Estimating empirical null distributions for Chi-squared and Gamma statistics with application to multiple testing in RNA-seq

Xing Ren[1], Jianmin Wang[1,2,*], Song Liu[1,2,*] and Jeffrey C. Miecznikowski[1,2,*]

[1]Department of Biostatistics, SUNY University at Buffalo, Buffalo NY 14214, USA
[2]Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute, Buffalo NY 14263, USA

Abstract

Genome and transcriptome studies using microarray and RNA-seq technologies often involve simultaneous hypothesis testing of thousands of genes or transcripts. A key step determining significant differential expression in such large-scale testing is obtaining the null distribution of the test statistics. We show by examples that the asymptotic null is often inappropriate for many of the χ^2 tests in RNA-seq analysis. Instead we propose a Gamma mixture model and new methods based on maximum likelihood and characteristic functions to estimate an empirical null distribution in these settings. We show by simulations and applications that our proposed methods perform favorably compared with other existing methods.

1 Introduction

Genome and transcriptome studies present thousands of genes for simultaneous hypothesis testing. The goal of these studies is to obtain a list of genes worth further investigation and meanwhile keep the number of type I errors low. A traditional way is to control the family-wise error rate (FWER), the probability of making one or more type I errors. FWER approach is usually considered too stringent for biological experiments. Alternatively, the k-FWER approach controls the probability of making k or more type I errors. Alternatively we can control the false discovery rate (FDR) (Benjamini and Hochberg, 1995), which is the expected proportion of type I errors. A key step in error control is to estimate the null distribution of the test statistics. It is pointed out that the test statistics in large scale testings may not accurately follow the theoretical null distribution (Efron, 2004). The large number of genes/transcripts in genome/transcriptome studies renders the possibility of estimating the density of the null distribution. For microarray experiments based on t-tests a normal mixture model is employed to estimate the empirical null distribution and two methods are proposed: maximum likelihood and mode matching (Efron, 2004). Adopting the same normal mixture model, Jin and Cai (2007) propose a method to estimate the empirical null based on characteristic functions.

Previous methods in (Efron, 2004; Jin and Cai, 2007) are based on a normal mixture model in application to microarray data. However in many other problems such as RNA-seq, the Chi-squared tests are commonly used (Li et al., 2012). The goal in many of these RNA-seq experiments is to find the genes that are significantly different between two (or more) conditions. Further, it is believed *a priori* that many of the genes are unchanged between the two conditions suggesting that many of the statistics should follow the null distribution. We show by four RNA-seq examples that the test statistics in these Chi-squared tests do not match the theoretical null distribution well. The first example is the B-cells RNA-Seq data (Cheung et al., 2010) containing transcriptome profiles of 17 male and 24 female immortalized B-cells samples. We perform Chi-squared tests to compare males and females by R package PoissonSeq (Li et al., 2012) and obtain 9688 score statistics. For two group comparison, the test statistics follows χ^2_1 theoretically under the null

[*] Corresponding authors: Jeffrey C. Miecznikowski - jcm38@buffalo.edu; Song Liu - song.liu@roswellpark.org; Jianmin Wang - jianmin.wang@roswellpark.org

hypothesis (Figure 1(a)). The data matches the the theoretical null fairly well, although it shows a heavy tail, i.e. 6% of the test statistics are greater than the 99% percentile of χ_1^2. Next we look at the prostate cancer data (Ren et al., 2012). The data contains the RNA-seq transcriptome profiles of 11 primary prostate cancer and 12 normal samples. We compare primary prostate cancer with normal tissue samples and obtain 33575 Chi-squared test statistics by PoissonSeq (Figure 1(b)). The data is over-dispersed compared to χ_1^2. One percent of the test statistics have very large values (> 50), making it a much heavier tail than χ_1^2. The third dataset we look at is the LNCap (lymph node carcinoma of the prostate) cells data (Li et al., 2008), which compare 3 androgen treated LNCap cells samples with 4 control samples. The 17162 test statistics obtained by PoissonSeq for the LNCap vs. control samples show a similar over-dispersion from χ_1^2 (Figure 1(c)). The fourth dataset is a human liver and kidney data (Marioni et al., 2008), which contains 7 technical replicates of a human liver sample and 7 technical replicates of a human kidney sample. The test statistics from a liver vs. kidney comparison are much more dispersed than χ_1^2 (Figure 1(d)). On the surface, these examples suggest that many of the statistics do not follow the null distribution. Hence, we propose new methods to empirically estimate null distributions.

The Chi-squared tests are also used for testing single nucleotide polymorphisms (SNPs) in genome-wide association studies (Schwartzman, 2008). Schwartzman (2008) has derived a method based on mode matching (MM) to estimate the empirical null distribution in general exponential families. In this manuscript, we propose the Gamma mixture model for the Chi-squared tests in large scale testing. We derive two estimation procedures based on maximum likelihood (MLE) and a characteristic function (CF) approach to estimate the null parameters in the Gamma mixture model. We also propose an adjustment approach based on local polynomial regression for the CF method to achieve better performance. We will discuss the MLE method in Section 2, our implementation of the MM method for Gamma distribution in Section 3 and the methods based on CF in Section 4. In Section 5 we conduct simulation studies to investigate the numerical performance of our proposed methods. In Section 6 we apply our methods to the four previously described RNA-seq datasets. We end with a discussion and conclusion.

2 Estimating empirical null by maximum likelihood

For all of the methods to estimate empirical null distributions we assume the test statistics follow a mixture model distribution. Mixture models have been commonly used to identify the empirical null distribution for FDR analysis (Efron, 2004) and k-FWER analyses (Miecznikowski and Gaile, 2014). To form the mixture model, we let $x_i, i = 1, \ldots, n$ be the collection of test statistics of n independent hypotheses testing and let f be the density of x_i. Then f can be considered as a mixture of two parts, a large proportion (p_0) of the null density f_0 and a small proportion ($p_1 = 1 - p_0$) of the non-null density f_1,

$$f = p_0 f_0 + p_1 f_1 \,. \tag{1}$$

Efron (2004) describes the framework for maximum likelihood method to estimate the null distribution of normal mixtures, i.e. f_0 belongs to the normal family. The mixture model parameter estimation is via a straightforward maximization of a likelihood derived from a normal mixture model. We extend the MLE method to Gamma mixture model in which f_0 is a Gamma density. Let $S = \{s_1, \ldots, s_n\}$ be the collection of n Chi-squared test statistics in an RNA-seq experiments. To estimate the empirical null, we adopt the mixture model in (1) and assume f_0 belongs to the Gamma family with shape parameter k_0 and scale parameter θ_0,

$$f_0(x|k_0, \theta_0) = \frac{x^{k_0-1} e^{-x/\theta_0}}{\theta_0^{k_0} \Gamma(k_0)} \,, \tag{2}$$

where $\Gamma()$ is the Gamma function. The goal throughout the remainder of the manuscript is estimate k_0 and θ_0. Note this null family also includes the Chi-squared family. Specifically, when $2k_0$ is an integer the Gamma distribution can be considered as a scaled Chi-squared distribution with degree of freedom $2k_0$. In

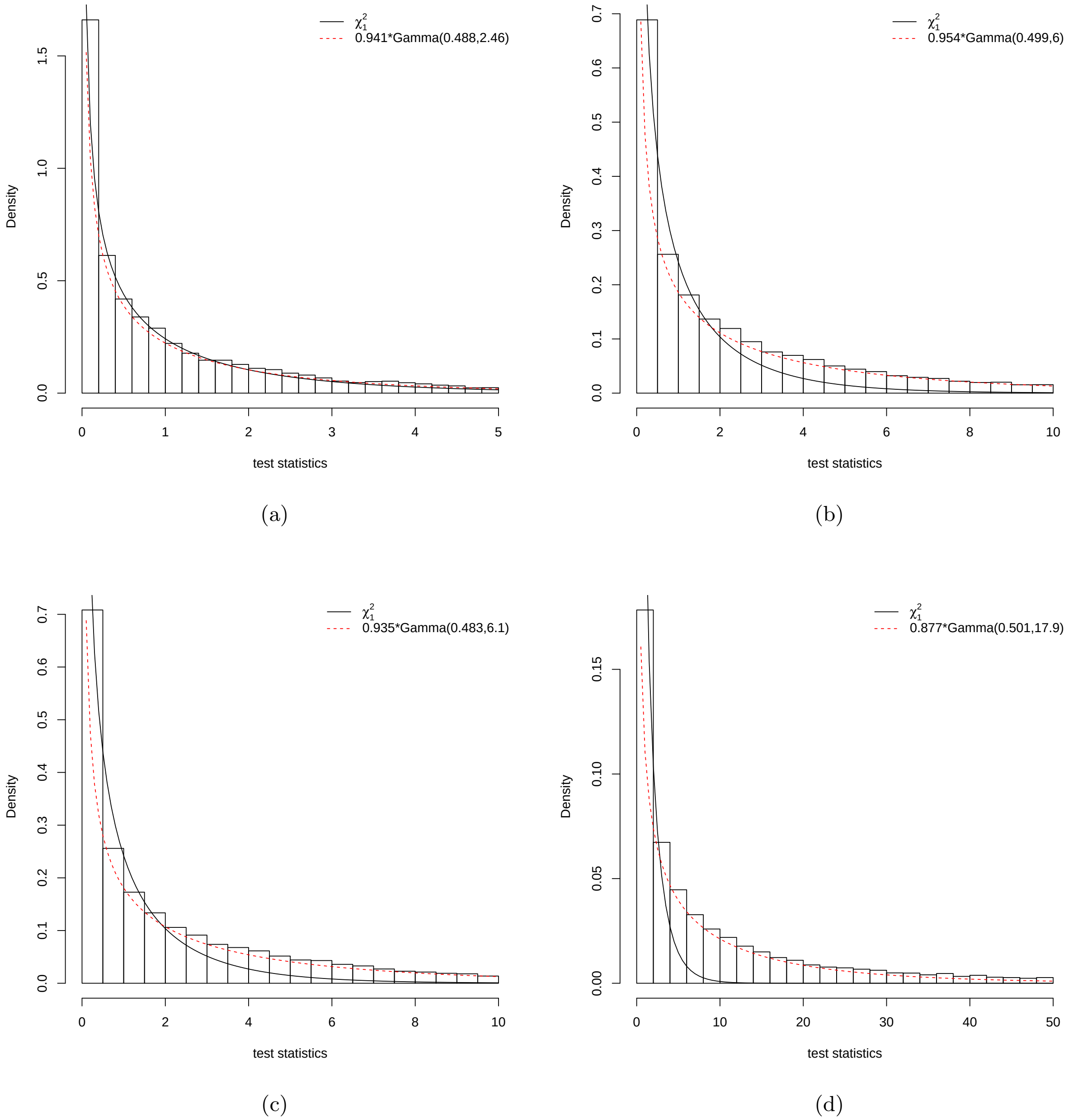


Figure 1: The Chi-squared test statistics for B-cells data (a), prostate cancer data (b), LNCaP cells data (c) and liver and kidney data (d). The solid curve is the density of χ^2_1, and the dashed curve is $\hat{p}_0\hat{f}_0$ obtained by MLE (see Section 2).

particular, $Gamma(\frac{k_0}{2}, 2)$ is the same as Chi-squared distribution with k_0 degrees of freedom. To estimate k_0 and θ_0 by maximum likelihood, we must make a zero assumption in interval A_0

$$p_1 f_1(x) = 0 \quad \forall\, x \in A_0 \,, \tag{3}$$

where A_0 is an interval $(0, q)$ for sample quantile q, e.g. in this manuscript we let q be the 3rd quartile. This zero assumption is necessary for identifiably of the mixture model. For any $s_i \in A_0$, it follows a truncated Gamma distribution:

$$f_{s_i}(x) = \frac{f_0(x)}{F_\Gamma(q|k_0, \theta_0)} \quad \forall\, x \in A_0 \,, \tag{4}$$

where $F_\Gamma(\cdot|k, \theta)$ is the CDF of Gamma distribution with shape k_0 and scale θ_0. We fit s_i within A_0 to (4) by maximum likelihood to obtain the estimates $\hat{k}_0$ and $\hat{\theta}_0$, which can be done by R package fitdistrplus (Delignette-Muller et al., 2014),

$$(\hat{k}_0, \hat{\theta}_0) = \text{argmax}_{k_0, \theta_0} \sum_{j, s_j < q} \log \Big[f_0(s_j | k_0, \theta_0) \Big] \,. \tag{5}$$

In addition, let n_0 be the number of statistics in $(0, q)$, then p_0 can be estimated by

$$\hat{p}_0 = \frac{n_0}{n F_\Gamma(q|\hat{k}_0, \hat{\theta}_0)} \,. \tag{6}$$

Thus the estimators in (5) and (6) will be our MLE estimators of the parameters in the null distribution and the proportion of the null effect.

3 Estimating empirical null by mode matching

Schwartzman (2008) proposes the mode matching (MM) method to estimate the parameters in general exponential families with the form,

$$f_0(x) = g_0(x) \exp(T(x)^T \eta - m(\eta)) \,, \tag{7}$$

where $g_0(x)$ is the base density, $T(x)$ is the sufficient statistic, η is the vector of canonical parameters, and $m(\eta)$ is the cumulant generating function. In the mode matching method the test statistics are partitioned into B bins of equal width d. Denote X_b as the bth bin, and let y_b be the number of test statistics in X_b ($y_b = \#\{x_i \in X_b\}$), c_b be the center point of X_b and v_b be the expected value of y_b, then

$$v_b \approx n d p_0 f_0(c_b) \quad \forall\, c_b \in A_0 \,, \tag{8}$$

where A_0 is the zero assumption interval same as (3). Thus

$$\log(v_b) = T(c_b)^T \eta + \log p_0 - m(\eta) + \log(n d g_0(c_b)) \quad \forall\, c_b \in A_0 \,. \tag{9}$$

Assume a Poisson model for y_b that

$$y_b \overset{\text{ind}}{\sim} \text{Poi}(v_b) \,. \tag{10}$$

Fitting (10) by Poisson regression yields the estimate $\hat{\eta}$ and $\hat{p}_0$.

We have implemented the MM method for our Gamma mixture model in R. For Gamma distribution with parameter k_0 and θ_0, $g_0(x) = 1$, $T(x) = (x, \log x)$, $\eta = (-\theta_0^{-1}, k_0 - 1)$ and $m(\eta) = k_0 \log \theta_0 + \log \Gamma(k_0)$. Therefore we fit the following generalized linear model with Poisson distribution by R function glm.

$$\log(v_b) = \beta_0 + \beta_1 c_b + \beta_2 \log c_b \quad \forall\, c_b \in A_0 \,. \tag{11}$$

We obtain $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, so that the moment matching estimators for the parameters in (2) are

$$\begin{aligned} \dot{k}_0 &= \hat{\beta}_2 + 1\,, \\ \dot{\theta}_0 &= -\hat{\beta}_1^{-1}, \\ \dot{p}_0 &= \frac{e^{\hat{\beta}_0}\Gamma(\dot{k}_0)\dot{\theta}_0^{\dot{k}_0}}{nd}\,. \end{aligned} \tag{12}$$

As discussed in (Schwartzman, 2008), a source of bias of the MM method is the approximation in (8). By a Taylor series expansion, we have

$$v_b - ndp_0f_0(c_b) \approx \frac{f_0''(c_b)}{24}nd^4 \quad \forall\, c_b \in A_0\,, \tag{13}$$

For a Gamma distribution,

$$f_0''(x) = -\theta_0^{-k_0}\Gamma(k_0)^{-1}e^{-\frac{x}{\theta_0}}\Big\{(k_0-2)(k_0-1)x^{k_0-3} + 2\theta_0^{-1}(k_0-1)x^{k_0-2} + \theta_0^{-2}x^{k_0-1}\Big\}. \tag{14}$$

The second derivative $f_0''(x)$ is unbounded at $x = 0$ for $k_0 < 3$ except $k_0 = 1$, meaning the error of the approximation in (8) large for the bins near 0.

4 Estimating empirical null by characteristic functions

Following the mixture model by Efron (2004), Jin and Cai (2007) propose a method to estimate the null distribution via the characteristic functions for a normal distribution. The authors show the CF estimators are consistent and perform better than the MLE method when the proportion of the non-null effect is not small, e.g. $p_0 < 0.9$. Their method is based on interpreting the mixture model in (1) in the form of the characteristic functions,

$$\psi(t) = p_0\psi_0(t) + p_1\psi_1(t)\,, \tag{15}$$

where $\psi_0(t)$, $\psi_1(t)$ and $\psi(t)$ are the characteristic functions of the null, non-null and all genes respectively. Given a sample $x_1, \dots, x_n$, the empirical characteristic function $\psi_n(t)$ is defined as,

$$\psi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itx_j}\,. \tag{16}$$

In large scale testings $\psi_n(t) \approx \psi(t)$ and at properly chosen t_o such that $p_0\psi_0(t_o) >> p_1\psi_1(t_o)$, we have

$$\psi_n(t_o) \approx \psi(t_o) \approx p_0\psi_0(t_o)\,. \tag{17}$$

This provides the basis for the CF estimators. To determine an appropriate t_o, Jin and Cai (2007) propose the following criterion that is adaptive to the magnitude of $\psi_n(t)$,

$$t_o = \hat{t}(\gamma) = \inf\{t : |\psi_n(t)| = n^{-\gamma}, 0 < t < \log n\}\,, \tag{18}$$

where γ is a tuning parameter.

4.1 The CF estimator for Gamma mixture model

Inspired by the work of Jin and Cai (2007), we propose the CF method to estimate the null parameters for Gamma mixture model. We derive two functionals,

$$\begin{aligned} \theta(f(t)) &= \frac{-|f(t)| \cdot \frac{d}{dt}|f(t)|}{t[\mathrm{Re}(f(t))\mathrm{Im}(f'(t)) - \mathrm{Re}(f'(t))\mathrm{Im}(f(t))]}\,, \\ k(f(t)) &= -t|f(t)|^{-1}\Big\{\frac{|f(t)|^{-2}\big[\mathrm{Re}(f(t))\mathrm{Im}(f'(t)) - \mathrm{Re}(f'(t))\mathrm{Im}(f(t))\big]^2}{\frac{d}{dt}|f(t)|} + \frac{d}{dt}|f(t)|\Big\}\, \textit{where}, \end{aligned} \tag{19}$$

evaluating these functionals at the characteristic function of $Gamma(k_0, \theta_0)$ ($\psi_0(t) = (1 - i\theta_0 t)^{-k_0}$) yields the following solutions for θ_0 and k_0 ,

$$\begin{aligned} \theta(p_0\psi_0(t)) &= \theta_0\,, \\ k(p_0\psi_0(t)) &= k_0\,. \end{aligned} \tag{20}$$

This situation is summarized in the following theorem:

Theorem 1. *If we assume the mixture model in* (1) *with f_0 specified in* (2) *and consider the characteristic function for a Gamma distribution $\psi_0(t) = (1 - i\theta_0 t)^{-k_0}$ then evaluating the functionals in* (19) *at $p_0\psi_0(t)$ is such that*

$$\begin{aligned} \theta(p_0\psi_0(t)) &= \theta_0\,, \\ k(p_0\psi_0(t)) &= k_0\,. \end{aligned} \tag{21}$$

Proof. Let $x = \tan^{-1}(\theta t)$, then $\frac{dx}{dt} = \frac{\theta}{1+\theta^2t^2}$,

$$\begin{aligned} \psi_0(t) &= (1 - i\theta t)^{-k} \\ &= [\sqrt{1+\theta^2t^2}(\text{cos}x - i\text{sin}x)]^{-k} \\ &= (1+\theta^2t^2)^{-\frac{k}{2}}[\text{cos}(-x) + i\text{sin}(-x)]^{-k} \\ &= (1+\theta^2t^2)^{-\frac{k}{2}}[\text{cos}(kx) + i\text{sin}(kx)]\,, \end{aligned} \tag{22}$$

$$\begin{aligned} |\psi_0(t)| &= (1+\theta^2t^2)^{-\frac{k}{2}} \\ \frac{d}{dt}|\psi_0(t)| &= -\frac{k}{2}(1+\theta^2t^2)^{-\frac{k}{2}-1}2\theta^2t \\ &= -(1+\theta^2t^2)^{-\frac{k}{2}-1}k\theta^2t\,, \end{aligned} \tag{23}$$

$$\begin{aligned} \text{Re}(\psi_0(t))\text{Im}(\psi_0'(t)) &= (1+\theta^2t^2)^{-\frac{k}{2}}\text{cos}(kx)\Big[-\frac{k}{2}(1+\theta^2t^2)^{-\frac{k}{2}-1}2\theta^2t\cdot\text{sin}(kx) + (1+\theta^2t^2)^{-\frac{k}{2}}\text{cos}(kx)k\frac{\theta}{1+\theta^2t^2}\Big] \\ &= (1+\theta^2t^2)^{-k-1}k\theta\text{cos}(kx)[-\text{tan}x\cdot\text{sin}(kx) + \text{cos}(kx)] \\ &= (1+\theta^2t^2)^{-k-1}k\theta\frac{\text{cos}(kx)}{\text{cos}x}[\text{cos}x\cdot\text{cos}(kx) - \text{sin}x\cdot\text{sin}(kx)] \\ &= (1+\theta^2t^2)^{-k-1}k\theta\frac{\text{cos}(kx)}{\text{cos}x}\text{cos}(kx+x)\,, \end{aligned} \tag{24}$$

$$\begin{aligned} \text{Re}(\psi_0'(t))\text{Im}(\psi_0(t)) &= (1+\theta^2t^2)^{-\frac{k}{2}}\text{sin}(kx)\Big[-\frac{k}{2}(1+\theta^2t^2)^{-\frac{k}{2}-1}2\theta^2t\cdot\text{cos}(kx) - (1+\theta^2t^2)^{-\frac{k}{2}}\text{sin}(kx)k\frac{\theta}{1+\theta^2t^2}\Big] \\ &= -(1+\theta^2t^2)^{-k-1}k\theta\text{sin}(kx)[\text{tan}x\cdot\text{cos}(kx) + \text{sin}(kx)] \\ &= -(1+\theta^2t^2)^{-k-1}k\theta\frac{\text{sin}(kx)}{\text{cos}x}[\text{sin}x\cdot\text{cos}(kx) + \text{cos}x\cdot\text{sin}(kx)] \\ &= -(1+\theta^2t^2)^{-k-1}k\theta\frac{\text{sin}(kx)}{\text{cos}x}\text{sin}(kx+x)\,. \end{aligned} \tag{25}$$

Therefore,

$$\begin{aligned} \text{Re}(\psi_0(t))\text{Im}(\psi_0'(t)) - \text{Re}(\psi_0'(t))\text{Im}(\psi_0(t)) &= (1+\theta^2t^2)^{-k-1}k\theta\frac{\text{cos}(kx+x)\cdot\text{cos}(kx) + \text{sin}(kx+x)\cdot\text{sin}(kx)}{\text{cos}x} \\ &= (1+\theta^2t^2)^{-k-1}k\theta\frac{\text{cos}x}{\text{cos}x} \\ &= (1+\theta^2t^2)^{-k-1}k\theta\,. \end{aligned} \tag{26}$$

$$
\begin{aligned}
\theta(\psi_0(t)) &= \frac{-|\psi_0(t)| \cdot \frac{d}{dt}|\psi_0(t)|}{t[\mathrm{Re}(\psi_0(t))\mathrm{Im}(\psi_0'(t)) - \mathrm{Re}(\psi_0'(t))\mathrm{Im}(\psi_0(t))]} \\
&= \frac{-(1+\theta^2t^2)^{-\frac{k}{2}}\Big[-(1+\theta^2t^2)^{-\frac{k}{2}-1}k\theta^2t\Big]}{(1+\theta^2t^2)^{-k-1}k\theta t} \\
&= \theta\,,
\end{aligned}
\tag{27}
$$

$$
\begin{aligned}
k(\psi_0(t)) &= -t|\psi_0(t)|^{-1}\Big\{\frac{|\psi_0(t)|^{-2}\big[\mathrm{Re}(\psi_0(t))\mathrm{Im}(\psi_0'(t)) - \mathrm{Re}(\psi_0'(t))\mathrm{Im}(\psi_0(t))\big]^2}{\frac{d}{dt}|\psi_0(t)|} + \frac{d}{dt}|\psi_0(t)|\Big\} \\
&= -t(1+\theta^2t^2)^{\frac{k}{2}}\Big\{\frac{(1+\theta^2t^2)^{k}\big[(1+\theta^2t^2)^{-k-1}k\theta\big]^2}{-(1+\theta^2t^2)^{-\frac{k}{2}-1}k\theta^2t} - (1+\theta^2t^2)^{-\frac{k}{2}-1}k\theta^2t\Big\} \\
&= t(1+\theta^2t^2)^{\frac{k}{2}}\Big[(1+\theta^2t^2)^{-\frac{k}{2}-1}kt^{-1} + (1+\theta^2t^2)^{-\frac{k}{2}-1}k\theta^2t\Big] \\
&= (1+\theta^2t^2)^{-1}(k + k\theta^2t^2) \\
&= k\,.
\end{aligned}
\tag{28}
$$

End of proof.

At properly chosen t_o that satisfies (17), plugging $\psi_n(t)$ in (19) yields the CF estimates of θ_0 and k_0,

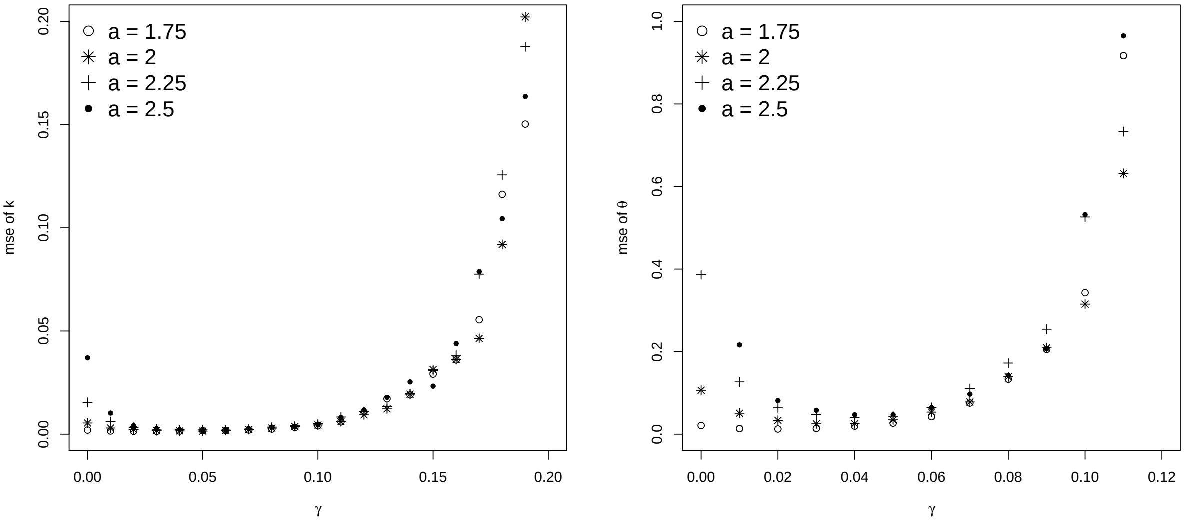
$$
\begin{aligned}
\tilde{\theta}_0 &= \theta(\psi_n(t_o))\,, \\
\tilde{k}_0 &= k(\psi_n(t_o))\,.
\end{aligned}
\tag{29}
$$

□

4.2 Choice of γ and t_o

We use the criterion in (18) to determine t_o. As t increases from 0, the second approximation in (17) becomes more accurate but the first one becomes less accurate, hence t_o must be chosen from an interval where both approximations are reasonable. For our estimator of θ in (19), there is another issue that the denominator $\mathrm{Re}(\psi_n(t))\mathrm{Im}(\psi_n'(t)) - \mathrm{Re}(\psi_n'(t))\mathrm{Im}(\psi_n(t))$ approaches 0 as t increases, which will make our estimator unstable. Therefore the candidate interval for t_o should be restricted in a smaller interval. We investigate the effect of γ in (18) via a simulation study. Let $n = 10000$, $p_0 = 0.9$, $k_0 = 1$, $\theta_0 = 3$ and $a = 1.75, 2, 2.25, 2.5$, where we simulate data as follows,
1) generate np_0 null statistics from $f_0 = Gamma(k_0, \theta_0)$,
2) generate $n(1-p_0)$ pairs of (k_j, σ_j) with k_j from $Unif(0.5, 1)$ and σ_j from $Unif(a, a+0.5)$,
3) let $\theta_j = \sigma_j^2$ and generate a statistic from $Gamma(k_j, \theta_j)$ for each pair of (k_j, θ_j).

We compute $\tilde{k}$ and $\tilde{\theta}$ at different values of γ and examine the mean squared error (mse). Based on 1000 simulations, the mse results suggest that the best choice of γ is in an interval around 0.05. The errors are fairly consistent in this interval and to different values of a (Figure 2). Therefore we choose $\gamma = 0.05$ throughout the rest of the manuscript, i.e. $t_o = \hat{t}(0.05)$ in (18).



(a) (b)

Figure 2: The mse for $\tilde{k}_0$ and $\tilde{\theta}_0$ at different γ. Some points are not shown since the mse is out of range.

4.3 Smoothed estimators for the CF method

Since the estimator in (19) involves the derivative of $\psi_n(t)$, we also need the approximation that

$$\psi_n'(t) \approx \psi'(t)\,. \tag{30}$$

For data following a Gamma mixture model the approximation in (30) is not accurate even under moderately large sample size. To illustrate this problem, we simulate $n = 10000$ statistics from the mixture density in (1) where $f_0 = Gamma(1, 2)$, $f_1 = Gamma(2, 5)$ and $p_0 = 0.9$. Although $\psi_n(t)$ is very close to $\psi(t)$, its derivative is variable (Figure 3) and thus increases the variance of our estimator. To address this issue, we propose a smoothed version of the CF estimators based on local polynomial regression (Wand and Jones, 1994). We take a large number of equally spaced points t_i from 0 to $\log N$. Then we fit a local polynomial $\psi_s(t)$ of order J to the pairs of t_i and $\psi_n(t_i)$, which is implemented in R package KernSmooth (Wand, 2013). In particular, at given t and bandwidth h, the algorithm minimizes the following quantity over β_j,

$$\sum_i \phi\big(\frac{t_i - t}{h}\big)\big(\psi_n(t_i) - \sum_{j=0}^{J} \beta_j t^j\big)^2\,, \tag{31}$$

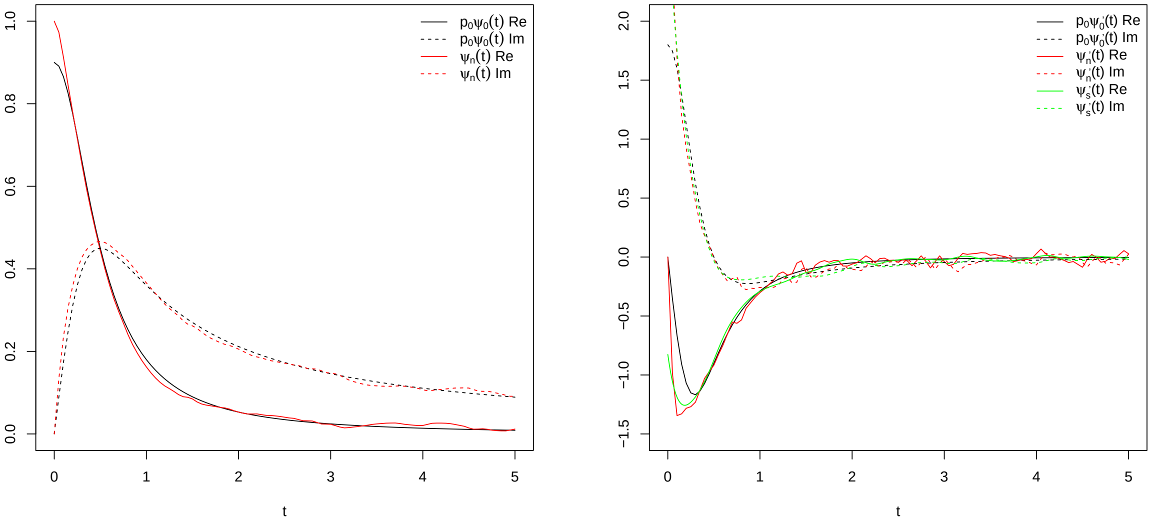
where $\phi(\frac{t_i - t}{h})$ is the normal kernel. We obtain $\hat{\beta}_j$, so that

$$\begin{aligned}\psi_s(t) &= \sum_{j=0}^{J} \hat{\beta}_j t^j \\ \psi_s'(t) &= \sum_{j=1}^{J} j\hat{\beta}_j t^{j-1}\end{aligned} \tag{32}$$

The smoothed derivative $\psi_s'(t)$ is much closer to $p_0\psi'(t)$ (Figure 3). Replacing $\psi_n'(t)$ by $\psi_s'(t)$ of the CF estimator in (29) yields the smoothed CF estimators, denoted by $\tilde{k}^*$ and $\tilde{\theta}^*$,

$$\tilde{\theta}^* = -\frac{\mathrm{Re}(\psi_n(t))\mathrm{Re}(\psi_s'(t)) + \mathrm{Im}(\psi_s'(t))\mathrm{Im}(\psi_n(t))}{t[\mathrm{Re}(\psi_n(t))\mathrm{Im}(\psi_s'(t)) - \mathrm{Re}(\psi_s'(t))\mathrm{Im}(\psi_n(t))]},$$

$$\tilde{k}^* = -t|\psi_n(t)|^{-2}\Big\{\frac{\big[\mathrm{Re}(\psi_n(t))\mathrm{Im}(\psi_s'(t)) - \mathrm{Re}(\psi_s'(t))\mathrm{Im}(\psi_n(t))\big]^2}{\mathrm{Re}(\psi_n(t))\mathrm{Re}(\psi_s'(t)) + \mathrm{Im}(\psi_s'(t))\mathrm{Im}(\psi_n(t))} + \mathrm{Re}(\psi_n(t))\mathrm{Re}(\psi_s'(t)) + \mathrm{Im}(\psi_s'(t))\mathrm{Im}(\psi_n(t))\Big\}. \tag{33}$$



(a)

(b)

Figure 3: Comparison of $p_0\psi_0(t)$ (black) and $\psi_n(t)$ (red) (Panel (a)) and their derivatives (Panel (b)). The solid and dashed curves represent the real parts and imaginary parts of the functions. The green curves in Panel (b) are the smooth derivatives obtained by local polynomial regression. The smoothed derivative is much closer to the true derivative.

We investigate the accuracy of the smoothed derivatives $\psi_s'(t)$ under different polynomial order J and kernel bandwidth h via simulations. To do so, we fix the parameters in the non-null distribution so that $\psi'(t)$ can be derived numerically. In this simulation we let $k_0 = 1$ and $\theta_0 = 3$ and set $n = 10000$, $p_0 = 0.9$, $k_1 = 0.8$ and $\theta_1 = 6$. We simulate np_0 statistics from $Gamma(k_0, \theta_0)$ and $n(1-p_0)$ statistics from $Gamma(k_1, \theta_1)$. The theoretical derivative is

$$\psi'(t_o) = ip_0k_0\theta_0(1 - i\theta_0t_o)^{-k_0-1} + i(1-p_0)k_1\theta_1(1 - i\theta_1t_o)^{-k_1-1}. \tag{34}$$

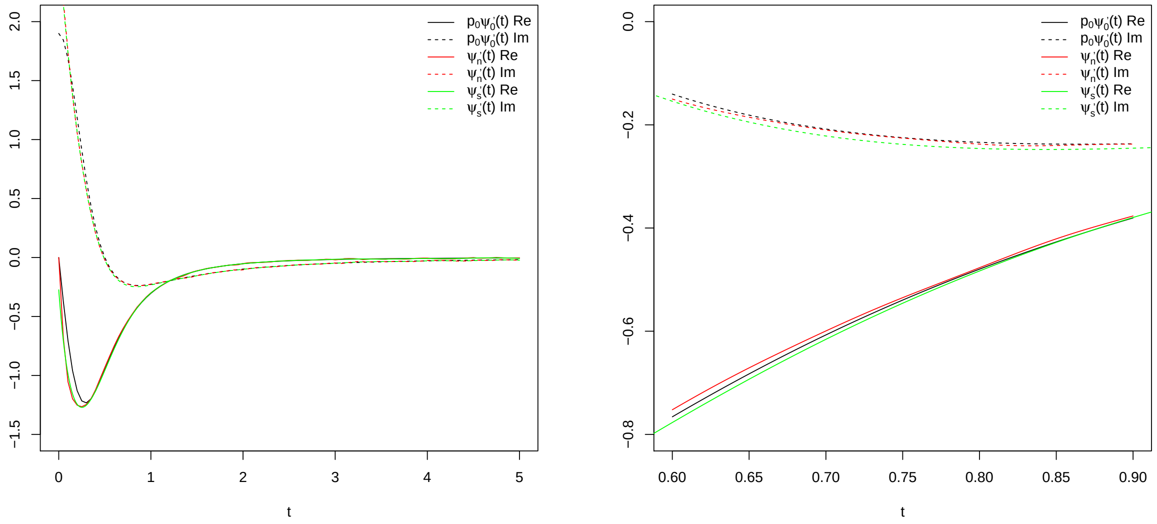
We obtain the smoothed derivative $\psi_s'(t_o)$ and compute the error $\psi_s'(t_o) - \psi'(t_o)$. The real and imaginary parts of the errors are divided by $\mathrm{Re}(\psi'(t_o))$ and $\mathrm{Im}(\psi'(t_o))$ respectively so they can be compared on the same scale. Based on 1000 simulations, results show that the errors of the imaginary part are significantly higher than the real part (Table 1). We choose $J = 4$ and $h = 0.2$ for our smoothed estimator as it has the smallest error for imaginary part and also fairly small error for the real part.

Re	J=1	J=2	J=3	J=4	J=5
$h = 0.10$	1.15×10^{-3}	1.16×10^{-3}	1.10×10^{-3}	1.10×10^{-3}	9.39×10^{-4}
$h = 0.20$	3.72×10^{-3}	7.57×10^{-4}	7.13×10^{-4}	5.06×10^{-3}	2.11×10^{-3}
$h = 0.30$	4.09×10^{-2}	5.02×10^{-4}	1.76×10^{-3}	1.42×10^{-3}	5.50×10^{-3}
$h = 0.40$	1.11×10^{-1}	1.49×10^{-3}	2.57×10^{-3}	1.46×10^{-3}	2.66×10^{-3}
$h = 0.50$	1.93×10^{-1}	7.49×10^{-3}	2.60×10^{-3}	4.06×10^{-3}	7.39×10^{-4}
Im					
$h = 0.10$	0.342	0.342	0.022	0.0224	0.0268
$h = 0.20$	1.534	2.735	0.178	0.0137	0.0148
$h = 0.30$	1.070	3.771	1.318	0.0620	0.0150
$h = 0.40$	0.616	3.226	2.651	0.0479	0.0334
$h = 0.50$	0.383	2.326	3.294	1.295	0.164

Table 1: The mse of $\psi'_s(t_o)$ (real part and imaginary part) at different polynomial order J and kernel bandwidth h.

4.4 Smoothing

The smoothing strategy described in Section 4.3 yields an improved estimate of $\psi'(t)$ under moderate sample size n or when p_0 is not close to 1. When n is sufficiently large or p_0 is very close to 1, $\psi'_n(t)$ is very smooth and the smoothing may not help and may even increase error slightly. To show this, we simulate $n = 640000$ statistics from the mixture density in 1 where $f_0 = Gamma(1, 2)$, $f_1 = Gamma(2, 5)$ and $p_0 = 0.95$. The imaginary part of the non-smoothed derivative $\psi'_n(t)$ is closer to the true derivative than the smoothed $\psi'_s(t)$, while the real part errors of the two are similar (Figure 4). We provide further advice in the Discussion section for when to consider using the smoothed version of the characteristic function estimators.



(a)

(b)

Figure 4: The issue of over-smoothing when n is very large and p_0 is close to 1. The solid and dashed curves represent the real parts and imaginary parts of the functions. Panel (b) is a zoom-in of (a) around t_o. The imaginary part of derivative $\psi_n'(t)$ has less error than the smoothed $\psi_s'(t)$, while the errors of the real part of the two are similar.

5 Simulations

We conduct four simulation studies to compare the MLE, CF, smoothed CF and MM method. The R code is available at *http://sphhp.buffalo.edu/biostatistics.html*. For the MLE and MM, we let $A_0 = (0, q)$ where q is the 3rd sample quartile and let the bin width $d = 0.1$ for MM. For CF and smoothed CF we choose $\gamma = 0.05$ and set $J = 4$, $h = 0.2$ for smoothed CF. In Simulation 1 we consider the setting that $f_0 = Gamma(0.5, 2)$ which is a standard Chi-squared distribution with degree of freedom 1. In Simulations 2 - 4 we let $f_0 = Gamma(1, 3)$ and for the non-null density we consider three settings 1) $k_1 = 2$ and $\theta_1 = 15$ (Simulation 2) under this setting the zero assumption in (3) is valid. In Simulation 3 we set $k_1 = 0.7$ and $\theta_1 = 10$, which violates the zero assumption slightly and in Simulation 4 the zero assumption is slightly violated and f_1 is not identical for non-null statistics. The difference in the null distribution between Simulations 2 and 3 are shown in Figure 5, where $(1 - p_0)f_1 \approx 0$ within the 3rd quartile of the data for Simulation 1 while $(1 - p_0)f_1 > 0$ for Simulation 2.

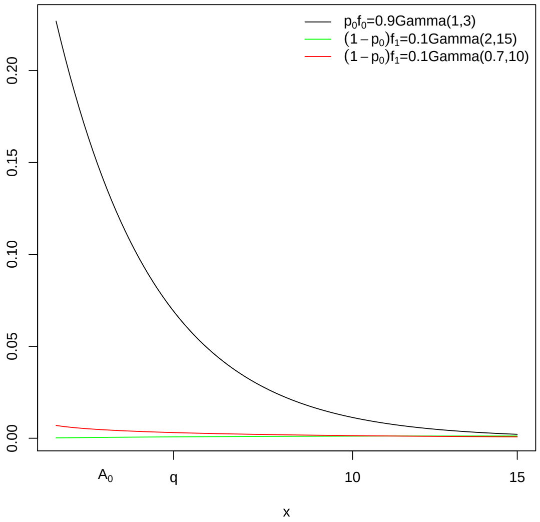


Figure 5: Comparison of $p_0 f_0$ with $(1-p_0)f_1$ for $p_0 = 0.9$ when the zero assumption holds (green) and the zero assumption is violated (red). The interval $(0, q)$ is the zero assumption interval A_0.

5.1 Simulation 1

We let $f_0 = Gamma(0.5, 2)$ and $f_1 = Gamma(2, 6)$. For $n = 10000$ and $p_0 = 0.8, 0.85, 0.9, 0.95$ we simulate np_0 statistics from f_0 and $n(1-p_0)$ statistics from f_1. We use the mse of 1000 simulations to summarize our results (see Table 2). The MLE estimator has significantly smaller mse than the other three estimators for k_0. MLE and smoothed CF have similar performance on θ_0. The MM method has significantly larger mse for both parameters and the mse increase as p_0 increases, suggesting the MM method has a significant bias for the null parameters when $k_0 = 0.5$.

	$p_0 = 0.80$	0.85	0.90	0.95
$\hat{k}_0$	8.92×10^{-5}	6.00×10^{-5}	6.04×10^{-5}	6.96×10^{-5}
$\tilde{k}_0$	7.19×10^{-3}	6.44×10^{-3}	3.03×10^{-3}	2.49×10^{-3}
$\tilde{k}_0^*$	8.18×10^{-3}	1.17×10^{-3}	5.72×10^{-4}	4.23×10^{-4}
$\dot{k}_0$	3.49×10^{-2}	3.73×10^{-2}	4.08×10^{-2}	4.30×10^{-2}
$\hat{\theta}_0$	1.11×10^{-1}	4.14×10^{-2}	3.05×10^{-2}	3.19×10^{-2}
$\tilde{\theta}_0$	7.82×10^{-1}	6.72×10^{-1}	3.70×10^{-1}	3.38×10^{-1}
$\tilde{\theta}_0^*$	6.78×10^{-2}	3.51×10^{-2}	4.47×10^{-2}	4.55×10^{-2}
$\dot{\theta}_0$	3.45	4.14	5.89	11.9

Table 2: The mse for MLE $(\hat{k}_0, \hat{\theta}_0)$, CF $(\tilde{k}_0, \tilde{\theta}_0)$, smoothed CF $(\tilde{k}_0^*, \tilde{\theta}_0^*)$ and MM $(\dot{k}_0, \dot{\theta}_0)$ for $k_0 = 0.5$, $\theta_0 = 3$, $k_1 = 2$ and $\theta_1 = 6$.

5.2 Simulation 2

In the next simulation we set $k_0 = 1$ so that $f_0 = Gamma(1, 3)$ and $f_1 = Gamma(2, 15)$. For $n = 10000$ and $p_0 = 0.8, 0.85, 0.9, 0.95$ we simulate np_0 statistics from f_0 and $n(1-p_0)$ statistics from f_1. Contrary to Simulation 1, the accuracy of the MM estimator is improved (Table 3) for the k_0 estimate, the mse of MLE and MM are much smaller than the CF and smoothed CF by approximately one to two magnitudes. For θ_0, the CF estimator has the largest mse while the other three estimators have similar performance. The smoothed CF approach reduces the estimation errors under most of the settings, suggesting smoothing is effective when the sample size is moderately large. The mse of CF estimator for k_0 and θ_0 both decrease as p_0 increases. However this is not the case for the other three methods. For k_0, the mse of $\hat{k}_0$ does not decrease, indicating a non-negligible bias for the MLE method. The mse of $\dot{k}_0$ surprisingly increases, suggesting an increase of bias.

	$p_0 = 0.80$	0.85	0.90	0.95
$\hat{k}_0$	3.93×10^{-4}	3.27×10^{-4}	3.87×10^{-4}	3.46×10^{-4}
$\tilde{k}_0$	1.67×10^{-2}	1.16×10^{-2}	9.51×10^{-3}	3.59×10^{-3}
$\tilde{k}_0^*$	2.75×10^{-2}	8.66×10^{-3}	6.09×10^{-4}	2.08×10^{-3}
$\dot{k}_0$	4.57×10^{-4}	4.26×10^{-4}	5.21×10^{-4}	6.22×10^{-4}
$\hat{\theta}_0$	5.74×10^{-2}	3.02×10^{-2}	3.11×10^{-2}	2.13×10^{-2}
$\tilde{\theta}_0$	3.26×10^{-1}	2.16×10^{-1}	1.59×10^{-1}	8.28×10^{-2}
$\tilde{\theta}_0^*$	6.76×10^{-2}	1.46×10^{-2}	2.74×10^{-2}	1.64×10^{-2}
$\dot{\theta}_0$	4.37×10^{-2}	2.16×10^{-2}	2.87×10^{-2}	3.99×10^{-2}

Table 3: The mse for MLE $(\hat{k}_0, \hat{\theta}_0)$, CF $(\tilde{k}_0, \tilde{\theta}_0)$, smoothed CF $(\tilde{k}_0^*, \tilde{\theta}_0^*)$ and MM $(\dot{k}_0, \dot{\theta}_0)$ when the zero assumption holds.

5.3 Simulation 3

Next we consider the scenario that the zero assumption is slightly violated. We let $f_0 = Gamma(1, 3)$ and $f_1 = Gamma(0.7, 10)$. For $n = 10000$ and $p_0 = 0.8, 0.85, 0.9, 0.95$ we simulate np_0 samples from f_0 and $n(1 - p_0)$ samples from f_1. Based on 1000 simulations, results show that smoothed CF has smallest mse for both k and θ_0 when $p_0 \leq 0.9$. Similar to simulation 1, the mse of $\tilde{k}_0^*$ increases when $p_0 = 0.95$, suggesting the issue of over-smoothing. In this case the mse of MLE is the smallest for k_0 and about the same as CF for θ_0.

	$p_0 = 0.80$	0.85	0.90	0.95
$\hat{k}_0$	3.43×10^{-3}	1.78×10^{-3}	7.54×10^{-4}	1.71×10^{-4}
$\tilde{k}_0$	4.86×10^{-3}	2.34×10^{-3}	1.01×10^{-3}	1.64×10^{-4}
$\tilde{k}_0^*$	1.01×10^{-3}	1.08×10^{-4}	2.01×10^{-4}	3.84×10^{-3}
$\dot{k}_0$	2.28×10^{-3}	9.63×10^{-3}	1.93×10^{-4}	5.57×10^{-5}
$\hat{\theta}_0$	2.17×10^{-1}	1.08×10^{-1}	3.81×10^{-2}	9.41×10^{-3}
$\tilde{\theta}_0$	1.26×10^{-1}	4.89×10^{-2}	2.54×10^{-2}	3.79×10^{-3}
$\tilde{\theta}_0^*$	1.15×10^{-1}	5.90×10^{-2}	2.06×10^{-2}	6.20×10^{-4}
$\dot{\theta}_0$	8.18×10^{-2}	2.72×10^{-2}	5.36×10^{-5}	9.66×10^{-3}

Table 4: The mse for MLE $(\hat{k}_0, \hat{\theta}_0)$, CF $(\tilde{k}_0, \tilde{\theta}_0)$, smoothed CF $(\tilde{k}_0^*, \tilde{\theta}_0^*)$ and MM $(\dot{k}_0, \dot{\theta}_0)$ when the zero assumption strictly is slightly violated.

5.4 Simulation 4

In practice it is unlikely that all non-null statistics follow the same distribution, therefore we look into a more realistic setting that allows the non-null samples generated from non-identical distributions. We keep $f_0 = Gamma(1, 3)$. For $p_0 = 0.8, 0.85, 0.9, 0.95$ and $n = 10000, 40000, 160000, 640000$, we simulate n samples in the same fashion as Section 4.2. We first generate np_0 null statistics from $f_0 = Gamma(k_0, \theta_0)$. Then we generate $n(1-p_0)$ pairs of (k_j, σ_j) with k_j from $Unif(0.5, 1)$ and σ_j from $Unif(3.5, 4)$. Let $\theta_j = \sigma_j^2$ and generate a statistic from $Gamma(k_j, \theta_j)$ for each pair of (k_j, θ_j). This setting is similar to simulation 2 such that the zero assumption is slightly violated.

From examining Table 5 all three estimators proposed have significantly better accuracy than MM. The mse of all estimators are about one or two magnitude smaller than MM in most of the settings. Under the same p_0, the mse of MM does not decrease as sample size n increases, which indicates a significant bias, especially when p_0 is not close to 1. The MLE estimate shows a similar but less significant bias, i.e. at $p_0 = 0.8$, the mse of $\hat{k}$ are 3.04×10^{-3}, 3.14×10^{-3}, 2.91×10^{-3}, 2.94×10^{-3} at 4 values of n. On the other hand the error for CF always reduces as n increases, suggesting that it is more consistent than MLE and MM.

Smoothing reduces the estimation errors for the CF method under moderate n and smaller p_0, i.e. at $n = 10000$ and $p_0 = 0.8$, $\tilde{\theta}_0^*$ is two magnitudes better than $\tilde{\theta}_0$ (4.23×10^{-3} vs. 2.11×10^{-1}). Similar to the previous simulations, however, when n is very large or when p_0 is close to 1, over-smoothing can cause increased error, i.e. for $n = 640000$ and $p_0 = 0.95$ $\tilde{k}_0^*$ is two magnitudes worse than $\tilde{k}_0$ (1.12×10^{-2} vs. 1.23×10^{-4}).

Overall the MLE, CF and smoothed CF estimators have similar performance. In particular their performances depend on p_0. When p_0 is not close to 1 smoothed CF is the most accurate for both k and θ, while under large p_0 MLE and CF are better. With p_0 close to 1, MLE is best under moderate n and CF is best under very large n.

n		$p_0 = 0.80$	0.85	0.90	0.95
10^4	$\hat{k}_0$	3.04×10^{-3}	1.97×10^{-3}	9.71×10^{-4}	4.37×10^{-4}
	$\tilde{k}_0$	8.14×10^{-3}	6.13×10^{-3}	2.57×10^{-3}	1.66×10^{-3}
	$\tilde{k}_0^*$	1.96×10^{-3}	7.21×10^{-4}	3.87×10^{-4}	1.81×10^{-3}
	$\dot{k}_0$	2.70×10^{-3}	1.71×10^{-3}	7.95×10^{-4}	5.48×10^{-4}
4×10^4	$\hat{k}_0$	3.14×10^{-3}	1.79×10^{-3}	8.19×10^{-4}	2.94×10^{-4}
	$\tilde{k}_0$	5.86×10^{-3}	2.57×10^{-3}	1.24×10^{-3}	5.22×10^{-4}
	$\tilde{k}_0^*$	2.77×10^{-3}	2.83×10^{-4}	7.61×10^{-4}	4.14×10^{-3}
	$\dot{k}_0$	2.59×10^{-3}	1.05×10^{-3}	4.58×10^{-4}	1.56×10^{-4}
16×10^4	$\hat{k}_0$	2.91×10^{-3}	1.63×10^{-3}	7.08×10^{-4}	1.89×10^{-4}
	$\tilde{k}_0$	3.54×10^{-3}	1.89×10^{-3}	8.48×10^{-4}	2.82×10^{-4}
	$\tilde{k}_0^*$	1.51×10^{-3}	1.20×10^{-4}	3.12×10^{-3}	4.18×10^{-3}
	$\dot{k}_0$	2.55×10^{-3}	8.44×10^{-4}	1.54×10^{-4}	4.42×10^{-5}
64×10^4	$\hat{k}_0$	2.94×10^{-3}	1.61×10^{-3}	7.36×10^{-4}	1.87×10^{-4}
	$\tilde{k}_0$	2.65×10^{-3}	1.46×10^{-3}	5.67×10^{-4}	1.23×10^{-4}
	$\tilde{k}_0^*$	7.50×10^{-5}	9.23×10^{-4}	2.56×10^{-3}	1.12×10^{-2}
	$\dot{k}_0$	2.53×10^{-3}	8.77×10^{-4}	1.27×10^{-4}	2.08×10^{-5}
10^4	$\hat{\theta}_0$	2.50×10^{-1}	1.47×10^{-1}	7.83×10^{-2}	3.34×10^{-2}
	$\tilde{\theta}_0$	2.11×10^{-1}	1.54×10^{-1}	7.58×10^{-2}	3.82×10^{-2}
	$\tilde{\theta}_0^*$	4.23×10^{-3}	1.18×10^{-2}	1.93×10^{-2}	4.39×10^{-2}
	$\dot{\theta}_0$	1.33×10^{-1}	8.97×10^{-2}	5.25×10^{-2}	4.41×10^{-2}
4×10^4	$\hat{\theta}_0$	2.37×10^{-1}	1.28×10^{-1}	5.87×10^{-2}	2.22×10^{-2}
	$\tilde{\theta}_0$	1.52×10^{-1}	6.44×10^{-2}	3.42×10^{-2}	1.57×10^{-2}
	$\tilde{\theta}_0^*$	2.51×10^{-2}	2.46×10^{-3}	5.79×10^{-3}	2.45×10^{-2}
	$\dot{\theta}_0$	1.30×10^{-1}	3.77×10^{-2}	2.42×10^{-2}	2.01×10^{-2}
16×10^4	$\hat{\theta}_0$	2.16×10^{-1}	1.15×10^{-1}	4.54×10^{-2}	1.28×10^{-2}
	$\tilde{\theta}_0$	8.79×10^{-2}	5.22×10^{-2}	1.83×10^{-2}	6.96×10^{-3}
	$\tilde{\theta}_0^*$	6.79×10^{-2}	1.31×10^{-2}	9.46×10^{-4}	1.04×10^{-3}
	$\dot{\theta}_0$	1.21×10^{-1}	1.81×10^{-2}	5.34×10^{-3}	1.72×10^{-2}
64×10^4	$\hat{\theta}_0$	2.20×10^{-1}	1.11×10^{-1}	4.77×10^{-2}	1.17×10^{-2}
	$\tilde{\theta}_0$	7.55×10^{-2}	3.93×10^{-2}	1.61×10^{-2}	3.41×10^{-3}
	$\tilde{\theta}_0^*$	7.79×10^{-2}	5.04×10^{-2}	4.80×10^{-2}	5.27×10^{-3}
	$\dot{\theta}_0$	1.20×10^{-2}	1.79×10^{-2}	2.16×10^{-3}	1.72×10^{-2}

Table 5: The mse for MLE $(\hat{k}_0, \hat{\theta}_0)$, CF $(\tilde{k}_0, \tilde{\theta}_0)$, smoothed CF $(\tilde{k}_0^*, \tilde{\theta}_0^*)$ and MM $(\dot{k}_0, \dot{\theta}_0)$ at different sample size n and null proportion p_0 when the zero assumption is violated.

6 Applications to RNA-seq analysis

We apply the proposed empirical null estimation procedures to the analysis of the B-cells data discussed in Section 1. The dataset contains transcriptome profiles of 17 male and 24 female immortalized B-cells samples. We make comparison between genders by PoissonSeq and obtain 9688 score statistics. The theoretical null distribution is Chi-squared with 1 degree of freedom ($Gamma(0.5, 2)$). We estimate the null parameters for the score statistics using each of the proposed methods. For the MLE and MM method we let $A_0 = (0, q)$ where q is the 3rd quartile of the test statistics and let bin width $d = 0.1$ for MM. For the CF method we set $\gamma = 0.05$ and for smoothing we let $J = 4$ and $h = 0.2$. The estimates are: $(\hat{k}_0, \hat{\theta}_0) = (0.488, 2.46)$, $(\tilde{k}_0, \tilde{\theta}_0) = (0.473, 5.27)$ and $(\tilde{k}_0^*, \tilde{\theta}_0^*) = (0.480, 2.71)$. The estimate $\tilde{\theta}_0$ is quite different from the other two, suggesting it may be affected by the non-smoothness of $\psi_n'(t)$.

Next we estimate the proportion of the null density p_0 using (6). This yields $\hat{p}_0 = 0.941$, $\tilde{p}_0 = 1.19$ and $\tilde{p}_0^* = 0.962$. Obviously the non-smoothness of the empirical characteristic function makes the CF estimator unstable. Therefore we only consider MLE and smoothed CF for FDR analysis. At given test statistic value x, the FDR can be estimated by

$$\widehat{FDR}(x) = p_0 \frac{1 - F_\Gamma(x|k_0, \theta_0)}{1 - \hat{F}(x)} \,. \tag{35}$$

To control the tail FDR, we find the minimum x_0 such that the $\widehat{FDR}(x_0) < 0.1$ and report the test statistics greater than x_0. The MLE method reports total 302 discoveries while the smoothed CF yields 232. In comparison, if we choose the theoretical null parameter $k_0 = 0.5, \theta_0 = 2$, the number of total discoveries is 474.

Similarly we apply the estimators to the other three datasets in Section 2. For the prostate cancer data $(\hat{k}_0, \hat{\theta}_0, \hat{p}_0) = (0.499, 6.00, 0.954)$, $(\tilde{k}_0, \tilde{\theta}_0, \tilde{p}_0) = (0.460, 6.81, 0.927)$ and $(\tilde{k}_0^*, \tilde{\theta}_0^*, \tilde{p}_0^*) = (0.514, 5.75, 0.967)$. Controlling for tail FDR at 0.1, the three methods yield 1141, 890 and 1226 discoveries out of 33575 genes respectively.

For the LNCap data $(\hat{k}_0, \hat{\theta}_0, \hat{p}_0) = (0.483, 6.10, 0.935)$, $(\tilde{k}_0, \tilde{\theta}_0, \tilde{p}_0) = (0.410, 5.48, 0.769)$ and $(\tilde{k}_0^*, \tilde{\theta}_0^*, \tilde{p}_0^*) = (0.478, 6.86, 0.979)$. Out of 17162 genes, the number of discoveries by the three methods are 618, 1001 and 495 respectively.

For the liver and kidney data $(\hat{k}_0, \hat{\theta}_0, \hat{p}_0) = (0.501, 17.9, 0.878)$, $(\tilde{k}_0, \tilde{\theta}_0, \tilde{p}_0) = (0.526, 18.5, 0.975)$ and $(\tilde{k}_0^*, \tilde{\theta}_0^*, \tilde{p}_0^*) = (0.571, 12.6, 0.913)$. Out of 18173 genes, the number of discoveries are 1720, 1557 and 2492 respectively.

7 Discussion

We have derived estimators based on maximum likelihood and characteristic functions to estimate the null density for Gamma distributions in large scale multiple testings.The methods are applied to RNA-seq experiment data. We show by examples that the theoretical null in RNA-seq analysis does not accurately match the test statistics and over-dispersion relative to a standard Chi-squared distribution is common in practice. We therefore propose to use the Gamma family of distributions to model the null distributions in RNA-seq experiments. Further we can use empirical methods to estimate the null parameters of the Gamma distribution in a two-class mixture setting. The empirical null provides better FDR control and it will greatly reduce the number of false discoveries.

Our MLE method follows the framework by Efron (2004) for normal data. It is easy to implement and fairly robust to different data structures. On the other hand it can be biased when the proportion of the non null effect is not small. Under smaller p_0 the CF estimator has shown superior performance, however it is heavily dependent on on choosing an appropriate t value for the characteristic function, which can be difficult for some data.

The CF estimator relies on the approximation of the characteristic function as well as its derivative. The empirical characteristic function usually provides good approximation to the true characteristic function, however its derivative can be noisy even under moderately large sample size. We propose a way to remove the noise based on local polynomial regression. Simulation and application both show that the smoothing approach is very effective when the sample size is limited.

We only demonstrate the gamma mixture model and our methods on RNA-seq experiments, but the application is not limited to that. Our methods can be used in estimate the null distribution for various data as long as the theoretical null belongs to the Gamma family. For example Chi-squared tests are common for testing single nucleotide polymorphisms (SNPs) in genome-wide association studies (Schwartzman, 2008). The Gamma mixture model and our estimation procedures should be adequate and flexible to accommodate the data structure in such tests.

Two aspects of our methods can benefit from future studies. First, the Gamma distribution proposed in this manuscript can be improved by adding a non-central parameter. For normal data, Efron (2004) argues that the empirical null may not be centered about 0. Similarly the Gamma or Chi-squared data not may not be central either. Adding a small non-central parameter can improve the accuracy of our model. Another aspect is the estimation of the null proportion. The estimator derived by the MLE method is easy to implement, but it does not guarantee an upper bound by 1. Jin (2008) has proposed a CF estimator of the null proportion for normal data. How to adopt the their strategy to Gamma data requires thorough considerations.

Acknowledgment

This work is supported by Roswell Park Cancer Institute and National Cancer Institute (NCI) grant #P30 CA016056.

References

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 (Methodological)*, pages 289–300.

Cheung, V. G., Nayak, R. R., Wang, I. X., Elwyn, S., Cousins, S. M., Morley, M., and Spielman, R. S. (2010). Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol*, 8(9):e1000480.

Delignette-Muller, M. L., Pouillot, R., Denis, J.-B., and Dutang, C. (2014). *fitdistrplus: help to fit of a parametric distribution to non-censored or censored data.* R package version 1.0-2.

Efron, B. (2004). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99(465).

Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):461–493.

Jin, J. and Cai, T. T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478):495–506.

Li, H., Lovci, M. T., Kwon, Y.-S., Rosenfeld, M. G., Fu, X.-D., and Yeo, G. W. (2008). Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proceedings of the National Academy of Sciences*, 105(51):20179–20184.

Li, J., Witten, D. M., Johnstone, I. M., and Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, 13(3):523–538.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517.

Miecznikowski, J. C. and Gaile, D. P. (2014). A novel characterization of the generalized family wise error rate using empirical null distributions. *Statistical applications in genetics and molecular biology*, 13(3):299–322.

Ren, S., Peng, Z., Mao, J.-H., Yu, Y., Yin, C., Gao, X., Cui, Z., Zhang, J., Yi, K., Xu, W., et al. (2012). Rna-seq analysis of prostate cancer in the chinese population identifies recurrent gene fusions, cancer-associated long noncoding rnas and aberrant alternative splicings. *Cell research*, 22(5):806–821.

Schwartzman, A. (2008). Empirical null and false discovery rate inference for exponential families. *The Annals of Applied Statistics*, 2(4):1332–1359.

Wand, M. (2013). *KernSmooth: Functions for kernel smoothing for Wand & Jones (1995).* R package version 2.23-10.

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*, volume 60. Crc Press.