Communications of the Korean Statistical Society
2009, Vol. 16, No. 2, 389–396

Use of Beta-Polynomial Approximations for Variance Homogeneity Test and a Mixture of Beta Variates

Hyung-Tae Ha[1,a], ChungAh Kim[b]

[a]Dept.of Applied Statistics, Kyungwon Univ., [b]Dept.of Tourism Management, Kyungwon Univ.

Abstract

Approximations for the null distribution of a test statistic arising in multivariate analysis to test homogeneity of variances and a mixture of two beta distributions by making use of a product of beta baseline density function and a polynomial adjustment, so called beta-polynomial density approximant, are discussed. Explicit representations of density and distribution approximants of interest in each case can easily be obtained. Beta-polynomial density approximants produce good approximation over the entire range of the test statistic and also accommodate even the bimodal distribution using an artificial example of a mixture of two beta distributions.

Keywords: Test statistic, p-values, density approximation, variance equality, moments, mixture of distributions, bimodality.

1. Introduction

Test statistic to test the hypothesis that variances of variates are equal is one of the principal test statistics used in multivariate analysis; p-values of the test are obtained from their null distributions. Many techniques for deriving the exact null and non-null distributions of multivariate test criteria were examined. Interestingly, determining their generating functions or moments or cumulants is often not as complicated as obtaining distribution functions. When the moment generating functions or cumulant generating functions of the distribution are given, the exact distributions can be obtained by making use of well known mathematical techniques such as method of direct integration, method of characteristic functions, method of convolutions, method of differential equations and the method of calculus of residues. It should be noted that it is often the case that the exact null distributions of tests statistics in multivariate analysis are analytically intractable or difficult to obtain in closed forms from those generating functions.

In these circumstances, it is desirable to obtain approximation techniques to provide such statistical quantities of the test statistics of interests. Several approximation techniques such as saddlepoint approximation and Edgeworth expansion, which are respectively based on cumulant generating functions and cumulants, have been extensively discussed in statistical inferences. For example, Butler and Wood (2002, 2004) approximated the noncentral distributions for three multivariate tests by making use of the sequential saddlepoint method, and Kolassa (2003) approximated the tail probability of sufficient statistics from a regression model with exponential errors using saddlepoint approximation. Edgeworth series approximation techniques have also been extensively applied in many fields of scientific areas. Since many test criteria under the null hypotheses can be expressed in terms of a convolution, Edgeworth series approximation might be a good option for approximating a density

[1] Corresponding author: Assistant Professor, Department of Applied Statistics, Kyungwon University, Sungnam-ci, Kyunggi-do 461-701, Korea. E-mail: htha@kyungwon.ac.kr

function when the normal approximation does not provide enough accuracy. The saddlepoint approximation methods are usually quite accurate in the tail areas of the target density. However, it should be pointed out that it is difficult to apply Edgeworth series approximation in the cases that the statistics of interest do not follow normal distributions. As pointed out in Reid (1988), saddlepoint approximation techniques are not widely used in many scientific applications because it may not be easy to understand the concepts of the techniques nor apply them in many types of situations although they are very accurate approximation tool in tail probability.

The density approximants applied in this paper, proposed in Ha and Provost (2007), which are not only conceptually simple but also easy to implement, are expressed as the product of initial baseline density on the basis of beta density function and a polynomial adjustment. Two parameters of the baseline density and the polynomial coefficients are both determined by making use of a matching-moment technique. Finite mixture distributions arise in a variety applications and many types of finite mixture distributions have been discussed in last several decades. For instance, the length distribution of a certain type of fish, as studied in Bhattacharya (1967), was found to split the observations into age categories, with each category contributing a normal component distributions to yield an overall mixture. Mixtures of beta distributions were discussed in statistical literatures, for instance, Tretter and Walster (1975) and Barlow *et al.* (1972). An artificial example of a mixture of two beta distributions has also been used in this paper to illustrate the approximation accuracy and flexibility to accommodate bimodality of the target distribution via beta-polynomial density approximants.

Beta-polynomial approximation technique is briefly reviewed in Section 2. The distribution of a test for testing the hypothesis of homogeneity of variances, are approximated in Section 3. It is shown in Section 4 that the technique can also be applied to accommodate bimodality of a mixture of two beta distributions. Finally, concluding remarks and certain computational aspects are discussed in Section 5.

2. Beta-Polynomial Approximation

A general semi-parametric approach to density approximation is proposed in Ha and Provost (2007). In this section, we review a technique on the basis of beta baseline density function for approximating density and distribution functions.

Let X be a random variable whose support is the interval $[0, 1]$ and let its raw moments $E(X^h)$ be denoted by $\mu_X(h)$, $h = 0, 1, \ldots$. We are interested in approximating the density and distribution functions of the random variable X, denoted by $f_X(x)$ and $F_X(x)$, respectively. A *beta-polynomial* density approximant of degree d, denoted by $f_{X_d}(x)$, is

$$f_{X_d}(x) = \psi(x) \sum_{i=0}^{d} \xi_i x^i . \tag{2.1}$$

This density approximant is expressed as the product of an initial approximation, $\psi(x)$ and a polynomial adjustment, $\sum_{i=0}^{d} \xi_i x^i$. That is, the beta baseline density function is

$$\psi(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \, I_{[0,1]}(x), \tag{2.2}$$

where $I_A(x)$ denotes the indicator function, which is equal to 1 when $x \in A$ and 0 otherwise and $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. The parameters α and β of the beta baseline density function are estimated from

the first two moments of X as follows:

$$\alpha = \frac{\mu_X(1)(\mu_X(1) - \mu_X(2))}{\mu_X(2) - \mu_X(1)^2} \quad \text{and} \quad \beta = \frac{(1 - \mu_X(1))(\alpha + 1)}{\mu_X(1)}, \tag{2.3}$$

see for instance Johnson *et al.* (1995, Section 25). The j^{th} moments of the beta baseline density function is denoted by $m(j)$, that is,

$$\int_0^1 x^j \psi(x) dx \equiv m(j). \tag{2.4}$$

The j^{th} moment of this beta baseline distribution can be expressed as

$$m(j) = \frac{\Gamma(\alpha+\beta)\,\Gamma(\alpha+j)}{\Gamma(\alpha)\,\Gamma(\alpha+j+\beta)} = \frac{\prod_{k=0}^{j-1}(k+\alpha)}{\prod_{k=0}^{j-1}(k+\alpha+\beta)}, \quad j = 1, 2, \ldots. \tag{2.5}$$

From the moment matching technique between the moments of the target distribution and the estimated beta baseline distribution, we can obtain the coefficients ξ_i of the polynomial adjustment. That is, the coefficients ξ_i satisfy the following matrix form equation,

$$\begin{pmatrix} \xi_0 \\ \xi_1 \\ \vdots \\ \xi_d \end{pmatrix} = \begin{pmatrix} m(0) & m(1) & \cdots & m(d-1) & m(d) \\ m(1) & m(2) & \cdots & m(d) & m(d+1) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m(d) & m(d+1) & \cdots & m(2d-1) & m(2d) \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ \mu_X(1) \\ \vdots \\ \mu_X(d) \end{pmatrix}. \tag{2.6}$$

The approximate cumulative distribution function of X, denoted by $F_{X_d}(x)$, evaluated at a is then

$$\begin{aligned} F_{X_d}(a) &= \sum_{i=0}^{d} \xi_i \int_0^a x^i \psi(x) \mathrm{d}x \\ &= \sum_{i=0}^{d} \frac{\xi_i\,\Gamma(\alpha+\beta)\,\mathrm{B}(a, i+\alpha, \beta)}{\Gamma(\alpha)\,\Gamma(\beta)}, \quad 0 \le a \le 1, \end{aligned} \tag{2.7}$$

where

$$\mathrm{B}(a, b, \beta) = \int_0^a t^{b-1}(1-t)^{\beta-1} dt \tag{2.8}$$

denotes the incomplete beta function.

3. Test for Homogeneity of Variances

Suppose that there are random samples of size n, x_{ij}, $i = 1, \ldots, p$, $j = 1, \ldots, n$, taken from p normal populations with unknown means μ_i and variances σ_i, $i = 1, \ldots, p$. One may be interested in testing the null hypothesis H_0: $\sigma_1 = \cdots = \sigma_p$ against the alternative H_a : $\sigma_i \neq \sigma_j$ for some $i = j$. The test

statistic proposed by Neyman and Pearson (1931), denoted by H, with significance level α has critical region $0 < c < H < 1$, where

$$H = \prod_{i=1}^{p} \left(\frac{\sum_{j=1}^{n} \left(x_{ij} - \frac{1}{n} \sum_{j=1}^{n} x_{ij} \right)^2}{\frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{n} \left(x_{ij} - \frac{1}{n} \sum_{j=1}^{n} x_{ij} \right)^2} \right)^{\frac{n}{2}} \tag{3.1}$$

and $\Pr(c < H < 1) = \alpha$. Interestingly, its integrands such as its generating functions or moments or cumulants are not as complicated as obtaining distribution functions. As shown in Provost and Rudiuk (1996), the k^{th} moment of this test statistic when the hypothesis is true is

$$E\left(H^k\right) = \frac{p^{kp\frac{n}{2}} \Gamma\left(\frac{k(n-1)}{2}\right) \Gamma^k\left(\frac{kn}{2} + \frac{n-1}{2}\right)}{\Gamma^k\left(\frac{n-1}{2}\right) \Gamma\left(\frac{kpn}{2} + \frac{p(n-1)}{2}\right)}. \tag{3.2}$$

It is known that, given the explicit representation of the moments, inverse Mellin transform technique is widely used for obtaining the exact density function. Gupta and Rathie (1982) obtained the exact density of H in terms of Meijer's G-function in the finite support of (0, 1) as

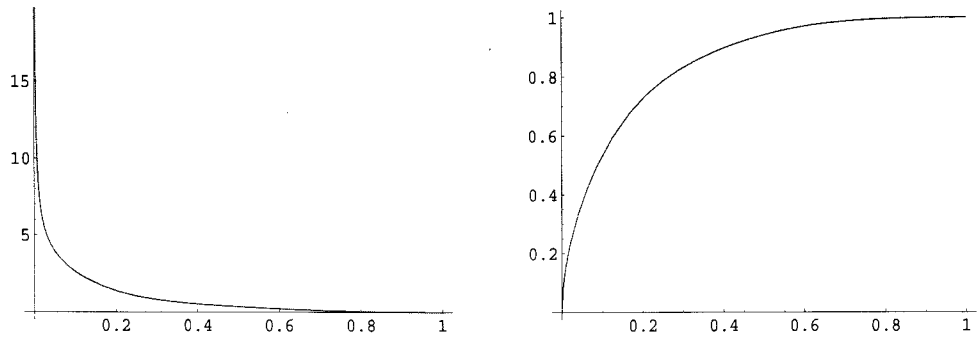
$$f(x) = \frac{2p^{\frac{p-np+1}{2}} \Gamma\left(\frac{p(n-1)}{2}\right)}{n(2\pi)^{\frac{1-p}{2}} \Gamma^p\left(\frac{n-1}{2}\right)} x^{-\frac{1}{n}} G_{p-1,p-1}^{p-1,0}\left\{ x^{\frac{2}{n}} \middle| \begin{matrix} \frac{1}{p}, \ldots, \frac{p-1}{p} \\ 0, \ldots, 0 \end{matrix} \right\}, \tag{3.3}$$

where the G-function can, in general, be defined with the following integral on the complex plane:

$$G_{m,n}^{p,q}\left\{ z \middle| \begin{matrix} a_1, \ldots, a_p \\ b_1, \ldots, b_q \end{matrix} \right\} = \frac{1}{2\pi i} \int_L \frac{\prod_{j=1}^{m} \Gamma(b_j - s) \prod_{j=1}^{n} \Gamma(1 - a_j + s)}{\prod_{j=m+1}^{q} \Gamma(1 - b_j + s) \prod_{j=n+1}^{p} \Gamma(a_j - s)} z^s ds \tag{3.4}$$

on denoting L as the path to follow while integrating. This explicit representation for the density function of H in Equation (3.3) is somewhat complicated in terms of analytical and computational aspects. One needs to compute another integral involving the Meijer's G-function, which contains integral over the path of L, in order to obtain statistical quantities such as p-values, confidence intervals and percentiles. It is the case that the exact null distribution of this tests statistic is difficult to obtain in closed forms.

In such circumstances, it is desirable to use approximation techniques to obtain the simple forms for the density and distribution functions, from which one could easily provide statistical quantities. We use the beta-polynomial approximation technique to density and distribution functions proposed in Ha and Provost (2007), which is based on the moment-matching technique. Let us consider a numerical example. When $p = 6$ and $N = 12$, one can easily obtain the integer moments of the test



(a) $f_{H_5}(h)$ (b) $F_{H_5}(h)$

Figure 1: $f_{H_5}(h)$ and $F_{H_5}(h)$

statistic from Equation (3.2). The following density approximant was obtained from Equation (2.1) on the basis of five moments:

$$f_{H_5}(h) = h^{-0.48988}(1-h)^{1.86665}(0.84033 + 5.17208h - 43.44577h^2 + 132.04935h^3 - 168.81722h^4 + 77.38472h^5). \tag{3.5}$$

The corresponding distribution function can easily be obtained from Equation (2.7) or integration of the density approximant unlike the exact probability density function expressed in terms of the Meijer G-function. The approximate 95^{th} and 99^{th} percentiles of this distribution of H for $p = 6$ and $N = 12$ are 0.52542 and 0.72685, respectively. Figure 1 shows $f_{H_5}(h)$ and $F_{H_5}(h)$, the approximants of the density and distribution functions with polynomial adjustment of degree of five. It should be mentioned that the approximants for the density and distribution functions are very simple and efficient with respect to computation unlike the exact density and distribution functions.

4. Mixture of Two Beta Distributions

The suitable degree can be determined by inspecting the plots of beta-polynomial density approximants of increasing degrees. In this section, the determination of a suitable degree is shown by making use of an example of a mixture of beta distributions. The more complicated situations arise, the more complex modelings such as finite mixtures need to be utilized. Finite mixture of distributions is often suitable to model population to consist of clusters with different parameters. A mixture model takes into account, for instance, optimal mixture of the ingredients for mixed fruit juices, weight or height of two different groups such as female and male and financial returns to often behave differently. Specially, the mixtures of beta distributions arise in a number of contexts such as the central cumulative density functions of Wilks' $\Lambda = |E|/|E + H|$ shown in Tretter and Walster (1975), which can be expressed as an infinite mixture of incomplete beta functions and significance levels for the ordered $\bar{E}^2$ tests of Barlow *et al.* (1972).

In this section, we use a mixture of two beta distributions to show that the bimodality can accurately be obtained by beta-polynomial approximation. The probability density function for a mixture of ℓ beta density functions is

$$p(x) = \sum_{i=1}^{\ell} \pi_i \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} x^{\alpha_i - 1}(1-x)^{\beta_i - 1} \mathcal{I}_{[0,1]}(x) \tag{4.1}$$

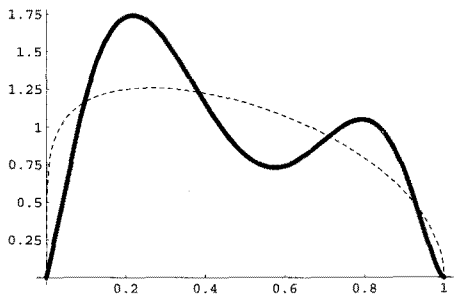
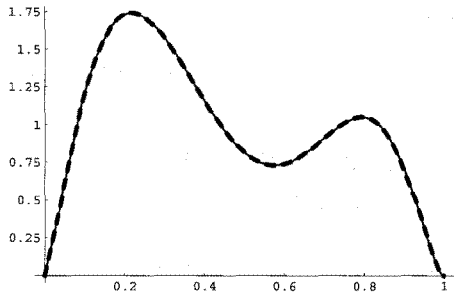
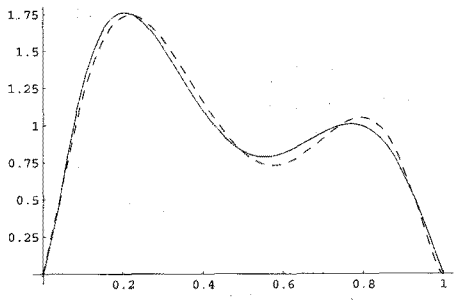


Figure 2: *Exact density(solid) and beta baseline density(dashed line)*



(a) *Exact density(solid) and* $f_{M_8}(x)$*(dashed line)*



(b) $f_{M_4}(x)$*(solid) and* $f_{M_8}(x)$*(dashed line)*

Figure 3: *Exact density,* $f_{M_4}(x)$ *and* $f_{M_8}(x)$

denoting π_i the weights for the corresponding distributions, where $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{\ell} \pi_i = 1$.

We consider a mixture of two beta random variables with parameters (7/3, 29/5) and (55/7, 19/7) and weights $\pi_1 = 2/3$, denoted by M, whose raw moments, $\mu_M(j)$, $j = 0, 1, \ldots$, can be expressed as follows:

$$\mu_M(j) = \int_0^1 x^j p(x)\mathrm{dx} \tag{4.2}$$
$$= \frac{2\prod_{k=0}^{j-1}\left(k+\frac{7}{3}\right)}{3\prod_{k=0}^{j-1}\left(k+\frac{122}{15}\right)} + \frac{\prod_{k=0}^{j-1}\left(k+\frac{55}{7}\right)}{3\prod_{k=0}^{j-1}\left(k+\frac{74}{7}\right)}.$$

From matching the first two moments of the mixture of two beta distributions to the first two moments of beta baseline density with the parameters α and β as shown in Equation (2.3), we obtain the parameters of the beta baseline density function, that is, $\alpha = 1.1817$ and $\beta = 1.5101$. The exact density for the mixture of two beta distributions and the estimated beta baseline density are shown in Figure 2.

As mentioned in Ha and Provost (2007), a suitable degree for a density approximation can be determined by a *de visu* inspection of the density plots of approximants of successive degrees. A density approximant of degree d is adequate if no noticeable differences are observed when comparing the plots of approximants of degrees $d - j$ and $d + j$ where j is a positive integer, a very small value of j being indicated if convergence occurs with relatively few moments. For the case of the mixture

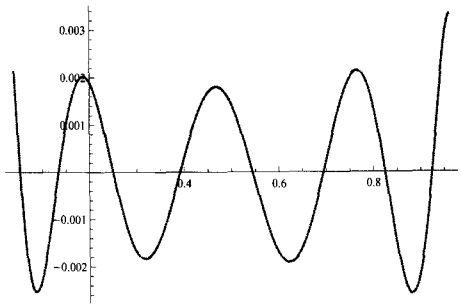


Figure 4: *PDF difference between exact density and* $f_{M_8}(x)$

of two beta distributions, it seems appropriate to set $j = 1$. From applying *de visu*, beta-polynomial density approximant of degree of eight appears to provide a satisfactory approximation. The resulting density approximant is superimposed on the exact density function in Figure 3(a). As can be seen in Figure 3(a) and 3(b), the exact density function is more irregular than those test statistics previously considered and thus the higher degree of the polynomial adjustment is required in order to obtain a suitable approximation. The difference between the exact and approximated probability density functions is also plotted in Figure 4, from which it is clearly seen that the approximation is very accurate.

5. Concluding Remarks

The symbolic computational package *Mathematica* was utilized for obtaining the density approximants and graphs for those examples. The density function of the test statistic of variance homogeneity was approximated by making use of beta-polynomial density approximation. And the corresponding distribution can also be easily computed by integration. One artificial numerical example is proposed and approximated, which is required a relatively large number of moments since the distributions considered exhibited irregular features. When the density functions to be approximated are the more irregular, one will require the more moments. The convergence behaviors of the beta-polynomial approximants need to be investigated in future studies.

Acknowledgements

The authors are grateful to two anonymous referees for very constructive comments to have led to a much improved version of the paper.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions*, John Wiley & Sons, London.

Bhattacharya, C. G. (1967). A simple method of resolution of a distribution into Gaussian components, *Biometrics*, **23**, 115–135.

Butler, R. W. and Wood, A. T. A. (2002). Laplace approximations for hypergeometric functions with matrix argument, *The Annals of Statistics*, **30**, 1155–1177.

Butler, R. W. and Wood, A. T. A. (2004). Mixture representations of noncentral distributions in multivariate analysis with application to a dimensional CLT, *Scandinavian Journal of Statistics*, **31**, 631–650.

Gupta, A. K. and Rathie, A. K. (1982). Distribution of the likelihood ratio criterion for the problem of K samples, *Metron*, **40**, 147–156.

Ha, H. T. and Provost, S. B. (2007). A viable alternative to resorting to statistical tables, *Communication in Statistics: Simulation and Computation*, **36**, 1135–1151.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions,* John Wiley & Sons, New York.

Kolassa, J. E. (2003). Multivariate saddlepoint tail probability approximations, *The Annals of Statistics*, **31**, 274–286.

Neyman, J. and Pearson, E. S. (1931). On the problem of k-samples, *Bulletin of Academic Polonaise Science Letters*, **3**, 460–481.

Provost, S. B. and Rudiuk, E. M. (1995). Moments and densities of test statistics for covariance structures, *International Journal of Mathematical and Statistical Sciences*, **4**, 85–104.

Reid, N. (1988). Saddlepoint methods and statistical inference, *Statistical Science*, **3**, 213–227.

Tretter, M. J. and Walster, G. W. (1975). Central and noncentral distributions of Wilks' statistic in Manova as mixtures of incomplete Beta functions, *The Annals of Statistics*, **3**, 467–472.

Received August 2008; Accepted October 2008