and bounded density ($q_1(\theta) \geq c_0$ and $q_2(r) \geq c_0$) conditions in B1-B3, there will be a $c_\epsilon > 0$ proportion of the signal-bearing responses with signal strength larger than $\tilde{r} - \epsilon$ for any small $\epsilon > 0$. This cluster of the responses around $\tilde{r}$ determines the power of the proposed test as revealed in the following theorem.

Theorem 3. *Under H_a in (2.2), A1-A3, B1-B3, (3.4), $\log p = o(n^{1/3})$ and ω small enough, for a series of slowly varying type I error rates converging to 0 as $n \to \infty$, with probability approaching 1,*

(i) if $\tilde{r} < \mathrm{DB}(\kappa)$, the power of the multi-level thresholding test $\to 0$;

(ii) if $\tilde{r} > \mathrm{DB}(\kappa)$, the power of the multi-level thresholding test $\to 1$.

The theorem indicates that $\mathrm{DB}(\kappa)$ is the detection boundary of the multi-level thresholding test. This detection boundary cannot be achieved by the standard L_2 tests, for instance that in Chen and Qin (2010), due to the lack of a thresholding component to screen out dimensions bearing no signal. Thresholding retains the most informative part of the signal while removing the non-informative dimensions.

The optimality of the detection boundary $\mathrm{DB}(\kappa)$ can be established when we confine to the linear regression with normally distributed response. Consider the linear regression model, for $i = 1, \cdots, n$ and $j = 1, \cdots, p$,

$$y_{ij} = z_i'\beta_j + \varepsilon_{ij} \quad \text{for} \quad \varepsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2). \tag{4.9}$$

Under this model, $\tilde{r} = \max_{r \in \mathcal{G}} \lim_{n\to\infty} r'D'\{D(Z'Z)^{-1}D'\}^{-1}Dr/(n\sigma^2)$ for $Z = (z_1, \cdots, z_n)'$.

Theorem 4. *Assume the responses for each observation are independent. For the hypothesis (2.2), under B2, B3 and the linear model (4.9), if* $\tilde{r} < \mathrm{DB}(\kappa)$, *Type I Error + Type II Error* $\to 1$ *for any test as* $n, p \to \infty$.

Theorem 4 shows that $\mathrm{DB}(\kappa)$ is the detection lower boundary of any test for the hypothesis (2.2) under the linear model (4.9) and independence among responses. From Theorem 3, we see that the proposed test obtains the optimal detection boundary under the conditions of Theorem 4. Although the assumption of independent responses is used here for deriving the optimal detection boundary and showing the optimality of the proposed test, the proposed test is still valid without the independence assumption (see A3). We would also like to point out that the optimal detection boundary under the case of dependent responses could be lower than the one given in Theorem 4, as discovered in Hall and Jin (2010) for testing high-dimensional means. We believe that the optimality under this dependent case could be achieved by the proposed test by implementing a data transformation first.

Arias-Castro et al. (2011) showed that $\mathrm{DB}(\kappa)$ is the optimal detection boundary for the linear regression model with high-dimensional covariates but low dimensional (univariate) response. The model we consider in (4.9) has low-dimensional covariates but high-dimensional responses.

For non-Gaussian distributions that satisfy A1 and A2, the following Theorem 5 indicates that $\mathrm{DB}(\kappa)$ is a detection lower boundary in the sense of (4.3). To formulate the statement, we define a new quantity r_0 that reflects the discrepancy between β_j and 0. Recall that $f_{h_j}(y_{ij}; z_i, \theta_j)$ is the density of the jth response and $\beta_{j,a} = r_j\sqrt{2(\log p)/n}$ for $r_j \in \mathcal{G}$. Let

$$\begin{aligned} I_{\beta,h_j,\infty}(\cdot) &= -\lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{E} \frac{\partial^2}{\partial\beta\partial\beta'} \log f_{h_j}(y_{ij}; z_i, \cdot)/n \quad \text{and} \\ r_0 &= \max_{h\in\mathcal{H}, r\in\mathcal{G}, \theta\in\mathcal{K}} r' I_{\beta,h,\infty}(\theta) r. \end{aligned} \tag{4.10}$$

Note that $\tilde{r}$ measures the signal strength of the targeting linear combinations $D\beta_j$, while r_0 is the squared standardized distance of r_j from 0. The latter

may involve nuisance parameters which are not of our interest. Since D has full row rank, and $I_{\beta,h,\infty}(\theta)$ and $I^{-1}_{h,\infty}(\theta)$ are the upper $m \times m$ sub-matrices of $I_{\theta,h,\infty}(\theta)$ and $I^{-1}_{\theta,h,\infty}(\theta)$, respectively, it can be shown that $r_0 \geq \tilde{r}$ under any distribution of the response.

Theorem 5. *Assume the responses for each observation are independent. Under A1, A2, and B1-B3, for the hypothesis (2.2), if* $r_0 < \mathrm{DB}(\kappa)$, *Type I Error + Type II Error* $\to 1$ *for any test as* $n, p \to \infty$.

Since $r_0 \geq \tilde{r}$, the undetectable region $r_0 < \mathrm{DB}(\kappa)$ given in Theorem 5 for any distribution is smaller than that given in Theorem 4 written in terms of $\tilde{r}$, which is specifically for the linear model (4.9). When the dispersion parameter ϕ_j is known and D is the identity matrix, we have $I_{\beta,h,\infty}(\theta) = I_{h,\infty}(\theta)$ and $r_0 = \tilde{r}$. Under this form of the simple null hypothesis, it can be shown that the multi-level thresholding test attains the detection lower boundary. Hence, it is optimal under this scenario. However, for a general composite hypothesis (2.2), the proposed test may not attain this lower detection boundary written in terms of r_0, since there may not exist a simple hypothesis equivalent to (2.2) under the non-Gaussian case.

5. Signal Identification. If hypothesis (2.1) is rejected, we are interested in locating the dimensions of signals. This is equivalent to considering

$$H_{j,0} : D\beta_j = 0 \quad \text{vs.} \quad H_{j,a} : D\beta_j \neq 0, \tag{5.1}$$

for $j = 1, \cdots, p$, and identifying the dimensions with $D\beta_j \neq 0$. Let p_0 be the number of true null hypotheses. For sparse signals, p_0 is close to p. Let V and R be the numbers of false positives and rejected null hypotheses, respectively. The false discovery proportion $\mathrm{FDP} = V/\max\{R, 1\}$ is the proportion of falsely rejected null hypotheses among all rejected null hypotheses, and the false discovery rate (FDR) is the expectation of the FDP.

Benjamini and Hochberg (1995) (BH) considered FDR control at a level $\alpha \in (0, 1)$ for (5.1) under dimension-wise independence. For each dimension j, let $p_j = P(\mathcal{X}^2_d > n(D\hat{\beta}_j)'\hat{V}_j^{-1}(D\hat{\beta}_j))$ be the p-value for testing $H_{j,0}$ based on the Wald test, where $\mathcal{X}^2_d$ denotes the chi-square distribution with d degrees of freedom. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(p)}$ be the ordered p-values, and $\pi(j)$ be the dimension label of the jth smallest p-value. BH's procedure rejects $H_{\pi(1),0}, \cdots, H_{\pi(M),0}$ in (5.1) for $M = \max\{j : p_{(j)} \leq \alpha j/p\}$.

Controlling FDR bounds the expected FDP over repeated experiments. Genovese and Wasserman (2006) suggested controlling the probability that

FDP exceeds a specific value, that is, to control $P(\text{FDP} > c) \le \alpha$ for a given c in $(0, 1)$, For each subset $W \subset \{1, \cdots, p\}$, they considered testing

$$H_{W,0} : D\beta_l = 0 \text{ for all } l \in W \quad \text{vs.} \quad H_{W,a} : D\beta_l \neq 0 \text{ for some } l \in W \tag{5.2}$$

at level α. Let $\mathcal{U}$ be the collection of all subsets W not rejected in (5.2). For any subset $A \subset \{1, \cdots, p\}$, they define $\bar{\Gamma}(A) = \max_{B \in \mathcal{U}} \frac{\#(B \cap A)}{\#(A)}$ to be a $1 - \alpha$ confidence envelope for FDP, where $\#(A)$ is the cardinality of A. Then, choose the rejection set R_0 such that $\bar{\Gamma}(R_0) \le c$ to control the FDP exceedance rate. Genovese and Wasserman (2006) proposed to test the overall hypothesis (5.2) via the minimum p-value test (GW1). Extensions to tests based on the kth smallest p-value and an approach to combine results from different k (GWcom) were proposed. See Sections 4 and 6 of Genovese and Wasserman (2006) for details.

Testing all the subsets of $\{1, \cdots, p\}$ is computational infeasible when p is large. Most importantly, by linking the results given in Donoho and Jin (2004) and Theorem 3, the test based on the minimum p-value cannot attain the optimal detection boundary $\text{DB}(\kappa)$. To translate the good detection property of the multi-level thresholding test to better signal detection, we apply the proposed test for the overall hypotheses (5.2) in a step-down formulation. Specifically, let $W_j = \{\pi(j), \pi(j+1), \cdots, \pi(p)\}$ for $j = 1, \cdots, p$. Consider testing at level α the hypothesis

$$H_{W_j,0} : D\beta_l = 0 \text{ for all } l \in W_j \quad \text{vs.} \quad H_{W_j,a} : D\beta_l \neq 0 \text{ for some } l \in W_j \tag{5.3}$$

The sequence of the tests in (5.3) serves as a step-down procedure for (5.1).

Let $\mathcal{T}(W_j)$ be the multi-level thresholding statistic computed using data in W_j. The following is the proposed multiple testing procedure for (5.1).

(i) Step-down: define $J = \min\{j : \mathcal{T}(W_j) \le a^{-1}_{p-j+1}(g_\alpha + b_{p-j+1}(\omega))\}$ or $p + 1$ if $\mathcal{T}(W_j) > a^{-1}_{p-j+1}(g_\alpha + b_{p-j+1}(\omega))$ for all $j = 1, \cdots, p$. Let $R_1 = \{\pi(1), \cdots, \pi(J-1)\}$ or the empty set if $J = 1$.

(ii) Augmentation: let $J^* = \min\{p, \lfloor (J-1)/(1-c) \rfloor\}$, $R^* = \{\pi(1), \cdots, \pi(J^*)\}$ or the empty set if $J^* = 0$.

(iii) Rejection set: our proposed procedure rejects the null hypothesis in (5.1) for all j in R^*.

Part (i) is a step-down procedure via the thresholding statistic (4.6) for the hypotheses (5.3). Essentially, J is obtained by repeatedly conducting the multi-level thresholding test on W_j while removing the most significant individual dimension one at a time until there is no rejection. Part (ii) is the augmentation step. Following Genovese and Wasserman (2006), the rejection

set R^* is obtained by augmenting R_1 from part (i) with the next $\lfloor (J-1)c/(1-c) \rfloor$ most significant dimensions whenever R_1 is nonempty.

The rationale for enlarging R_1 is that if we only choose R_1 as the set of signals, the FDP rate would diminish to zero with probability $1-\alpha$ as the number of signals increases with n and p. Augmenting the rejection set with the next $\lfloor (J-1)c/(1-c) \rfloor$ most significant dimensions increases power while still asymptotically controlling the rate of FDP exceeding c at level α.

Comparing to the BH procedure that only controls average FDP, our procedure can control a more stringent type I error rate without power loss. Comparing to the GW procedure, since the multi-level thresholding test is more powerful than the minimum p-value test in detecting rare and faint signals as confirmed in Figure 3, the proposed signal identification procedure enjoys higher power than the GW procedure. We will demonstrate those advantages of the proposed procedure by the following two theorems.

Let $S = R - V$ be the number of correctly discovered signals (true positives), and let S_{GW}, S_{BH} and S_{prop} be S for the GW, BH and the proposed procedures, respectively. Recall that $p - p_0$ is the total number of signals. The following theorem compares the ratios of signal selection of the proposed procedure with that of the GW and BH procedures. To simplify the presentation, we assume the standardized signal strength $\tilde{r}_{h_j}(r_j, \theta_j)$ in (4.7) is the same for all the responses in the following theorems.

Theorem 6. *Under H_a in (2.2) with $\tilde{r} > \kappa$, Conditions A1-A3 and $\log p = o(n^{1/3})$, as $\alpha \to 0$ slowly and $n, p \to \infty$,*

(i) $S_{GW}/(p-p_0) \xrightarrow{p} 0$ when $\tilde{r} < 1$;

(ii) $S_{GW}/(p-p_0) \xrightarrow{p} 1$ at the rate $p^{-(\sqrt{\tilde{r}}-\sqrt{\kappa})^2+o(1)}$ when $\tilde{r} > 1$;

(iii) $S_{prop}/(p-p_0)$ and $S_{BH}/(p-p_0)$ converge to 1 in probability at the rate $p^{-(\sqrt{\tilde{r}}-\sqrt{\kappa})^2+o(1)}$.

Theorem 6 shows that the signal identification by employing the proposed approach and the BH procedure is selection consistent as long as $\tilde{r} > \kappa$, whereas the GW procedure is selection consistent only for the strong signal case of $\tilde{r} > 1$. It is noted that $\tilde{r} > \kappa$ is a minimum requirement for identifying rare and faint signal since Ji and Jin (2012) discovered that the signal identification is impossible if $\tilde{r} < \kappa$. When the signal strength is stronger such that $\tilde{r} > 1$, all the three procedures attain the signal selection consistency with the same rate of convergence up to a factor of $p^{o(1)}$. While Theorem 6 shows that the true positive rates of the proposed procedure is comparable to that of the BH procedure, the following Theorem shows that the proposed procedure can control the FDP exceedance rate, which is more

stringent than the FDR control achieved by the BH procedure.

Theorem 7. *Under H_a in (2.2) with $\tilde{r} > \kappa$, Conditions A1-A3 and $\log p = o(n^{1/3})$, as $n, p \to \infty$, the proposed multiple testing procedure controls the FDP exceedance rate such that $P(\text{FDP} > c) \leq \alpha$.*

It is noted that the FDR based procedure (BH method) controls the average of FDP without considering its variation. Hence, it may not be suitable in some applications as pointed out by Genovese and Wasserman (2006). The proposed procedure provides a method for incorporating FDP variability under control without sacrificing the power in terms of signal selection consistency. Simulation studies reported in the next section confirm that the proposed procedure can control both FDR and the exceedance FDP rate, and outperform both the BH and GW procedures.

6. Simulation study. We studied the empirical performance of the proposed test under generalized linear models. Balanced designs with two treatments were considered. To mimic the "large p, small n" paradigm, we chose the total sample size $n = 20$ and 40, where the sub-sample sizes of each treatment group are 10 and 20, respectively. The dimension was chosen as $p = 100, 400, 1000$ and 10000. Three models were used to simulate data. In each model, the covariate vectors (z_i) take values $(1, 0)'$ or $(0, 1)'$, indicating the first and second treatment, respectively. The models are as follows:

- Poisson regression. For $i = 1, \cdots, n$ and $j = 1, \cdots, p$, the response y_{ij} follows Poisson distribution with mean $\mu_{ij} = \exp(z_i'\beta_j)$.
- Binomial regression. Suppose $y_{ij} \sim$binomial(n_{ij}, p_{ij}), where n_{ij} were randomly chosen from the integers between 20 and 40 according to a discrete uniform distribution, and $p_{ij} = \exp(z_i'\beta_j)/\{\exp(z_i'\beta_j) + 1\}$.
- Negative binomial regression. The response y_{ij} is generated from NB(μ_{ij}, ϕ_j) with $\log(\mu_{ij}) = z_i'\beta_j$, where the dispersion parameters $\phi_1, \cdots, \phi_p$ were set according to *i.i.d.* draws from the uniform$(3, 13)$ distribution.

We tested whether there are treatment effects for any of the response variables. Namely, with the D matrix in (2.2) equal to $(1, -1)$, we have

$$H_0 : \beta_{j1} = \beta_{j2} \text{ for all } j \quad \text{vs.} \quad H_a : \beta_{j2} \overset{i.i.d.}{\sim} (1-\epsilon)\nu_{\beta_{j1}} + \epsilon\nu_{\beta_{j1}+\beta_a}, \tag{6.1}$$

where $\epsilon = p^{-\kappa}$ and $\beta_a = \sqrt{(2r_a \log p)/n}$. Under H_0, $\beta_{j1} = \beta_{j2} = \beta_0$ for all response variables, where $\beta_0 = 2$ for Poisson regression, 2.5 for negative binomial regression and 0.5 for binomial regression.

We chose $\kappa = 0.6$ and 0.55 representing the sparsity of signals. The numbers of signals were kept at $7, 11, 16$ and 40 corresponding to $p =$

$100, 400, 1000$ and 10000 for $\kappa = 0.6$, and $8, 15, 22$ and 63 corresponding to $p = 100, 400, 1000$ and 10000 for $\kappa = 0.55$, respectively. The strength parameter r_a was chosen differently between different models to make the standardized signal strength within $(0, 1)$. We estimated $\mathrm{E}\{T_n(s)|H_0\}$ by its main order $\mu_0(s)$ given in (3.2), and set $\omega = 0.1$. The nominal size was 5%. All the simulation results reported below are based on 1000 replications.

For negative binomial regression, the MLE $\hat{\phi}_j$ usually overestimates ϕ_j when the sample size is small, leading to under-estimation of the standard deviation of $\hat{\beta}_j$. This enlarges the thresholding statistic and causes a size distortion. We use a parametric bootstrap to correct the bias of $\hat{\phi}_j$. Specifically, for each response variable, the MLEs $\hat{\beta}_j$ and $\hat{\phi}_j$ are first obtained based on the original sample. Bootstrap resamples of size n are drawn from the negative binomial model with parameters $\hat{\beta}_j$ and $\hat{\phi}_j$, and ϕ_j is re-estimated based on the resample. The process is repeated B times to obtain the bootstrapped MLEs $\hat{\phi}^*_{j,1}, \cdots, \hat{\phi}^*_{j,B}$. The bias corrected estimator is $\tilde{\phi}_j = 2\hat{\phi}_j - \bar{\phi}^*_j$, where $\bar{\phi}^*_j = \sum_{i=1}^B \hat{\phi}^*_{j,i}/B$. We use $\tilde{\phi}_j$ to approximate the Fisher information of β_j and to compute the thresholding statistics $T_n(s)$ in (3.1).

The empirical size and power of the multi-level thresholding test are displayed in Figure 1. We observe that the proposed test had reasonable size around the nominal level 5% in most cases. The sizes for Poisson and negative binomial regression were slightly conservative under $n = 20$. When n was increased to 40, the sizes increased to around 5%. The powers of the proposed test were satisfactory under all the scenarios and increased rapidly with the increase of signal strength and number of signals. There were dips in the power between the fourth and fifth index, which were due to the decrease in the signal strength r_a that was not compensated by the increase in the number of signals. The simulation setting provides one example where fewer large signals are easier to detect than many small signals.

To better understand the different performances of the proposed test under the three models, we provide in Table 1 the value of r_a that defines $\beta_a = \sqrt{(2r_a \log p)/n}$ and the corresponding maximal standardized signal strength $\tilde{r}$ in (4.8). It shows that the negative binomial regression has the highest $\tilde{r}$ among the three models, which is due to the fact that the r_a for the negative binomial regression is larger than those in the Poisson and binomial regression. Having the largest $\tilde{r}$ is the reason why the empirical power was higher for negative binomial regression than for Poisson or binomial regression. It is observed that the empirical power is not very responsive to the changing sample size, but is sensitive to the dimensionality p. This is because, as shown in Proposition S1 in the supplementary material, the signal to noise ratio (SNR) of the thresholding test is largely determined by

p, κ and $\tilde{r}$. And the SNR increases as p increases as long as $\tilde{r} > \mathrm{DB}(\kappa)$.

TABLE 1
The maximal standardized signal strength $\tilde{r}$ at different r_a for the three models.

Poisson $\beta = (2,2)'$		Binomial $\beta = (0.5,0.5)'$		Negative binomial $\beta = (2.5,2.5)'$	
r_a	$\tilde{r}$	r_a	$\tilde{r}$	r_a	$\tilde{r}$
0.15	0.28	0.15	0.29	0.4	0.63
0.2	0.37	0.22	0.42	0.5	0.78
0.25	0.46	0.3	0.59	0.6	0.94

In addition to (6.1), we also considered scenarios motivated by the experiment described in Section 7. That experiment involves a Latin square design with two blocking factors (lanes and barcodes) and one treatment factor of interest (genotype). Suppose for $i = 1, \cdots, n$ and $j = 1, \cdots, p$,

$$g(\mu_{ij}) = \nu_j + X'_{g,i}\alpha_j + X'_{\ell,i}\tau_j + X'_{b,i}\gamma_j,$$

where $g(\cdot)$ is a link function, ν_j is an intercept parameter, $X_{g,i}$, $X_{\ell,i}$ and $X_{b,i}$ are vectors that indicates the genotype, lane and barcode of the ith experimental unit respectively, and $\alpha_j = (\alpha_{j1}, \alpha_{j2}, \alpha_{j3}, \alpha_{j4})'$, $\tau_j = (\tau_{j1}, \tau_{j2}, \tau_{j3}, \tau_{j4})'$ and $\gamma_j = (\gamma_{j1}, \gamma_{j2}, \gamma_{j3}, \gamma_{j4})'$ are vectors of genotype, lane and barcode effects, respectively. As discussed in Section 7, we are interested in testing the hypothesis $H_0 : \alpha_{j1} = \alpha_{j2} = \alpha_{j3} = 0$ for all j, where α_{j4} is set to zero for identifiability purposes. Recall that I_3 is the 3×3 identity matrix. The D matrix in (2.2) that corresponds to this hypothesis is $[0_{3\times 1}, I_3, 0_{3\times 3}, 0_{3\times 3}]$.

We consider Poisson, negative binomial and binomial regression with $n = 16$ (to match the sample size in the case study) and also $n = 32$, which doubles the number of observations for each combination of factors. The link function $g(\cdot)$ was set to log for Poisson and negative binomial cases and to logit for binomial regression. We set $\nu_j = 3.5, 2, 0.2$ for Poisson, negative binomial and binomial regression, respectively. We set $\tau_j = \gamma_j = (-0.5, 0, 0.5, 0)'$ for Poisson and negative binomial regression, and $\tau_j = \gamma_j = (-0.1, 0, 0.1, 0)'$ for binomial regression. In all our simulations, α_{j1}, α_{j2}, and α_{j3} were set to a common value denoted as α_{j0}. Under the null, α_{j0} was set to 0, and under the alternative, α_{j0} was generated according to

$$H_a : \alpha_{j0} \overset{i.i.d.}{\sim} (1-\epsilon)\nu_0 + \epsilon\nu_{\alpha_a}, \tag{6.2}$$

where $\alpha_a = \sqrt{(2r_a \log p)/n}$. The simulation results, reported in Figure 2, show that the multi-level thresholding test had reasonable size and good

power for detecting the alternative in (6.2). This shows that the proposed method works well for designed experiments more complex than the two-group comparisons covered in our other simulation scenarios.

To gain wider perspectives on our proposal, we compared the proposed test with two alternative formulations. One was a HC test in the spirit of Donoho and Jin (2004), which rejects H_0 in (6.1) if $HC^* > \sqrt{2\log\log p}$ where

$$HC^* = \max_{1\le j\le p/2} \sqrt{p}[j/p - p_{(j)}]/[p_{(j)}(1-p_{(j)})]^{1/2}.$$

The other was a test based on the minimum p-value, which reject H_0 in (6.1) if $p_{(1)} < B^{-1}_{1,p}(0.05) \approx 0.05/p$. This is equivalent to the test based on the max-norm statistics over all dimensions. We considered 30 and 40 signals. Figure 3 displays the powers of the three tests for Poisson regression with $p = 400$ and 1000. The results for other models and dimensions were similar. To make the power comparison fair, all the empirical sizes were adjusted to be 5%. It is observed that the proposed test had the best power among the three test procedures, and the power of the minimum p-value test was the lowest. In the context of testing for means, Donoho and Jin (2004) showed that the minimum p-value test is powerless for $\kappa \in (1/2, 3/4)$ and $\kappa - 1/2 < \tilde{r} < (1-\sqrt{1-\kappa})^2$, a region that lies above the optimal detection boundary of the proposed test. Similar detection boundary results for the minimum p-value test can be derived under our context, so the poorer power performance in our simulation is not surprising. The superior performance of the proposed test over the HC formulation suggests that the advantage of the L_2 formulation after thresholding as discovered in Zhong et al. (2014) for the mean parameter may be valid for general MLEs.

To compare the proposed multiple testing procedure with the procedures of Benjamini and Hochberg (1995) (BH) and Genovese and Wasserman (2006), we considered negative binomial regression under H_a in (6.1) with $\beta_{j1} = 2.5$, $n = 40$, $p = 10000$ and 4 different numbers of signals, 50, 100, 150 and 200. We adopt the GW procedure with $k = 1$ (GW1) and the combined k approach (GWcom), and set the FDP exceedance level $c = 0.1$ and the control rate $\alpha = 0.05$. Figure 4 shows the type I and type II errors of the four procedures. We see that the proposed procedure controlled the FDR and the exceedance FDP rate around 5% for all the cases. The BH procedure (which is designed to control FDR) was unable to control the exceedance FDP rate when the number of signals is 50. The non-discovery proportions of the proposed procedure were close to those of the BH procedure. However, the proposed procedure was more powerful when the signal strength is weak ($r_a = 0.8$). Although both the GW procedures controlled the exceedance

FDP rate, they were too conservative. Their type I error rates were around 0, which inevitably brought large type II errors.

7. Case study. In this section, we illustrate the proposed method in an analysis of maize RNA-seq data from Paschold et al. (2014). In an RNA-seq experiment, target mRNA molecules are first converted to cDNA fragments that are sequenced on a high-throughput next generation sequencing platform. Then, these sequences (known as reads) are aligned to a reference genome, and the number of reads mapped to a given gene measures its expression level. The data set we analyze consists of RNA-seq read counts from root cortex tissue of four maize genotypes with four replications per genotype. The four genotypes include two inbred parental lines (labeled B and M) as well as two hybrid genotypes formed by crossing B and M with B as the female parent and M as the male parent (BM) and vice versus (MB). Although these two reciprocal hybrids are genotypically indistinguishable, they may differ in some traits, including gene expression levels.

The four samples from any given block were sequenced together in a single Illumina flow cell lane. Four barcodes (AR001, AR003, AR008 and AR009) were used so that each read could be attributed to the correct sample within each lane. Table 2 illustrates the Latin square sequencing design employed to facilitate estimation of block/lane, barcode and genotype effects on gene expression levels.

TABLE 2
The Latin Square Sequencing Design for the Maize Study.

	Barcode			
	AR001	AR003	AR008	AR009
Block/lane 1	B	M	BM	MB
Block/lane 2	M	BM	MB	B
Block/lane 3	BM	MB	B	M
Block/lane 4	MB	B	M	BM

Consistent with standard practice in the analysis of RNA-seq data, we applied pre-screening to delete the genes with low read counts (average counts less than 10). For the jth gene included in the analysis, let $y_{1j}, \cdots, y_{16\,j}$ be the RNA-seq read counts. We assume y_{ij} follows a negative binomial distribution with dispersion parameter ϕ_j and mean μ_{ij} satisfying

$$\log(\mu_{ij}) = \nu_j + X'_{g,i}\alpha_j + X'_{\ell,i}\tau_j + X'_{b,i}\gamma_j,$$

where ν_j is an intercept parameter and $\alpha_j = (\alpha_{j1}, \alpha_{j2}, \alpha_{j3}, \alpha_{j4})'$, $\tau_j = (\tau_{j1}, \tau_{j2}, \tau_{j3}, \tau_{j4})'$ and $\gamma_j = (\gamma_{j1}, \gamma_{j2}, \gamma_{j3}, \gamma_{j4})'$ are the effects of genotypes,

blocks/lanes and barcodes for the jth gene. Without loss of generality, we set $\alpha_{j4} = \tau_{j4} = \gamma_{j4} = 0$ for identifiability purposes.

We begin our analysis by testing whether any gene is differentially expressed across genotypes. The relevant hypotheses are

$$\begin{aligned} &H_0 : \alpha_{j1} = \alpha_{j2} = \alpha_{j3} = 0 \ \text{ for all } j \ \text{ vs.} \\ &H_a : \text{at least one component of } \alpha_j \text{ is not equal to 0 for some } j. \end{aligned} \tag{7.1}$$

Let $\beta_j = (\nu_j, \alpha_{j1}, \alpha_{j2}, \alpha_{j3}, \tau_{j1}, \tau_{j2}, \tau_{j3}, \gamma_{j1}, \gamma_{j2}, \gamma_{j3})'$. As in the second setting of the simulation study, $D = [0_{3\times 1}, I_3, 0_{3\times 3}, 0_{3\times 3}]$.

We applied the proposed test for the hypotheses in (7.1). The value of the multi-level thresholding statistic in (4.6) was 1012.3. At the 5% significant level, we reject the null hypothesis when this statistic is larger than 3.08. Therefore, the proposed method provides a clear indication that the null hypothesis of (7.1) should be rejected. Next, we test whether any gene is differentially expressed between the reciprocal hybrid genotypes BM and MB. The value of the multi-level thresholding statistic was 37.83, which exceeds the critical value 3.08. Thus, there is evidence that some genes are differentially expressed between the reciprocal hybrids BM and MB.

We also applied the proposed multiple testing procedure to identify genes differentially expressed (DE) between the hybrids. We controlled at 5% the probability of FDP in excess of 0.1. The proposed method identified 32 DE genes between the reciprocal hybrids, while the BH procedure found only 23 of the 32 DE genes. Both the GW procedures based on the minimum p-value and the combined k approach found just 18 of the 23 identified by the BH procedure. These results are consistent with the findings of Theorem 6 and the simulation study: our proposed method tends to identify more genes as DE than does the GW approach, and it does not suffer power loss compared to the BH approach, while controlling the FDP exceedance rate.

8. Extension. The proposed multi-level thresholding test can be extended to more complicated scenarios. We discuss two possible extensions in this section: generalized linear mixed models (GLMMs) and high-dimensional predictors.

GLMM. We focus on a group random effects model. Suppose that the $n = u \times v$ observations come from u groups with v observations in each group. Let y_{ijk} be the value of the kth observation of the jth response variable in the ith group, and let z_{ik} be the corresponding vector of explanatory variables, where $i = 1, \cdots, u$, $j = 1, \cdots, p$ and $k = 1, \cdots, v$. Let $\eta_j = (\eta_{1j}, \cdots, \eta_{uj})'$ and $\beta_j = (\beta_{j1}, \cdots, \beta_{jm})'$ be random group effects and fixed treatment effects, respectively. Let $\mu_{ijk} = \mathrm{E}(y_{ijk}|\eta_{ij})$. For the link function

$g(\cdot)$ and an unknown positive variance component σ_j^2,

$$g(\mu_{ijk}) = z'_{ik}\beta_j + \eta_{ij} \quad \text{for} \quad \eta_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_j^2). \tag{8.1}$$

Define $\theta_j = (\beta'_j, \phi_j, \sigma_j^2)'$, $\widetilde{Y}_{ij} = (y_{ij1}, \cdots, y_{ijv})'$ and $Z_i = (z'_{i1}, \cdots, z'_{iv})'$. Let $\widetilde{Y}_j = (\widetilde{Y}'_{1j}, \cdots, \widetilde{Y}'_{uj})'$. For the jth response variable in the ith group, the marginal probability density function of $\widetilde{Y}_{ij}$ is

$$f_j(w_1, \cdots, w_v; Z_i, \theta_j) = \int \prod_{k=1}^{v} p_{ijk}(w_k|t; z'_{ik}, \beta_j, \phi_j)\phi_{\sigma_j^2}(t)dt, \tag{8.2}$$

where $p_{ijk}(\cdot|t; z'_{ik}, \beta_j, \phi_j)$ is the conditional density of y_{ijk} given $\eta_{ij} = t$, and $\phi_{\sigma_j^2}(t)$ is the $N(0, \sigma_j^2)$ density. Due to the random group effects, observations are independent between groups but dependent within groups. Let $\tilde{\theta}_j$ be the exact MLE. However, due to intractable integration in (8.2) for $f_j(\widetilde{Y}_{ij}; Z_i, \theta_j)$, $\tilde{\theta}_j$ may be unobtainable.

As the group random effects η_{ij} are Gaussian, Gauss-Hermite quadrature can be used to approximate $f_j(\widetilde{Y}_{ij}; Z_i, \theta_j)$. Let $\hat{f}_{j,G}(\widetilde{Y}_{ij}; Z_i, \theta_j)$ be its approximation by the Gauss-Hermite quadrature of degree G. The exact MLE $\tilde{\theta}_j$ can be approximated by $\hat{\theta}_{j,G}$, which maximizes the approximate log likelihood $\sum_{i=1}^{u} \log\{\hat{f}_{j,G}(\widetilde{Y}_{ij}; Z_i, \theta_j)\}$ (McCulloch et al., 2008). Since the approximation error of $\hat{f}_{j,G}(\widetilde{Y}_{ij}; Z_i, \theta_j)$ to $f_j(\widetilde{Y}_{ij}; Z_i, \theta_j)$ can be controlled by the quadrature degree G, a moderate deviation result similar to Lemma 1 could also hold for the approximate MLE $\hat{\theta}_{j,G}$ with a carefully chosen G. This indicates that the thresholding test procedure could be applied in conjunction with the Gaussian quadrature approximation method.

High dimensional predictors. The proposed procedure could be applied to the case of diverging number of predictors, namely, allowing $m \to \infty$. We illustrate the idea via the linear regression. For $i = 1, \cdots, n$ and $j = 1, \cdots, p$,

$$y_{ij} = z'_{ij}\beta_j + \epsilon_{ij} \quad \text{for} \quad \epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2), \tag{8.3}$$

where $z_{ij} = (z_{ij1}, \cdots, z_{ijm})'$, $\beta_j = (\beta_{j1}, \cdots, \beta_{jm})'$, and both the covariates and responses are high-dimensional such that $m \gg n$ and $p \gg n$. For some covariate k, a hypothesis of interest is

$$H_0 : \beta_{jk} = 0 \ \text{ for all } j \ \text{ vs. } \ H_a : \beta_{jk} \neq 0 \ \text{ for some } j. \tag{8.4}$$

Let $\widetilde{Y}_j = (y_{1j}, \cdots, y_{nj})'$ and $Z_j = (z_{1j}, \cdots, z_{nj})'$. For each response, we estimate β_j by the desparsified Lasso estimator

$$\hat{b}_j = \hat{\beta}_j + \hat{\Theta}_j Z'_j(\widetilde{Y}_j - Z_j\hat{\beta}_j)/n$$

of Zhang and Zhang (2014) and van de Geer et al. (2014), where $\hat{b}_j = (\hat{b}_{j1}, \cdots, \hat{b}_{jm})'$, $\hat{\beta}_j$ is the Lasso estimator, and $\hat{\Theta}_j$ is from the node-wise regression of each covariate in the design matrix Z_j on all other covariates. See equation (7) and (8) of van de Geer et al. (2014). By Theorem 2.1 of van de Geer et al. (2014), the moderate deviation result for the desparsified Lasso estimators $\hat{b}_j$ similar to Lemma 1 could be established under some suitable conditions. Based on this, the proposed procedures for signal detection and identification could be applied on $\hat{b}_{jk}$ for the hypothesis (8.4).

Appendix. Here we provide the proof of Theorem 4, which is the key in the detection boundary analyzes. Proofs of the other theorems are given in the supplementary material.

Proof of Theorem 4. Consider the hypotheses (2.2) under the linear model

$$y_{ij} = z_i'\beta_j + \varepsilon_{ij} \quad \text{for} \quad \varepsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2), \ i = 1, \cdots, n, \ j = 1, \cdots, p.$$

Let $Z = (z_1, \cdots, z_n)'$, and $C_Z \subset \mathbf{R}^n$ be its column space with dimension m. Let $\mu_{ij} = \mathrm{E}(y_{ij})$. Then, for each j, the jth response mean, $(\mu_{1j}, \cdots, \mu_{nj})'$, is in C_Z. Let $N_D \subset \mathbf{R}^m$ be the null space of D. Because D has d linearly independent rows, the dimension of N_D is $m-d$. Let E be an $m \times (m-d)$ matrix whose column space is N_D. Under H_0 in (2.2), we see that $(\mu_{1j}, \cdots, \mu_{nj})'$ is contained in the column space of ZE for each $j = 1, \cdots, p$.

Following the argument for linear hypotheses in page 266 of Lehmann (1959), we would like to construct an $n \times n$ orthogonal matrix G in such a way that the first m rows of G span C_Z with the $(d+1)$th row to the mth row spanning the column space of ZE. Transform the responses by G. Let $(y^*_{1j}, \cdots, y^*_{nj})' = G(y_{1j}, \cdots, y_{nj})'$ and $\eta_{ij} = \mathrm{E}(y^*_{ij})$ for $i = 1, \cdots, n$ and $j = 1, \cdots, p$. Then, testing $D\beta_j = 0$ is equivalent to testing $\eta_{1j} = \cdots = \eta_{dj} = 0$ for each j.

Let $A = \sigma^{-1}\{D(Z'Z)^{-1}D'\}^{-1/2}D(Z'Z)^{-1}Z'$. Note that $AA' = \sigma^{-2}I_{d\times d}$. It can be shown that the first d rows of G can be chosen as σA. Then, the first d transformed responses under the linear model are

$$y^*_{ij} = B_i\beta_j + \varepsilon^*_{ij} \quad \text{for} \quad \varepsilon^*_{ij} \overset{i.i.d.}{\sim} N(0, 1),$$

where B_i is the ith row of $B = \sigma^{-1}\{D(Z'Z)^{-1}D'\}^{-1/2}D$ for $i = 1, \cdots, d$. Let $r_t = Br/\sqrt{n}$. The hypotheses (2.2) are equivalent to

$$\begin{aligned} &H_0 : \eta_{1j} = \cdots = \eta_{dj} = 0 \ \text{ for } j = 1, \cdots, p \ \text{ vs.} \\ &H_a : \eta_{ij} \overset{ind}{\sim} (1-\epsilon)\nu_0 + \epsilon\nu_{a_i} \ \text{ for } i = 1, \cdots, d \text{ and } j = 1, \cdots, p, \end{aligned} \tag{A.1}$$

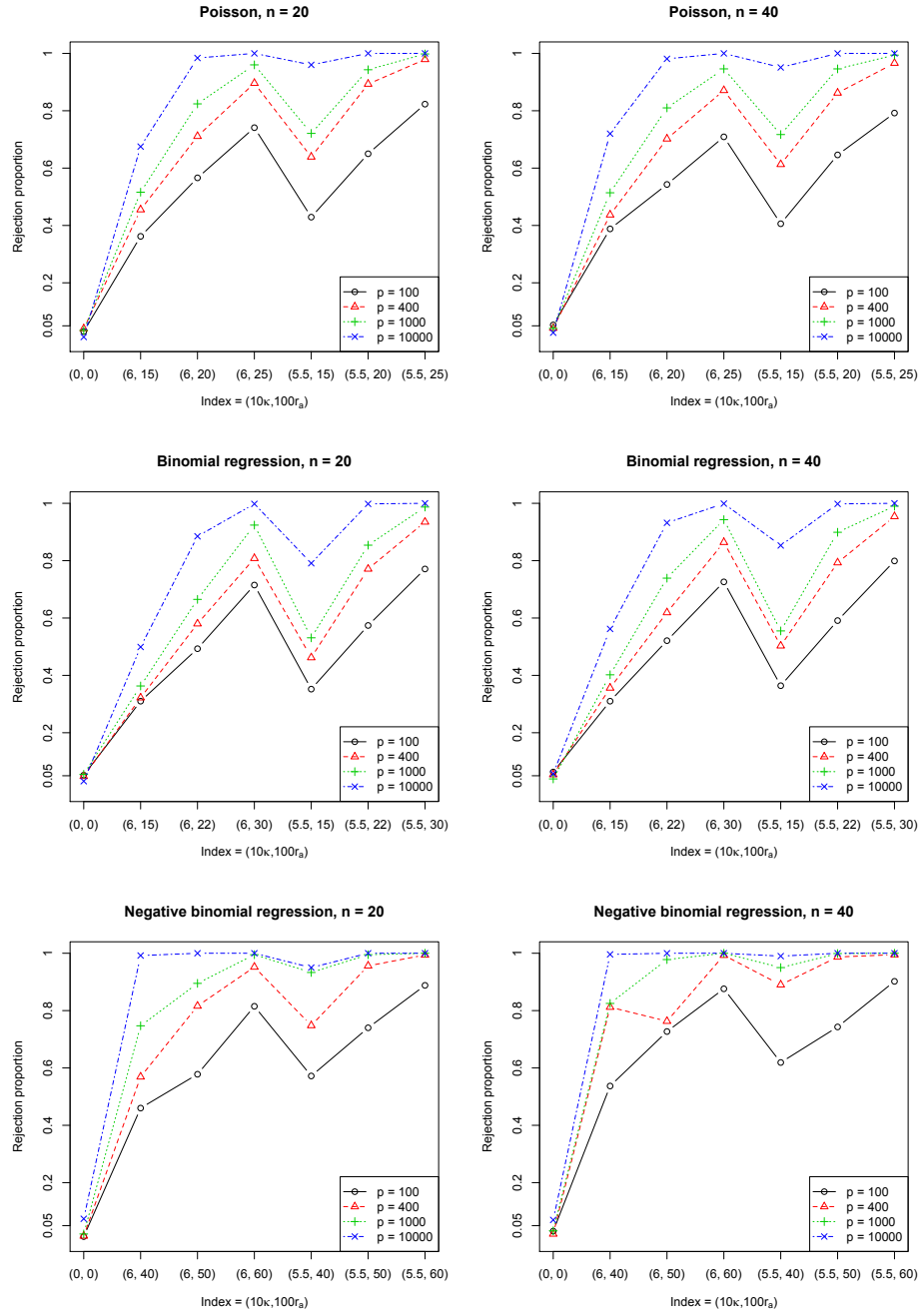

Fig 1: Empirical sizes and powers of the multi-level thresholding test for the hypothesis (6.1) under Poisson, binomial and negative binomial regression. The vertical axis shows the proportion of rejections. The horizontal axis gives the null hypothesis, represented by (0, 0), and six alternative hypotheses. The first and second index of the horizontal axis give the values of 10κ and $100r_a$, respectively, providing signal sparsity and strength.

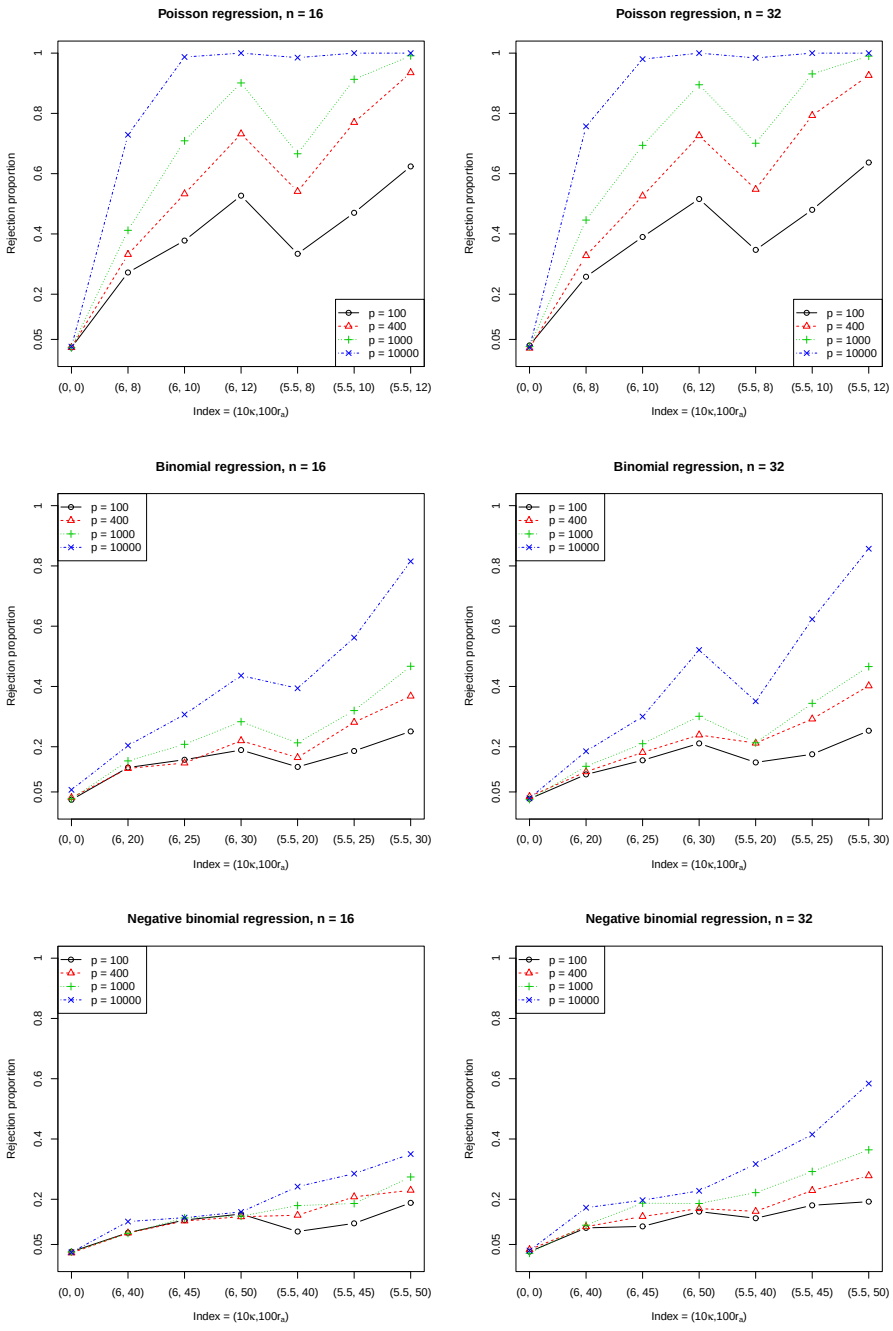

Fig 2: Empirical sizes and powers of the multi-level thresholding test for the hypothesis (6.2) under Poisson, binomial and negative binomial regression. The vertical axis shows the proportion of rejections. The horizontal axis gives the null hypothesis, represented by (0, 0), and six alternative hypotheses. The first and second index of the horizontal axis give the values of 10κ and $100r_a$, respectively, providing signal sparsity and strength.

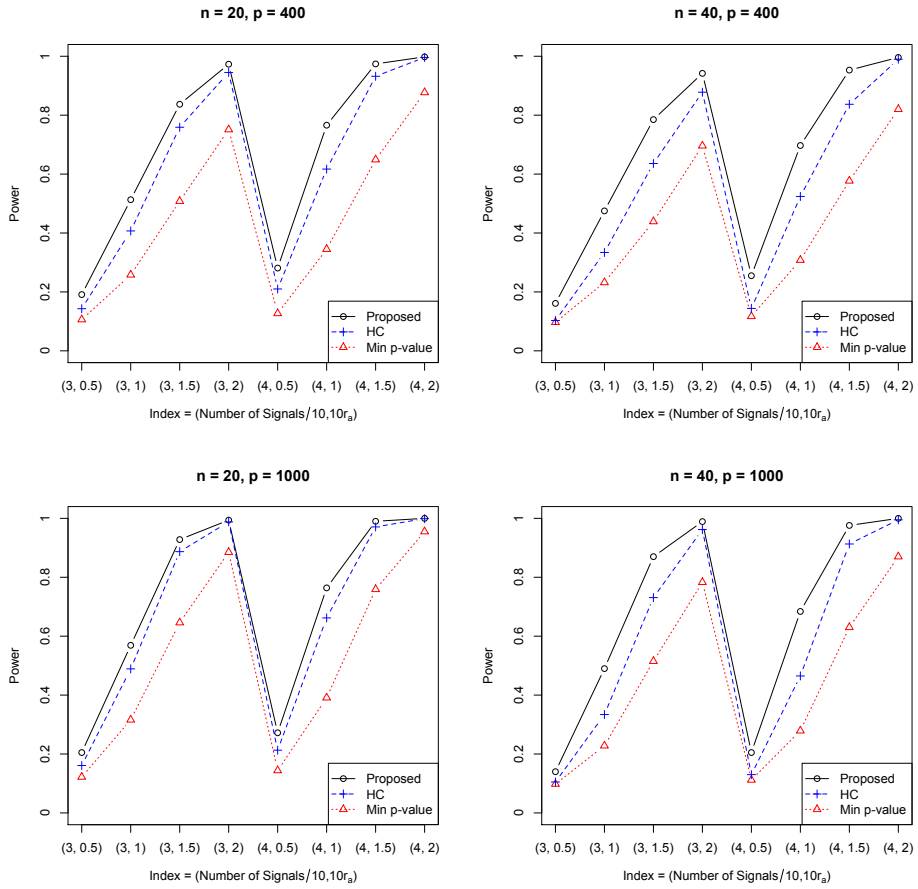

Fig 3: Power comparison between the multi-level thresholding test, HC test and minimum p-value test. The vertical axis gives the empirical powers of the three tests for hypothesis (6.1) under Poisson regression and $p = 400, 1000$. The first and second index of the horizontal axis give one tenth of the number of signals and $10r_a$, respectively.

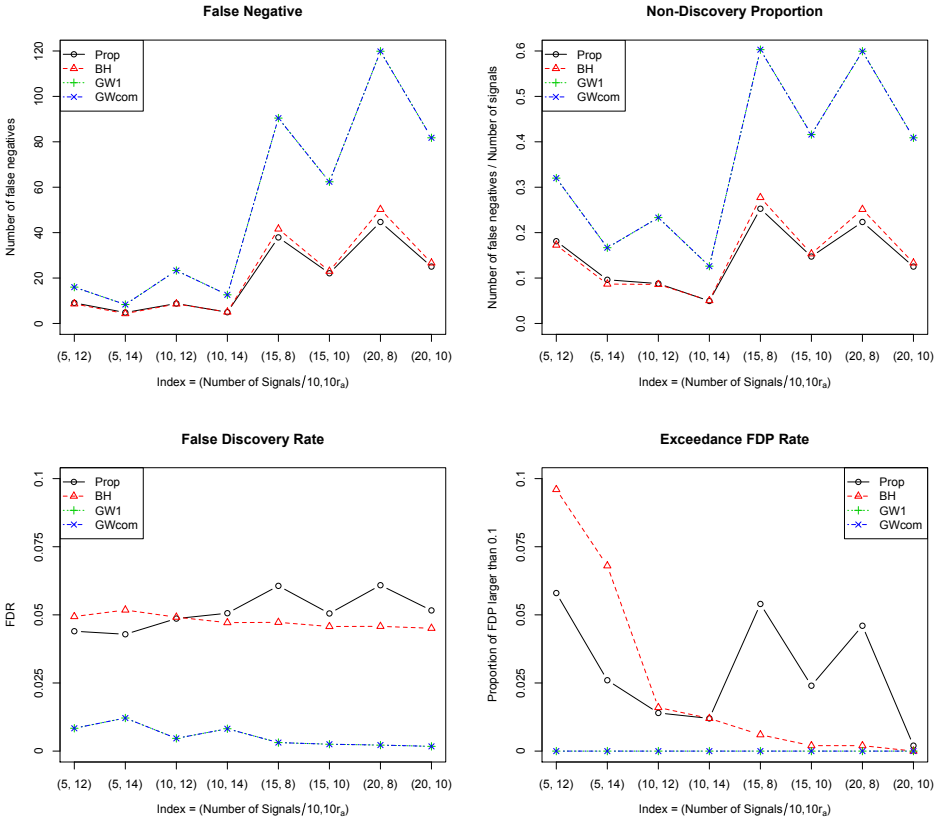

Fig 4: Averages of the number of false negatives (type II errors), the non-discovery proportion (number of false negatives/number of signals), FDR and the proportion of FDP in excess of 0.1 for the proposed multiple testing procedure, BH procedure and GW procedures with $k = 1$ and the combined k approach for the negative binomial regression under H_a in (6.1) with $\beta_{j1} = 2.5$, $n = 40$ and $p = 10000$. The first and second index of the horizontal axis give one tenth of the number of signals and $10r_a$, respectively.

where $\epsilon = p^{-\kappa}$ for $\kappa \in (0,1)$, $a_i = r_{t,i}\sqrt{2\log p}$ and $r_{t,i}$ is the ith row of r_t for $i = 1, \cdots, d$.

Let $\mu = (a_1, \cdots, a_d)'$. Note that $|\mu|^2 = r_t' r_t (2\log p)$. Let P and Q be the distribution under H_0 and H_a of (A.1). Due to the independence between responses, it follows that $P = P_1^p$ and $Q = Q_1^p$, where P_1 and Q_1 are the distributions of the jth response under H_0 and H_a, respectively. We have

$$H^2(P,Q) = 2 - 2\left(1 - \frac{H^2(P_1,Q_1)}{2}\right)^p \quad \text{for} \quad H^2(P_1,Q_1) = \int \left(\sqrt{\frac{dQ_1}{d\Phi}} - 1\right)^2 d\Phi,$$

where Φ is the d-dimensional standard normal distribution.

It can be shown that, if $H^2(P_1,Q_1) = o(p^{-1})$, then $H^2(P,Q) \to 0$, and no test can distinguish H_0 and H_a of (A.1), asymptotically. Let $L(y) = dQ_1/d\Phi$ be the likelihood ratio. We have

$$\begin{aligned} L(y) &= \frac{(1-\epsilon)\exp(-y'y/2) + \epsilon \int \exp\{-(y-\mu)'(y-\mu)/2\} dF(r)}{\exp(-y'y/2)} \\ &= (1-\epsilon) + \epsilon \int \exp(y'\mu - ||\mu||^2/2) dF(r). \end{aligned}$$

Let L_p be a multi-$\log(p)$ term which may change from case to case. Define $f = \mu/\sqrt{2\log p}$. By Jensen's inequality, it follows that

$$H^2(P_1,Q_1) \le L_p \int\int \left\{\sqrt{1 + p^{-\kappa}\left(p^{2w'f - r_*} - 1\right)} - 1\right\}^2 p^{-w'w} dw dF(r),$$

where $r_* = r_t' r_t$. Given r, it can be shown that the leading order of the inner integration in the term above is

$$\int p^{\{2w'f - r_* - \kappa\} \wedge \{4w'f - 2r_* - 2\kappa\} - w'w} dw \cong p^{\max_w\{g(w,r_*)\}},$$

where $g(w,r_*) = \{2w'f - r_* - \kappa\} \wedge \{4w'f - 2r_* - 2\kappa\} - w'w$. Hence, the optimal detection boundary is determined by whether $\max_{w, r\in\mathcal{G}}\{g(w,r_*)\}$ is larger or smaller than -1. It can be shown that

$$\max_{w\in\mathbf{R}^d}\{g(w,r_*)\} = \begin{cases} -\kappa & \text{if } r_* \ge \kappa \\ -(\kappa + r_*)^2/4r_* & \text{if } \kappa/3 \le r_* < \kappa \\ 2r_* - 2\kappa & \text{if } r_* < \kappa/3, \end{cases}$$

and $\max_w\{g(w,r_*)\}$ is an increasing function of r_*. The optimal detection boundary $\mathrm{DB}(\kappa)$ in Theorem 4 follows by noting $\tilde{r} = \max_{r\in\mathcal{G}} \lim_{n\to\infty} r_*$. □

Acknowledgments. The authors thank the AE and two reviewers for comments and suggestions which led to improvement in the presentation of the paper, and Professor Frank Hochholdinger of the University of Bonn for sharing the maize RNA-seq data. The authors acknowledge support from China's National Key Research Special Program Grants SQ2016ZY01002112 and 2015CB856000, China's National Natural Science Foundation grants 71371016 and 71532001, and NSF grants DSM-1309210 and DMS-1127914.

References.

[1] Anders, S. and Huber, W. (2010). Differential Expression Analysis for Sequence Count Data. *Genome Biology*, **11**.

[2] Arias-Castro, E., Candes, E. and Plan, Y. (2011). Global Testing under Sparse Alternatives: ANOVA, Multiple Comparison and the Higher Criticism. *The Annals of Statistics*, **39**, 2533 - 2556.

[3] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B*, **57**, 289-300.

[4] Bradley, R. (2005). Basic Properties of Strong Mixing Conditions: A Survey and Some Open Questions. *Probability Surveys*, **2**, 107 - 144.

[5] Chen, S. X., and Qin, Y.-L. (2010). A Two Sample Test for High Dimensional Data with Applications to Gene-set Testing. *The Annals of Statistics* **38**, 808-835.

[6] Delaigle, A., Hall, P. and Jin, J. (2011). Robustness and Accuracy of Methods for High Dimensional Data Analysis Based on Student's t-statistic. *Journal of the Royal Statistical Society Series B*, **73**, 283 - 301.

[7] Donoho, D. and Jin, J. (2004). Higher Criticism for Detecting Sparse Heterogeneous Mixtures. *The Annals of Statistics*, **32**, 962 - 994.

[8] Fan, Y., Jin, J. and Yao Z. (2013). Optimal Classification in Sparse Gaussian Graphic Model. *The Annals of Statistics*, **41**, 2537 - 2571.

[9] Fan, J. and Song, R. (2010). Sure Independent Screening in Generalized Linear Models with NP-dimensionality. *The Annals of Statistics*, **38**, 3567 - 3604.

[10] Genovese, C., and Wasserman, L. (2006). Exceedance Control of the False Discovery Proportion. *Journal of the American Statistical Association*, **101**, 1408 - 1417.

[11] 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.

[12] Guo, B. and Chen, S. X. (2016). Tests for High Dimensional Generalized Linear Models. *Journal of the Royal Statistical Society Series B*, to appear.

[13] Hall, P. and Jin, J. (2008). Properties of Higher Criticism under Strong Dependence. *The Annals of Statistics*, **36**, 381 - 402.

[14] Hall, P. and Jin, J. (2010). Innovated Higher Criticism for Detecting Sparse Signals in Correlated Noise. *The Annals of Statistics*, **38**, 1686-1732.

[15] Inglot, T. and Kallenberg, W. (2003). Moderate Deviations of Minimum Contrast Estimators under Contamination. *The Annals of Statistics*, **31**, 852 - 879.

[16] Ingster, Y. I. (1997). Some Problems of Hypothesis Testing Leading to Infinitely Divisible Distributions. *Mathematical Methods of Statistics*, **6**, 47 - 69.

[17] Jensen, J. L. and Wood, A. T. A. (1998). Large Deviation and Other Results for Minimum Contrast Estimators. *Ann. Inst. Statist. Math.*, **50**, 673 - 695.

[18] Ji, P. and Jin, J. (2012). UPS Delivers Optimal Phase Diagram in High-dimensional Variable Selection. *The Annals of Statistics*, **40**, 73-103.
[19] Lehmann, E.L. (1959), *Testing Statistical Hypotheses*. John Wiley & Sons, New York.
[20] Lund, S.P., Nettleton, D., McCarthy, D.J. and Smyth, G.K. (2012). Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates. *Statistical Applications in Genetics and Molecular Biology*, **11**.
[21] McCulloch, C., Searle, S. and Neuhaus, J. (2008). *Generalized, Linear, and Mixed Models*. John Wiley and Sons Press, Hoboken.
[22] Paschold, A., Larson, N. B., Marcon, C., Schnable, J. C., Yeh, C. T., Lanz, C., Nettleton, D., Piepho, H.-P., Schnable, P. S., Hochholdinger, F. (2014). Non-syntenic Genes Drive Highly Dynamic Complementation of Gene Expression in Maize Hybrids. *The Plant Cell*, Accepted.
[23] Robinson M. D. and Smyth G. K. (2007). Moderated Statistical Tests for Assessing Differences in Tag Abundance. *Bioinformatics*, **23**, 2881 - 2887.
[24] Robinson M. D. and Smyth G. K. (2008). Small-sample Estimation of Negative Binomial Dispersion, with Applications to SAGE Data. *Biostatistics*, **9**, 321 - 332.
[25] Saulis, L. and Statulevičius, V. A. (1991). *Limit Theorems for Large Deviations*. Kluwer Academic Publishers, The Netherlands.
[26] van de Geer, S., Buhlmann, P., Ritov, Y. and Dezeure, R. (2014). On Asymptotically Optimal Confidence Regions and Tests for High-dimensional Models. *The Annals of Statistics*, **42**, 1166-1202.
[27] van der Vaart, A.W. (2000), *Asymptotic Statistics*. Cambridge University Press, Cambridge.
[28] Zhang, C.-H. and Zhang, S. S. (2014). Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models. *Journal of the Royal Statistical Society Series B*, **76**, 217-242.
[29] Zhong, P.-S. and Chen, S. X. (2011). Tests for High-Dimensional Regression Coefficients with Factorial Designs. *Journal of the American Statistical Association*, **106**, 260 - 274.
[30] Zhong, P.-S., Chen, S. X. and Xu, M. (2013). Tests Alternative to Higher Criticism for High Dimensional Means under Sparsity and Column-wise Dependence. *The Annals of Statistics*, **41**, 2820 - 2851.

Department of Statistics
University of Nebraska Lincoln
Lincoln, Nebraska 68583-0963, USA
E-mail: yumouqiu@unl.edu

Guanghua School of Management and
Center for Statistical Science
Peking University
Beijing 100871, China
E-mail: songchen@iastate.edu

Department of Statistics
Iowa State University
Ames, Iowa 50011-1210, USA
E-mail: dnett@iastate.edu