Web Appendix for Two-sample test with *g*-modeling and its applications

Jingyi Zhai and Hui Jiang

S1 | MATHEMATICAL DERIVATIONS

S1.1 | Asymptotic distribution of test statistics in simple case

Considering the simple case as defined in section 2,

$$\begin{aligned} X_i \overset{ind}{\sim} p_i(X_i|\Theta_i), \quad Y_k \overset{ind}{\sim} p_k(Y_k|\Lambda_k), \quad i = 1, \dots, N_X, \quad k = 1, \dots, N_Y, \\ \text{and} \qquad \Theta_i \overset{ind}{\sim} G, \quad \Lambda_k \overset{ind}{\sim} H, \end{aligned} \tag{1}$$

There are two groups of random sample X and Y follow any exponential family distribution. Under the g-modeling framework, we assume G and H with semi-parametric exponential distributions as defined in section 2.1,

$$\begin{aligned} Pr(\Theta_i = \theta_j) = g_j(\boldsymbol{\alpha}_X) = \exp\{\boldsymbol{Q}_j^T \boldsymbol{\alpha}_X - \phi(\boldsymbol{\alpha}_X)\}, \text{ for } j = 1, \dots, m, \\ Pr(\Lambda_k = \theta_j) = h_j(\boldsymbol{\alpha}_Y) = \exp\{\boldsymbol{Q}_j^T \boldsymbol{\alpha}_Y - \phi(\boldsymbol{\alpha}_Y)\}, \text{ for } j = 1, \dots, m, \end{aligned} \tag{2}$$

With the MLE algorithm, we can estimate the parameter vector $\boldsymbol{\alpha}$ in the semi-parametirc exponential distributions as $\hat{\boldsymbol{\alpha}}$. Before the of two-sample test statistics, we first review the asymptotic distributions of $\hat{\boldsymbol{\alpha}}$ and the corresponding PDF $g(\hat{\boldsymbol{\alpha}})$ and CDF $G(\hat{\boldsymbol{\alpha}})$ from Efron (2016)[1]. Based on the penalized likelihood framework, the asymptotic distribution of $\hat{\boldsymbol{\alpha}}$ is a multivariate normal distribution defined as follows,

$$\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \doteq N[-\{\mathcal{I}(\boldsymbol{\alpha}_0) + \ddot{s}(\boldsymbol{\alpha}_0)\}^{-1} \dot{s}(\boldsymbol{\alpha}_0), \{\mathcal{I}(\boldsymbol{\alpha}_0) + \ddot{s}(\boldsymbol{\alpha}_0)\}^{-1} \mathcal{I}(\boldsymbol{\alpha}_0) \{(\boldsymbol{\alpha}_0) + \ddot{s}(\boldsymbol{\alpha}_0)\}^{-1}],$$

where

$$s(\boldsymbol{\alpha}) = c_0 ||\boldsymbol{\alpha}||, \quad \dot{s}(\boldsymbol{\alpha}) = c_0 \frac{\boldsymbol{\alpha}}{||\boldsymbol{\alpha}||}, \quad \ddot{s}(\boldsymbol{\alpha}) = c_0 \frac{c_0}{||\boldsymbol{\alpha}||} (I_p - \frac{\boldsymbol{\alpha}\boldsymbol{\alpha}^T}{||\boldsymbol{\alpha}||^2}).$$

Here $\boldsymbol{\alpha}_0$ is the true value of $\boldsymbol{\alpha}$ and I_p is a $p \times p$ identity matrix. $\mathcal{I}(\boldsymbol{\alpha}_0)$ is the corresponding Fisher information matrix which can be calculated as follows,

$$\mathcal{I}(\boldsymbol{\alpha}) = \boldsymbol{Q}^T [W_i(\boldsymbol{\alpha}) W_i(\boldsymbol{\alpha})^T + W_i(\boldsymbol{\alpha}) g(\boldsymbol{\alpha})^T + g(\boldsymbol{\alpha}) W_i(\boldsymbol{\alpha})^T - Diag\{W_i(\boldsymbol{\alpha})\}] \boldsymbol{Q},$$

where $W_i(\boldsymbol{\alpha})$ is an m-vector with the j-th element defined as

$$w_{ij}(\boldsymbol{\alpha}) = g_j(\boldsymbol{\alpha}) \{p_{ij} / f_j(\boldsymbol{\alpha}) - 1\},$$

and Diag$\{W_i(\boldsymbol{\alpha}))\}$ is an $m \times m$ diagonal matrix with $W_i(\boldsymbol{\alpha})$ as the diagonal components.

Since the true value $\boldsymbol{\alpha}_0$ is unknown in practice, we replace $\boldsymbol{\alpha}_0$ with $\hat{\boldsymbol{\alpha}}$ to approximately compute the bias and covariance matrix as follows,

$$\text{Bias}(\hat{\boldsymbol{\alpha}}) = -\{\mathcal{I}(\hat{\boldsymbol{\alpha}}) + \ddot{s}(\hat{\boldsymbol{\alpha}})\}^{-1} \dot{s}(\hat{\boldsymbol{\alpha}}), \quad \text{Cov}(\hat{\boldsymbol{\alpha}}) = \{\mathcal{I}(\hat{\boldsymbol{\alpha}}) + \ddot{s}(\hat{\boldsymbol{\alpha}})\}^{-1} \mathcal{I}(\hat{\boldsymbol{\alpha}}) \{\mathcal{I}(\hat{\boldsymbol{\alpha}}) + \ddot{s}(\hat{\boldsymbol{\alpha}})\}^{-1}.$$

Furthermore, from the Delta method, the asymptotic distribution of the the estimated density function can be derived as follows,

$$g(\hat{\boldsymbol{\alpha}}) - g(\boldsymbol{\alpha}_0) \doteq N\{D(\hat{\boldsymbol{\alpha}}) \boldsymbol{Q} \text{Bias}(\hat{\boldsymbol{\alpha}}), D(\hat{\boldsymbol{\alpha}}) \boldsymbol{Q} \text{Cov}(\hat{\boldsymbol{\alpha}}) \boldsymbol{Q}^T D(\hat{\boldsymbol{\alpha}})\},$$

where

$$D(\hat{\boldsymbol{\alpha}}) = \text{Diag}\{g(\hat{\boldsymbol{\alpha}})\} - g(\hat{\boldsymbol{\alpha}}) g(\hat{\boldsymbol{\alpha}})^T,$$

and Diag$\{g(\hat{\boldsymbol{\alpha}})\}$ is a $p \times p$ diagonal matrix with $g(\hat{\boldsymbol{\alpha}})$ as the diagonal components.

Since both $G(\hat{\boldsymbol{\alpha}})$ and $g(\hat{\boldsymbol{\alpha}})$ are evaluated on the grid $\boldsymbol{\tau} = (\theta_1, \dots, \theta_m)$, assuming it is an equally-spaced grid for simplicity, we have $G(\hat{\boldsymbol{\alpha}}) = \boldsymbol{A} g(\hat{\boldsymbol{\alpha}})$, where

$$\boldsymbol{A} = \begin{bmatrix} a & 0 & \dots & 0 \\ a & a & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \dots & a \end{bmatrix}.$$

and a is the grid size. Then we have

$$G(\hat{\boldsymbol{\alpha}}) - G(\boldsymbol{\alpha}_0) \doteq N[\boldsymbol{A}D(\hat{\boldsymbol{\alpha}})\boldsymbol{Q}\text{Bias}(\hat{\boldsymbol{\alpha}}), \boldsymbol{A}D(\hat{\boldsymbol{\alpha}})\boldsymbol{Q}\text{Cov}(\hat{\boldsymbol{\alpha}})\boldsymbol{Q}^T D(\hat{\boldsymbol{\alpha}})\boldsymbol{A}^T].$$

Based on the above asymptotic distribution, we can calculate the bias and covariance matrix of $G(\hat{\boldsymbol{\alpha}})$ by following formulas,

$$\text{Bias}\{G(\hat{\boldsymbol{\alpha}})\} = \boldsymbol{A}D(\hat{\boldsymbol{\alpha}})\boldsymbol{Q}\text{Bias}(\hat{\boldsymbol{\alpha}}), \quad \text{Cov}\{G(\hat{\boldsymbol{\alpha}})\} = \boldsymbol{A}D(\hat{\boldsymbol{\alpha}})\boldsymbol{Q}\text{Cov}(\hat{\boldsymbol{\alpha}})\boldsymbol{Q}^T D(\hat{\boldsymbol{\alpha}})\boldsymbol{A}^T.$$

Now we extend to the two-sample K-S test statistics defined as in section 2.2,

$$T = \max_j |\hat{G}_j(\hat{\boldsymbol{\alpha}}_X) - \hat{H}_j(\hat{\boldsymbol{\alpha}}_Y)|, \quad j = 1, \dots, m, \tag{3}$$

which is the maximum absolute difference between $G(\hat{\boldsymbol{\alpha}}_X)$ and $H(\hat{\boldsymbol{\alpha}}_Y)$. As the null hypothesis defined where $\boldsymbol{\alpha}_{X0} = \boldsymbol{\alpha}_{Y0}$, we can pool two groups of sample together to estimate the vector of parameters denoted as $\hat{\boldsymbol{\alpha}}_p$ for the calculate of biases and covariance matrices. Here we assume that the two groups of samples are independent between groups, we finally obtain the asymptotic null distribution of $G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y)$ as follows,

$$G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y) \doteq N[\text{Bias}\{G(\hat{\boldsymbol{\alpha}}_p)\} - \text{Bias}\{H(\hat{\boldsymbol{\alpha}}_p)\}, \text{Cov}\{G(\hat{\boldsymbol{\alpha}}_p)\} + \text{Cov}\{H(\hat{\boldsymbol{\alpha}}_p)\}].$$

S1.2 | Asymptotic distribution of test statistics in zero-inflated Poisson case

In this section, a more complicated case with zero-inflated Poisson distributions is discussed. With the zero-inflated Poisson assumption, The conditional probability of X_i given $\Theta_i = \theta_j$ is assumed to be a zero-inflated Poisson distribution represented as

$$Pr(X_i = x|\Theta_i = \theta_j, \pi) = p_{ij} = I(x = 0)[\pi + (1 - \pi)e^{-d_i\theta_j}] + [1 - I(x = 0)](1 - \pi)e^{-d_i\theta_j}.$$

In this case, we need to estimate π and $\boldsymbol{\alpha}$ simultaneously which can be estimated through the MLE algorithm. Hence we are interested in the joint distribution of π and $\boldsymbol{\alpha}$ and then denote the new vector of parameters as $\boldsymbol{\alpha}' = \begin{pmatrix} \pi \\ \boldsymbol{\alpha} \end{pmatrix}$. Similarly, we can obtain the joint asymptotic distribution of $\hat{\pi}$ and $g(\hat{\boldsymbol{\alpha}})$ by applying the same procedure on $\boldsymbol{\alpha}'$ as in the simple case. We start with the log likelihood of $\boldsymbol{\alpha}'$ for the i-th sample denoted as

$$l_i(\boldsymbol{\alpha}') = \log \boldsymbol{P}_i^T g(\boldsymbol{\alpha}) = log\{\sum_{j=1}^{m} p_{ij} g_j(\boldsymbol{\alpha})\}.$$

Then, the associated first derivatives of log likelihood with respect to π and $\boldsymbol{\alpha}$ can be calculated as follows,

$$\frac{\partial l_i(\boldsymbol{\alpha}')}{\partial \pi} = \frac{1}{\boldsymbol{P}_i^T g(\boldsymbol{\alpha})} \frac{\partial \boldsymbol{P}_i^T}{\partial \pi} g(\boldsymbol{\alpha}) = \frac{1}{f_i(\boldsymbol{\alpha}')} \frac{\partial \boldsymbol{P}_i^T}{\partial \pi} g(\boldsymbol{\alpha}),$$

$$\frac{\partial l_i(\boldsymbol{\alpha}')}{\partial \boldsymbol{\alpha}} = \boldsymbol{Q}^T W_i(\boldsymbol{\alpha}') = \sum_{j=1}^{m} \boldsymbol{Q}_j^T w_{ij}(\boldsymbol{\alpha}') = \sum_{j=1}^{m} \boldsymbol{Q}_j [g_j(\boldsymbol{\alpha})\{p_{ij}/f_i(\boldsymbol{\alpha}') - 1\}],$$

We can further represent the above first derivatives in terms of $\boldsymbol{\alpha}'$ in the joint format as

$$\dot{l}_i(\boldsymbol{\alpha}') = \frac{\partial l_i(\boldsymbol{\alpha}')}{\partial \boldsymbol{\alpha}'} = \begin{bmatrix} \frac{\partial \boldsymbol{P}_i^T}{\partial \pi} g(\boldsymbol{\alpha})\{f_i(\boldsymbol{\alpha}')\}^{-1} \\ \boldsymbol{Q}^T W_i(\boldsymbol{\alpha}') \end{bmatrix},$$

and the overall first derivatives among all samples can be computed as

$$\dot{l}(\boldsymbol{\alpha}') = \sum_{i=1}^{N} \dot{l}_i(\boldsymbol{\alpha}').$$

The second derivatives can also be obtained based on the above first derivatives as follows,

$$-\frac{\partial l_i^2(\boldsymbol{\alpha}')}{\partial \pi^2} = \{f_i^2(\boldsymbol{\alpha}')\}^{-1}\{\frac{\partial \boldsymbol{P}_i^T}{\partial \pi} g(\boldsymbol{\alpha})\}^2,$$

$$-\frac{\partial l_i^2(\boldsymbol{\alpha}')}{\partial \boldsymbol{\alpha}^2} = \boldsymbol{Q}^T[W_i(\boldsymbol{\alpha}')W_i(\boldsymbol{\alpha}')^T + W_i(\boldsymbol{\alpha}')g(\boldsymbol{\alpha})^T + g(\boldsymbol{\alpha})W_i(\boldsymbol{\alpha}')^T - diag\{W_i(\boldsymbol{\alpha}')\}]\boldsymbol{Q},$$

$$-\frac{\partial l_i^2(\boldsymbol{\alpha}')}{\partial \boldsymbol{\alpha}\partial \pi} = \sum_{j=1}^{m} \boldsymbol{Q}_j^T[g_j(\boldsymbol{\alpha})\{p_{ij}\frac{\partial f_i(\boldsymbol{\alpha}')}{\partial \pi} - \frac{\partial p_{ij}}{\partial \pi}f_i(\boldsymbol{\alpha}')\}\{f_i^2(\boldsymbol{\alpha}')\}^{-1}]$$

Based on the above second derivatives formulas, the Fisher information matrix for $\boldsymbol{\alpha}'$ can be obtained as

$$\mathcal{I}(\boldsymbol{\alpha}') = \begin{bmatrix} \partial l_i^2(\boldsymbol{\alpha}')/\partial \pi^2 & -\partial l_i^2(\boldsymbol{\alpha}')/\partial \boldsymbol{\alpha}\partial \pi \\ -\{\partial l_i^2(\boldsymbol{\alpha}')/\partial \boldsymbol{\alpha}\partial \pi\}^T & -\partial l_i^2(\boldsymbol{\alpha}')/\partial \boldsymbol{\alpha}^2 \end{bmatrix}.$$

The regularization term $s(\boldsymbol{\alpha}') = c_0||\boldsymbol{\alpha}'|| = c_0(\pi^2 + \sum_{h=1}^{p}\boldsymbol{\alpha}_h^2)^{1/2}$ is used to improve the accuracy of the estimation of $\boldsymbol{\alpha}'$. The final objective function for MLE algorithm is

$$m(\boldsymbol{\alpha}') = l(\boldsymbol{\alpha}') - s(\boldsymbol{\alpha}').$$

We denote $\boldsymbol{\alpha}_0'$ as the true value of $\boldsymbol{\alpha}'$ and $\hat{\boldsymbol{\alpha}}'$ is the MLE of $\boldsymbol{\alpha}'$ which maximizes $m(\boldsymbol{\alpha}')$. Hence we have the first derivatives of $m(\hat{\boldsymbol{\alpha}}')$ equal to zero as follows,

$$\begin{aligned} \dot{m}(\hat{\boldsymbol{\alpha}}') &\doteq \dot{m}(\boldsymbol{\alpha}_0') + \ddot{m}(\boldsymbol{\alpha}_0')(\hat{\boldsymbol{\alpha}}' - \boldsymbol{\alpha}_0') \\ &= \{\dot{l}(\boldsymbol{\alpha}_0') - \dot{s}(\boldsymbol{\alpha}_0')\} + \{-\ddot{l}(\boldsymbol{\alpha}_0') + \ddot{s}(\boldsymbol{\alpha}_0')\}(\hat{\boldsymbol{\alpha}}' - \boldsymbol{\alpha}_0') \\ &= 0. \end{aligned}$$

This gives

$$\hat{\boldsymbol{\alpha}}' - \boldsymbol{\alpha}_0' \doteq \{-\ddot{l}(\boldsymbol{\alpha}_0') + \ddot{s}(\boldsymbol{\alpha}_0')\}^{-1}\{\dot{l}(\boldsymbol{\alpha}_0') - \dot{s}(\boldsymbol{\alpha}_0')\}$$

As the expectation of $\dot{l}(\boldsymbol{\alpha}_0')$ is zero and $\mathrm{Cov}\{\dot{l}(\boldsymbol{\alpha}_0')\} = \mathcal{I}(\boldsymbol{\alpha}_0')$, and we can further represent the asymptotic distribution of $\hat{\boldsymbol{\alpha}}'$ as,

$$\hat{\boldsymbol{\alpha}}' - \boldsymbol{\alpha}_0' \doteq N[-\{\mathcal{I}(\boldsymbol{\alpha}_0') + \ddot{s}(\boldsymbol{\alpha}_0')\}^{-1}\dot{s}(\boldsymbol{\alpha}_0'), \{\mathcal{I}(\boldsymbol{\alpha}_0') + \ddot{s}(\boldsymbol{\alpha}_0')\}^{-1}\mathcal{I}(\boldsymbol{\alpha}_0')\{\mathcal{I}(\boldsymbol{\alpha}_0') + \ddot{s}(\boldsymbol{\alpha}_0')\}^{-1}],$$

where

$$\dot{s}(\boldsymbol{\alpha}') = c_0\frac{\boldsymbol{\alpha}'}{||\boldsymbol{\alpha}'||}, \quad \ddot{s}(\boldsymbol{\alpha}') = \frac{c_0}{||\boldsymbol{\alpha}'||}(I - \frac{\boldsymbol{\alpha}'\boldsymbol{\alpha}'^T}{||\boldsymbol{\alpha}'||^2}).$$

Here the true value $\boldsymbol{\alpha}_0'$ is unknown in practice. Therefore, we use $\hat{\boldsymbol{\alpha}}'$ instead of $\boldsymbol{\alpha}_0'$ to approximately estimate the bias and covariance matrix of $\hat{\boldsymbol{\alpha}}'$ as follows,

$$\mathrm{Bias}(\hat{\boldsymbol{\alpha}}') = -\{\mathcal{I}(\hat{\boldsymbol{\alpha}}') + \ddot{s}(\hat{\boldsymbol{\alpha}}')\}^{-1}\dot{s}(\hat{\boldsymbol{\alpha}}'),$$

$$\mathrm{Cov}(\hat{\boldsymbol{\alpha}}') = \{\mathcal{I}(\hat{\boldsymbol{\alpha}}') + \ddot{s}(\hat{\boldsymbol{\alpha}}')\}^{-1}\mathcal{I}(\hat{\boldsymbol{\alpha}}')\{\mathcal{I}(\hat{\boldsymbol{\alpha}}') + \ddot{s}(\hat{\boldsymbol{\alpha}}')\}^{-1},$$

Based on the asymptotic distribution of $\hat{\boldsymbol{\alpha}}'$, the joint asymptotic distribution of $\hat{\pi}$ and $g(\hat{\boldsymbol{\alpha}})$ can be obtained as we can represent $\begin{bmatrix} \hat{\pi} \\ g(\hat{\boldsymbol{\alpha}}) \end{bmatrix}$ in terms of $\hat{\boldsymbol{\alpha}}'$ with the transformation function K as follows,

$$\begin{bmatrix} \hat{\pi} \\ g(\hat{\boldsymbol{\alpha}}) \end{bmatrix} = K(\hat{\boldsymbol{\alpha}}').$$

We have the first derivative matrix of g with $\hat{\boldsymbol{\alpha}}'$ which can be computed by $\dot{g} = \boldsymbol{Q}^T D$ from Efron (2016)[1]. Therefore, we can further calculate the first derivative of K when takes value $\hat{\boldsymbol{\alpha}}'$ as

$$\dot{\boldsymbol{K}} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \dot{g} & \\ 0 & & & \end{bmatrix},$$

Based on the Delta method, the joint asymptotic null distribution of $\hat{\pi}$ and $g(\hat{\boldsymbol{\alpha}})$ can be derived as follows,

$$\begin{bmatrix} \hat{\pi} - \pi_0 \\ g(\hat{\boldsymbol{\alpha}}) - g(\boldsymbol{\alpha}_0) \end{bmatrix} \doteq N[\mathrm{Bias}\{K(\hat{\boldsymbol{\alpha}}')\}, \mathrm{Cov}\{K(\hat{\boldsymbol{\alpha}}')\}],$$

with

$$\mathrm{Bias}\{K(\hat{\boldsymbol{\alpha}}')\} = \dot{\boldsymbol{K}}^T\mathrm{Bias}(\hat{\boldsymbol{\alpha}}'), \quad \mathrm{Cov}\{K(\hat{\boldsymbol{\alpha}}')\} = \dot{\boldsymbol{K}}^T\mathrm{Cov}(\hat{\boldsymbol{\alpha}}')\dot{\boldsymbol{K}}.$$

Moreover, the joint asymptotic null distribution of $\hat{\pi}$ and $G(\hat{\boldsymbol{\alpha}})$ can be obtained as we can denote $C(\hat{\boldsymbol{\alpha}}') = \begin{bmatrix} \hat{\pi} \\ G(\hat{\boldsymbol{\alpha}}) \end{bmatrix}$ in terms of $\begin{bmatrix} \hat{\pi} \\ g(\hat{\boldsymbol{\alpha}}) \end{bmatrix}$ with the transformation matrix $\boldsymbol{B}$ as follows,

$$\begin{bmatrix} \hat{\pi} \\ G(\hat{\boldsymbol{\alpha}}) \end{bmatrix} = \boldsymbol{B} \begin{bmatrix} \hat{\pi} \\ g(\hat{\boldsymbol{\alpha}}) \end{bmatrix},$$

where

$$\boldsymbol{B} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \boldsymbol{A} & \\ 0 & & & \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} a & 0 & \dots & 0 \\ a & a & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \dots & a \end{bmatrix},$$

where a is the grid size for equally spaced grids.

Since $\boldsymbol{B}$ is a linear transform matrix, the associated bias and covariance of $C(\hat{\boldsymbol{\alpha}}')$ can be computed as

$$\text{Bias}\{C(\hat{\boldsymbol{\alpha}}')\} = \boldsymbol{B}\text{Bias}\{K(\hat{\boldsymbol{\alpha}}')\}, \quad \text{Cov}\{C(\hat{\boldsymbol{\alpha}}')\} = \boldsymbol{B}\text{Cov}\{K(\hat{\boldsymbol{\alpha}}')\}\boldsymbol{B}^T.$$

Extend to the two-sample test, a modified two-sample K-S test statistics is defined as in section 2,

$$T = \max\{\max_j |\hat{G}_j(\hat{\boldsymbol{\alpha}}_X) - \hat{H}_j(\hat{\boldsymbol{\alpha}}_Y)|, |\hat{\pi}_X - \hat{\pi}_Y|\}, \tag{4}$$

which includes both the absolute change in CDFs and the absolute difference in π between two groups. Similarly as in the simple case, we pool X and Y together to obtain the estimated augmented vector of parameters $\hat{\boldsymbol{\alpha}}_p'$ under the null hypotheis. Hence, the joint asymptotic distribution of $\hat{\pi}_X - \hat{\pi}_Y$ and $G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y)$ is addressed and it is a multivariate normal distribution with mean and covariance as follows,

$$\begin{bmatrix} \hat{\pi}_X - \hat{\pi}_Y \\ G(\hat{\boldsymbol{\alpha}}_X) - H(\hat{\boldsymbol{\alpha}}_Y) \end{bmatrix} \doteq N[\text{Bias}\{C(\hat{\boldsymbol{\alpha}}_p')\} - \text{Bias}\{C(\hat{\boldsymbol{\alpha}}_p')\}, \text{Cov}\{C(\hat{\boldsymbol{\alpha}}_p')\} + \text{Cov}\{C(\hat{\boldsymbol{\alpha}}_p')\}].$$

S2 | ADDITIONAL SIMULATIONS

We also run simulations on several different distributions besides zero-inflated Poisson distribution. There is one example for each type of X and Y distribution. X and Y distribution includes common distributions such as normal distribution, binomial distribution and poisson distribution. More complex types of X and Y distribution are also considered, for example, Poisson distribution with additional parameter.

S2.1 | Normal distribution

Suppose that there are groups of normal random samples X and Y as

$$X_i \sim N(\Theta_i, 1), \quad Y_j \sim N(\Lambda_j, 1), \quad i, j = 1, \dots 100$$

Under null hypothesis, we simulate Θ and Λ both from standard normal distribution, i.e.

$$\Theta_i \sim N(0, 1), \quad \Lambda_j \sim N(0, 1)$$

Under alternative hypothesis, Θ and Λ are simulated from two different normal distribution with different means but same standard deviations, i.e.,

$$\Theta_i \sim N(0, 1), \quad \Lambda_j \sim N(0 + \Delta, 1), \quad \Delta = 0.1, 0.2, \dots, 1$$

We ran 1000 simulations and $B = 99$ bootstraps for each simulation. Under the null hypothesis, p-values are uniformly distributed on $(0, 1)$ in theory. As shown in Figure S1a, the histogram of p-values from 1000 simulations is close to $U(0, 1)$. In addition, the Q-Q plot of p-values (Figure S1b) is nearly the diagonal line across the origin point. Both figures demonstrates that the K-S test is valid under the null distribution. In details, from Figure S1c, the estimated distributions of Θ and Λ are close to the true distribution $N(0, 1)$ except two tails and peak point.

Moreover, in order to access the performance of our K-S test under the alternative hypothesis, power plot was present with the difference in mean of two normal distributions. From Figure S3, the probability of type I error is 0.04 slightly less than

the predefined test level $\alpha = 0.05$. Therefore, the new K-S test controls the probability of type I error. The power of this new K-S test increases with the larger difference in the mean of normal distributions. The larger difference in the mean of normal distributions means that $g(\theta)$ and $h(\theta)$ are more different from each other. When the difference in mean is 0.8, the power of the test reaches 0.90. Hence, our K-S test is sensitive to the difference in mean for normal distributions.

S2.2 | Binomial distribution

If we have two groups of random counts where,

$$X_i \sim Bin(n_X, \Theta_i), \quad Y_j \sim Bin(n_Y, \Lambda_j,), \quad i, j = 1, \dots 1000$$

$$n_X \sim Poi(50), \quad n_Y \sim Poi(50)$$

For null hypothesis, we simulate the two unobserved parameters according to the following distributions

$$\Theta_i \sim Beta(2, 2), \quad \Lambda_j \sim Beta(2, 2)$$

For alternative hypothesis, two different Beta distributions are assumed to generate Θ and Λ i.e.,

$$\Theta_i \sim Beta(2, 2)$$

$$\Lambda_j \sim \pi Beta(2, 2) + (1 - \pi) Beta(2, 5), \quad \pi = 0.1, 0.2, \dots, 1$$

Here h is a mixture distribution of $Beta(2, 2)$ and $Beta(2, 5)$. π determines the closeness of the mixed distribution to $Beta(2, 2)$. g is more different from h as π increases.

One thousand simulations were ran, and each simulation involved $B = 99$ bootstraps for empirical p-value calculation. Under the null hypothesis, p-values is uniformly distributed on (0, 1). As shown in Figure S4a, the histogram of p-values from 1000 simulations is close to $U(0, 1)$. In addition, the Q-Q plot of p-values demonstrates that the K-S test is valid under the null distribution. In details, from Figure S4c, the estimated distributions of Θ and Λ are close to the true distribution $Beta(2, 2)$, though there are biases at the tail and peak point.

Moreover, in order to access the performance of our K-S test under the alternative hypothesis, power plot was shown with different π values. From Figure S5, the probability of type I error is 0.03 slightly less than the predefined test level $\alpha = 0.05$. Therefore, the new K-S test controls the probability of type I error. The power of this new K-S test increases with the larger difference in the mean of the normal distribution. When $\pi = 0.6$, the power of the test reaches 0.90.

S2.3 | Poisson distribution

S2.3.1 | Simple Poisson

Consider the simplest case in Poisson distribution, suppose we observe two groups of counts as follows,

$$X_i \sim Poi(\Theta_i), \quad Y_j \sim Poi(\Lambda_j), \quad i, j = 1, \dots 100$$

Under null hypothesis, we simulate the two unknown parameters from the following distribution

$$\Theta_i \sim \chi^2_{10}, \quad \Lambda_j \sim \chi^2_{10}$$

Under alternative hypothesis, Θ and Λ are simulated from two different chi-square distributions, i.e.,

$$\Theta_i \sim \chi^2_{10}, \quad \Lambda_j \sim \chi^2_{10+\Delta}, \quad \Delta = 1, \dots, 5$$

One thousand simulations were ran with $B = 99$ bootstraps for each simulation. Under the null hypothesis, p-values is uniformly distributed on (0, 1). As shown in Figure S6a, the histogram of p-values from 1000 simulations is close to $U(0, 1)$. In addition, the Q-Q plot of p-values demonstrates that the K-S test is valid under the null distribution. In details, from Figure S6c, the estimated distributions of Θ and η are close to the true distribution χ^2_{10}, though there are biases at the tail and peak point.

Moreover, in order to access the performance of our K-S test under the alternative hypothesis, power plot were present with difference in degree of freedom between two chi-square distributions. From Figure S7, the probability of type I error is 0.03 which is close to the predefined test level $\alpha = 0.05$. Therefore, the new K-S test has controlled probability of type I error. The power of this new K-S test increases with the larger difference in the degree freedom of chi-square distribution. When the difference in degree of freedom is 4, the power of the test is over 0.90.

S2.3.2 | Poisson distribution with additional parameter

We extended the previous simple Poisson case to a more complexed case with addition parameter d involved. d is known parameter but introduces more uncertainty in the Poisson distribution. Now we have

$$X_i \sim Poi(d_{Xi}\Theta_i), \quad Y_i \sim Poi(d_{Yj}\Lambda_j), \quad i, j = 1, \dots 100$$
$$d_{Xi} \sim Unif(0.5, 1), \quad d_{Yj} \sim Unif(0.5, 1).$$

Under null hypothesis, Θ and Λ are both generated from χ^2_{10},

$$\Theta_i \sim \chi^2_{10}, \quad \Lambda_j \sim \chi^2_{10}$$

Under alternative hypothesis, we simulate Θ and Λ from two different chi-square distributions,

$$\Theta_i \sim \chi^2_{10}, \quad \Lambda_j \sim \chi^2_{10+\Delta}, \quad \Delta = 1, \dots, 5$$

We ran 1000 simulations and $B = 99$ bootstraps for each simulation. Under the null hypothesis, p-values is uniformly distributed on (0, 1). As shown in Figure S8a, the histogram of p-values from 1000 simulations is close to $U(0, 1)$. In addition, the Q-Q plot of p-values demonstrates that the K-S test is valid under the null distribution. In details, from Figure S8c, the estimated distributions of Θ and Λ are close to the true distribution χ^2_{10}, though there are biases at the tail and peak point.

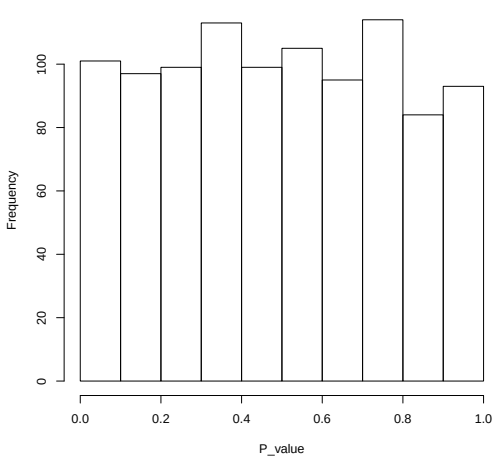
Moreover, in order to access the performance of our K-S test under the alternative hypothesis, power plot were present with difference in degree of freedom between two chi-square distributions. From Figure S9, the probability of type I error is 0.06 which is a bit higher than the predefined test level $\alpha = 0.05$. Therefore, the new K-S test has a slightly inflated type I error. The power of this new K-S test increases with the larger difference in the degree freedom of chi-square distribution. When the difference in degree of freedom is 4, the power of the test reaches 0.95.

S3 | DE GENE EXAMPLE

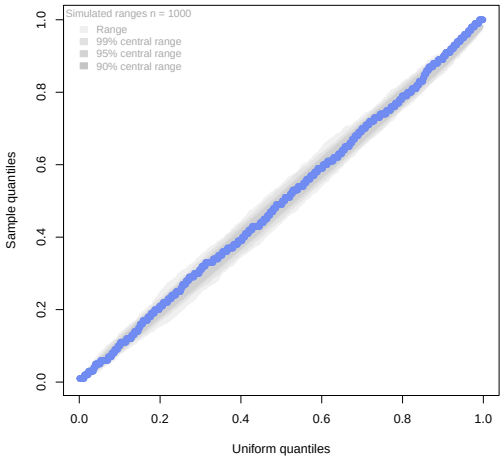
Here is an example of the DE gene that can only be detected by our proposed method (KS). Figure S10a and S10b shows the g-modeling based estimated density and c.d.f of this gene in each cell group (E3 v.s. E4). The average gene expression of this gene in the two groups are close to each other, while the shapes of distribution are different. In addition, the estimated excess probability of zero counts is 0.06 for E3 group and 0.03 for E4 group, so the two groups have the similar excess probability of zero counts. There is only the difference in the shape of gene distribution, which can only be detected by our method.

References

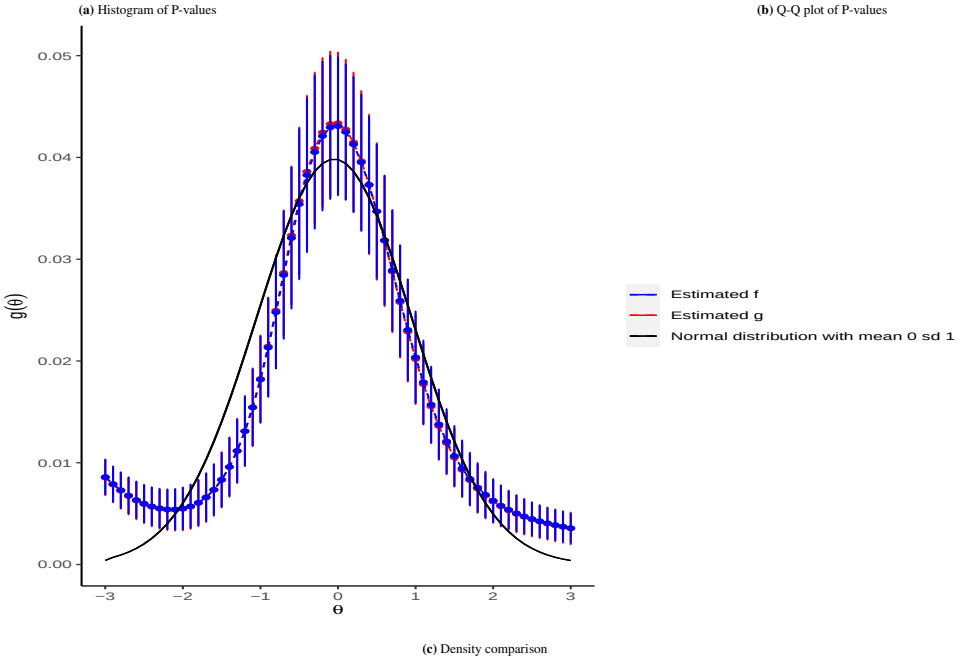
1. Efron B. Empirical Bayes deconvolution estimates. *Biometrika* 2016; 103(1): 1-20. doi: 10.1093/biomet/asv068



(a) Histogram of P-values



(b) Q-Q plot of P-values



(c) Density comparison

FIGURE S1 For the normal case, (a) histogram for estimated p-values from 1000 simulations, (b) Q-Q plot of estimated p-values against $U(0,1)$, (c) estimated densities for two simulated samples against the true density $N(0,1)$.

TABLE S1 Mean running time (SD) (in seconds) of proposed simple and accelerated bootstrap procedures with varying sample sizes in each group under null hypothesis for normal-based model.

Sample Size	Simple Bootstrap	Accelerated Bootstrap
100	3.37 (0.06)	0.21 (0.01)
200	6.54 (0.25)	0.34 (0.02)
400	12.21 (0.40)	0.63 (0.02)
800	24.45 (0.59)	1.26 (0.07)
1600	48.30 (1.67)	2.30 (0.08)

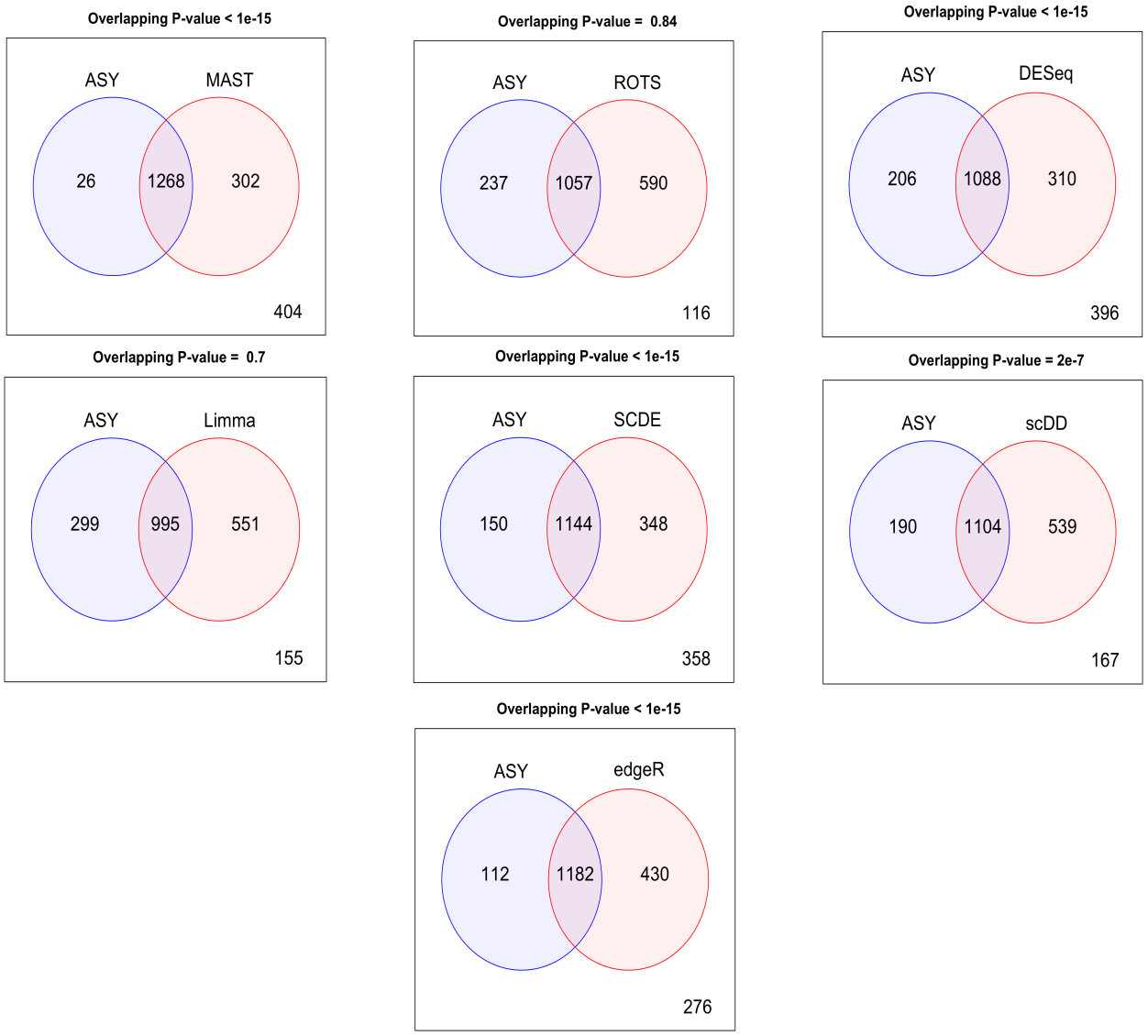


FIGURE S2 Venn diagrams for numbers of DE genes found by proposed method with accelerated bootstrap (ASY) versus other methods with p-values from hypergeometric tests.

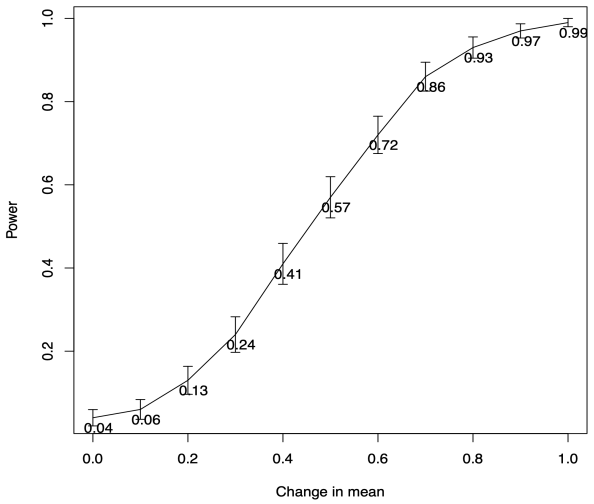
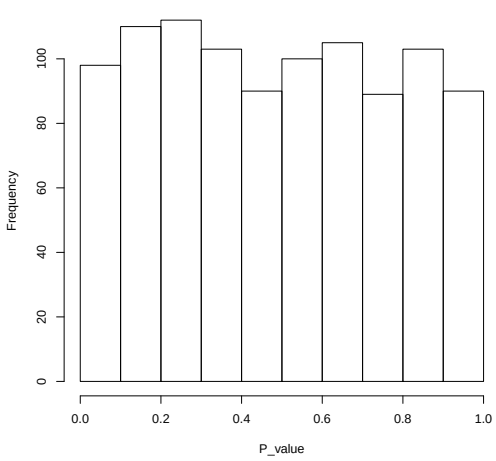
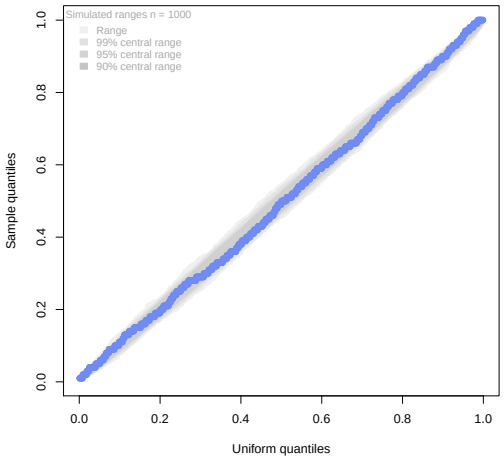


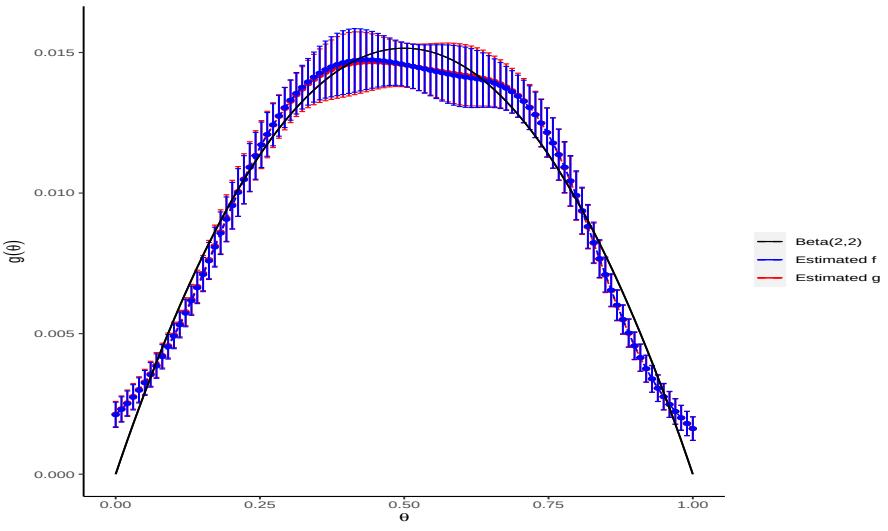
FIGURE S3 Power plots for the normal case. For the line, the power changes with the difference in mean varying from 0.1 to 1. The vertical bar represents the mean $\pm$ sd for each power value.



(a) Histogram of P-values



(b) Q-Q plot of P-values



(c) Density comparison

FIGURE S4 For the binomial case, (a) histogram for estimated p-values from 1000 simulations, (b) Q-Q plot of estimated p-values against $U(0,1)$, (c) estimated densities for two simulated samples against the true density $Beta(2,2)$.

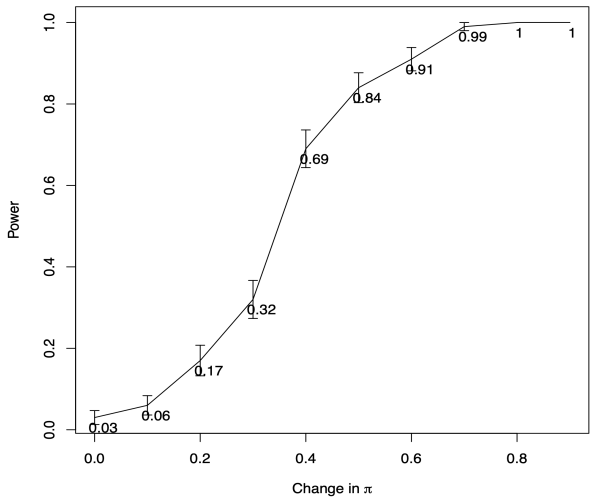
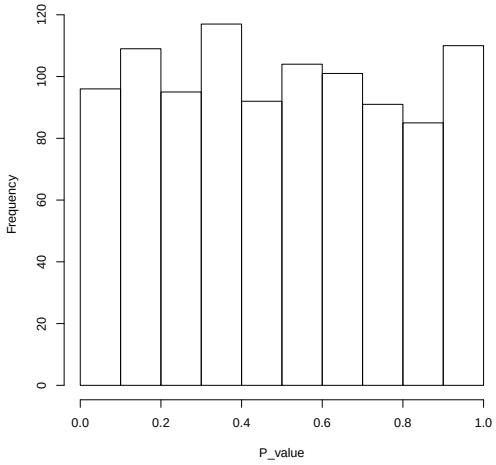
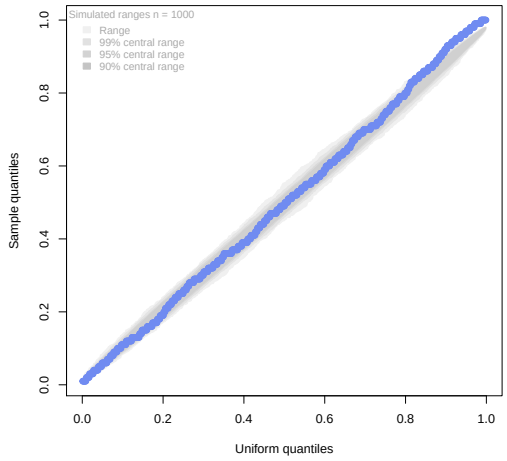


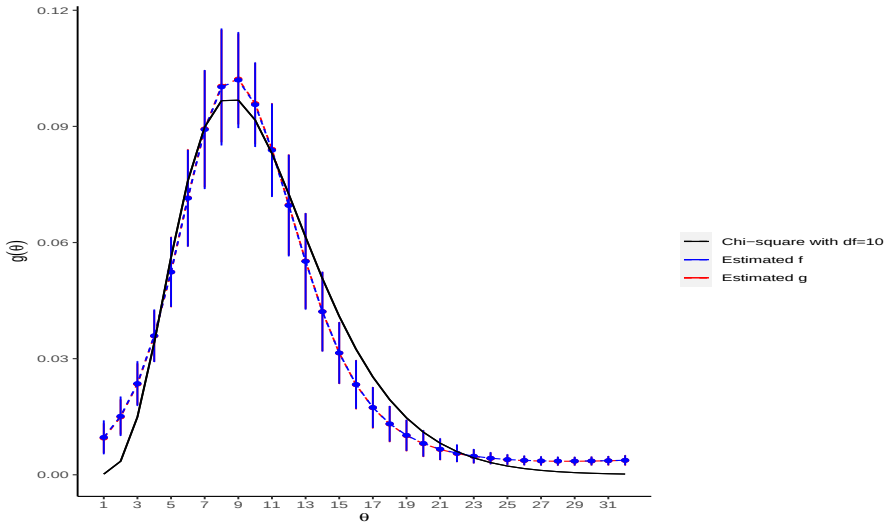
FIGURE S5 Power plots for the binomial case. For the line, the power changes with the difference in π varying from 0.1 to 1. The vertical bar represents the mean $\pm$ sd for each power value.



(a) Histogram of P-values



(b) Q-Q plot of P-values



(c) Density comparison

FIGURE S6 For the Poisson case, (a) histogram for estimated p-values from 1000 simulations, (b) Q-Q plot of estimated p-values against $U(0,1)$, (c) estimated densities for two simulated samples against the true density χ^2_{10}.

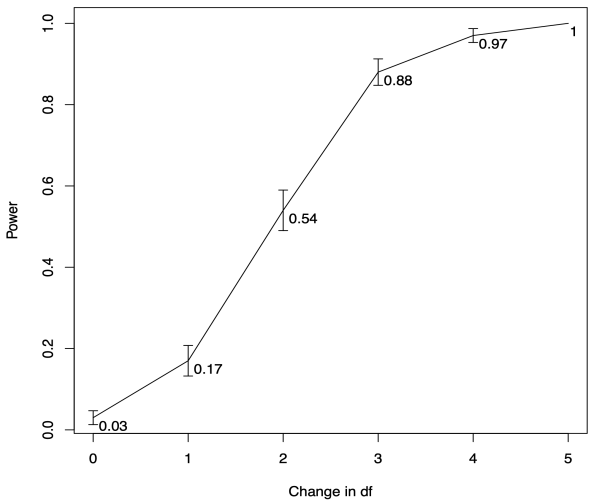
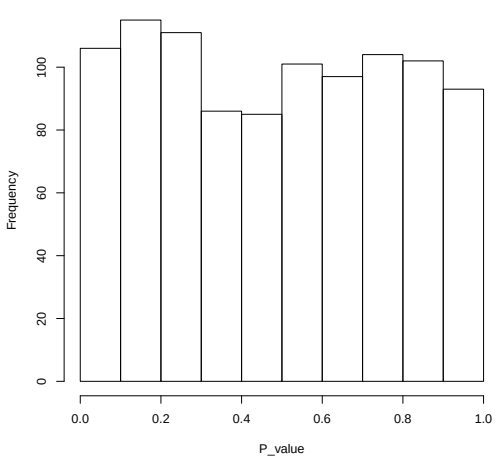
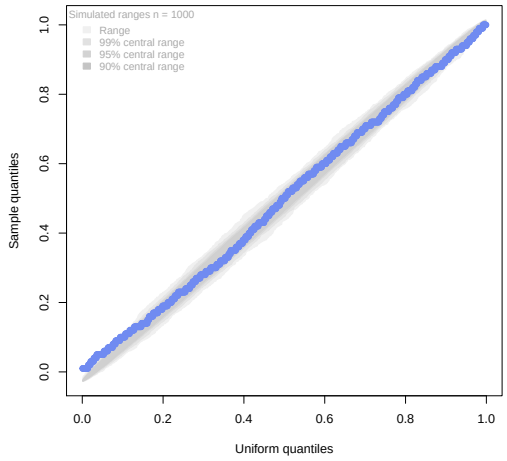


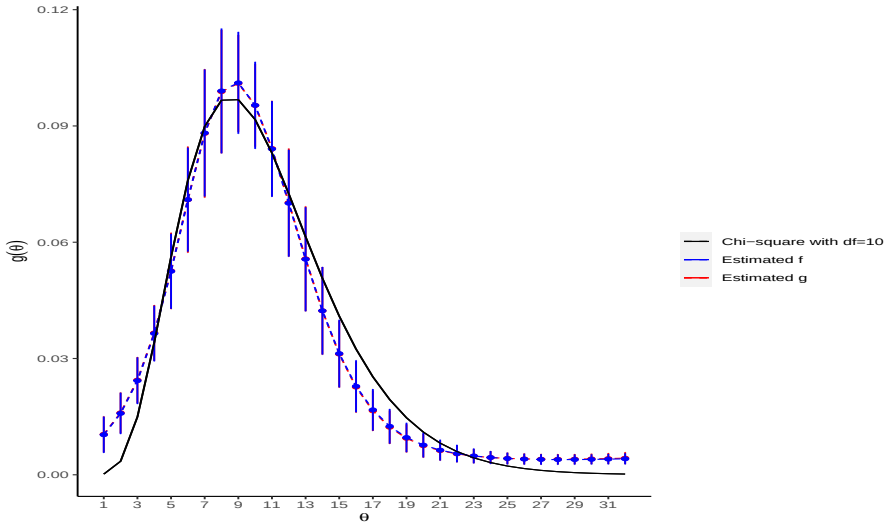
FIGURE S7 Power plots for the Poisson case. For the line, the power changes with the difference in degrees of freedom varying from 0 to 5. The vertical bars represent the mean $\pm$ sd for each power value.



(a) Histogram of P-values



(b) Q-Q plot of P-values



(c) Density comparison

FIGURE S8 For the Poisson with additional parameter case, (a) histogram for estimated p-values from 1000 simulations, (b) Q-Q plot of estimated p-values against $U(0,1)$, (c) estimated densities for two simulated samples against the true density χ^2_{10}.

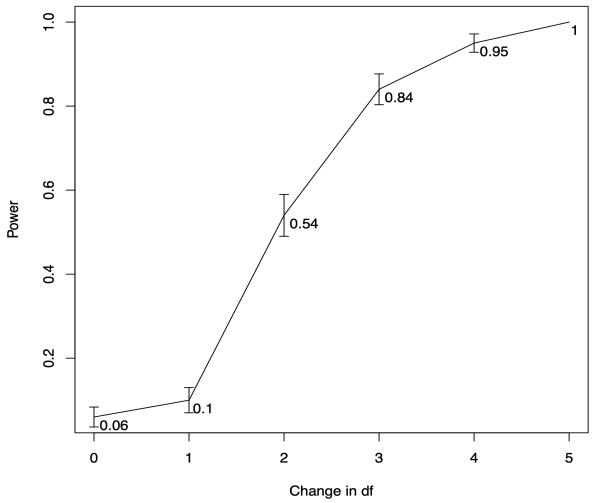
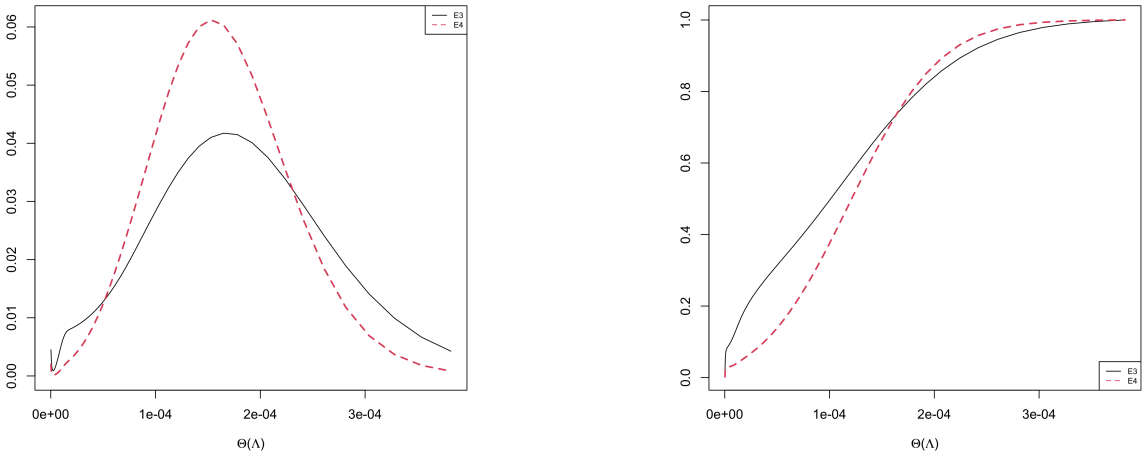


FIGURE S9 Power plots for the Poisson with additional parameter case. For the line, the power changes with the difference in degrees of freedom varying from 0 to 5. The vertical bar represents the mean ± sd for each power value.



(a) Estimated density

(b) Estimated cumulative density

FIGURE S10 An example of a DE gene which can only be detected by our proposed method (KS): (a) Estimated density from the g-modeling method. (b) Estimated cumulative density from the g-modeling method.

How to cite this article: