Scientific Research Publishing

Open Journal of Statistics, 2018, 8, 731-746
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Asymptotic Results for Goodness-of-Fit Tests Using a Class of Generalized Spacing Methods with Estimated Parameters

Andrew Luong

École d'actuariat, Université Laval, Ste Foy, Québec, Canada
Email: Andrew.Luong@act.ulaval.ca

How to cite this paper: Luong, A. (2018) Asymptotic Results for Goodness-of-Fit Tests Using a Class of Generalized Spacing Methods with Estimated Parameters. *Open Journal of Statistics*, **8**, 731-746.
https://doi.org/10.4236/ojs.2018.84048

Received: June 8, 2018
Accepted: August 24, 2018
Published: August 27, 2018

Abstract

A class of pseudo distances is used to derive test statistics using transformed data or spacings for testing goodness-of-fit for parametric models. These statistics can be considered as density based statistics and expressible as simple functions of spacings. It is known that when the null hypothesis is simple, the statistics follow asymptotic normal distributions without unknown parameters. In this paper we emphasize results for the null composite hypothesis: the parameters can be estimated by a generalized spacing method (GSP) first which is equivalent to minimize a pseudo distance from the class which is considered; subsequently the estimated parameters are used to replace the parameters in the pseudo distance used for estimation; goodness-of-fit statistics for the composite hypothesis can be constructed and shown to have again an asymptotic normal distribution without unknown parameters. Since these statistics are related to a discrepancy measure, these tests can be shown to be consistent in general. Furthermore, due to the simplicity of these statistics and they come a no extra cost after fitting the model, they can be considered as alternative statistics to chi-square statistics which require a choice of intervals and statistics based on empirical distribution (EDF) using the original data with a complicated null distribution which might depend on the parametric family being considered and also might depend on the vector of true parameters but EDF tests might be more powerful against some specific models which are specified by the alternative hypothesis.

Keywords

Density Based Tests, EDF Tests, Anderson-Darling Statistic, Hellinger Distance Statistic, Pseudo-Distance, Maximum Spacing Method

1. Introduction

Let $X_1,\cdots,X_{n-1}$ be a sample of size $n-1$ from a continuous distribution $F\in\{F_{\boldsymbol{\theta}}\}$ and let $X_1\le\cdots\le X_{n-1}$ be the order statistics and let the transformed data be defined as $U_i(\boldsymbol{\theta})=F_{\boldsymbol{\theta}}(X_i), i=1,\cdots,n-1$, $U_{(i)}(\boldsymbol{\theta})=F_{\boldsymbol{\theta}}\left(X_{(i)}\right), i=1,\cdots,n-1$ and define $F_{\boldsymbol{\theta}}\left(X_{(n)}\right)=1$ and $F_{\boldsymbol{\theta}}\left(X_{(0)}\right)=0$. The spacings are given by $D_i(\boldsymbol{\theta})=F_{\boldsymbol{\theta}}\left(X_{(i)}\right)-F_{\boldsymbol{\theta}}\left(X_{(i-1)}\right)$, $i=1,\cdots,n$

Ghosh and Jammalammadaka [1], Luong [2] have studied generalized spacing methods (GSP) of estimation with the vector of GSP estimators given by the vector $\hat{\boldsymbol{\theta}}$ which minimizes

$$\sum_{i=1}^{n} h(nD_i) \quad \text{with} \quad h(x)=-x^{\alpha} \tag{1}$$

Using this class of $h(x)$, it is shown that the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by

$$\frac{1}{n}\sigma_h^2(\alpha)\left[I(\boldsymbol{\theta})\right]^{-1} \quad \text{with} \quad \sigma_h^2(\alpha)\ge 0$$

and $\sigma_h^2(\alpha)$ depends on α but does not depend on the parametric family $\{F_{\boldsymbol{\theta}}\}$ and $\boldsymbol{I}(\boldsymbol{\theta})$ is the usual information matrix of maximum likelihood (ML) estimation.

Furthermore, by letting $\alpha\to 0_+$, $\sigma_h^2(\alpha)\to 1$. This result is interesting, as it means if we set $\alpha=0.01$ we then have $\sigma_h^2(\alpha)\approx 1.02$ and therefore, the loss of efficiency comparing to ML estimation or maximum spacing (MSP) method is around two percent no matter which parametric model is used. Luong [2] has also shown that this loss of efficiency is compensated by a gain in robustness and it might be preferred to use GSP estimation if ML and MSP estimation are not robust; see Remark 2 as given by Luong [2] (p 632). Furthermore, when there are tied observations, this implies some spacings will be equal to 0 and log of these spacings is undefined so that we might want to use GSP methods instead of maximum spacing method (MSP) method; see Section 5 for tied-observations. MSP method is also called maximum product of spacings method; see Cheng and Stephens [3].

In this paper, we focus on using this class of GSP methods for construction of goodness-of-fit tests statistics for testing the simple null hypothesis:

H_0: data come from a distribution $F_0=F_{\boldsymbol{\theta}_0}$; $\boldsymbol{\theta}_0$ is specified and for testing the composite null hypothesis.

H_0: data come from the parametric family $F\in\{F_{\boldsymbol{\theta}}\}$; $\boldsymbol{\theta}$ is unspecified. For the composite H_0, Cheng and Stephens [3] have shown that the Moran's statistic with parameters estimated by the MSP method has an asymptotic normal distribution which does not depend on the parametric family $\{F_{\boldsymbol{\theta}}\}$ and we shall show that similar properties hold for the class of test statistics constructed using the class of GSP methods being considered in this paper. In a previous paper, we have considered estimation using this class of GSP methods and parameter hypothesis testing. In this paper, we focus on model validation using this class of GSP methods since testing for goodness-of-fit for composite H_0 using GDP methods has not received much attention in the literature.

We adopt an approach using pseudo distances by showing the class of $h(x) = -x^{\alpha}, 0 < \alpha < 1$ induces a class of pseudo distances which we shall denote by $d_{\phi}(f_1, f_2)$, the function $\phi = 1 - x^{\alpha} = 1 + h(x)$, f_1 and f_2 are densities and $d_{\phi}(f_1, f_2)$ is a measure to quantify how close these densities are. Implicitly, for methods using spacings we work with transform data and if $\boldsymbol{\theta}_0$ is the true parameter then the transform data

$$U_i = F_0(X_i) = F_{\theta_0}(X_i), i = 1, \cdots, n-1$$

will follow a standard uniform distribution with density $f_U = 1, 0 \le x \le 1$ and $f_U = 0$, elsewhere.

Using the transformed data we can obtain an easily constructed elementary density estimate without requiring a kernel of the usual density estimate, this empirical density estimate is denoted by $\hat{f}_n$, see expression (6) and for testing the simple null hypothesis, a test statistic can be constructed which is based on

$$n^k d_{\phi}(f_n, f_U) \tag{2}$$

with the restriction on $k > 0$ and $n^k d_{\phi}(f_n, f_U)$ can be reexpressed equivalently as a simple function of spacings and numerically simple to compute; the statistic will follow an asymptotic normal distribution which does not depend on the parametric family. For the statistic to have good power for large samples, it appears that we should choose the scaling factor n^k so that an asymptotic distribution exists for the statistic given by expression (2) and at the same time $k > 0$ so that $n^k \to \infty$ as $n \to \infty$ and if $d_{\phi}(f_n, f_U)$ can be used to discriminate whether the sample is drawn from an assumed distribution, the test will be consistent and it is an advantage over chi-square tests which do not have the consistency property, in general.

For the composite hypothesis, we use a GSP method to obtain the GSP estimators given by the vector $\hat{\boldsymbol{\theta}}$ first but we shall see that minimizing expression (1) is equivalent to minimizing the following pseudo distance based on a function ϕ, the expression up to a positive multiplicative constant is given by $n^k d_{\phi}(f_n^{\boldsymbol{\theta}}, f_U)$, $f_n^{\boldsymbol{\theta}}$ is defined by expression (11) in Section (4).

Subsequently the statistic is based on

$$n^k d_{\phi}\left(f_n^{\hat{\boldsymbol{\theta}}}, f_U\right) \tag{3}$$

and after simplifications, it is reduced to a simple function of spacings with estimated parameters and it will be shown again the equivalent statistic to the one given by expression (3) will follow an asymptotic normal distribution without unknown parameters; this property will facilitate goodness-of-fit testing. Using this unified presentation, we would like to show that these statistics are density based and they are parallel to traditional test statistics based on distribution functions (EDF) such as the Anderson-Darling statistic, see Anderson Darling [4], Boos [5], Stephens [6] or chi-square goodness-of-fit statistics with parameters estimated with minimum chi-square methods as discussed by Greenwood and Nikulin [7] (p 70-159).

The approach used in this paper hopefully will unify estimation and model testing and facilitate the comparisons of these density based statistics with traditional EDF statistics and chi-square statistics which are more often used than these density based statistics. We note that these statistics can be computed easily and their null asymptotic distribution is normal without unknown parameters which make it easy to use these statistics and comparing to the related chi-square statistics, these statistics do not need a choice of intervals and they come as by products when fitting models using the corresponding GSP methods. This feature is not shared by maximum likelihood (ML) methods.

We also note that power analysis using theoretical works might not give a complete picture for these density based statistics as the analysis is often based on only one sequence of functions which belongs to the alternative hypothesis converging to the functions specified by the null distribution and there are so many sequences that can approach the functions of the hypothesis in a functional space; see Sethuraman and Rao [8] for Pitman efficiency analysis for these statistics for the simple null hypothesis.

In this paper, we shall concentrate on asymptotic distributions goodness-of-fit tests statistics based on GSP methods and emphasizing a class of GSP methods which complete the results on estimation and parameter testing given by a previous paper. Implicitly, GSP methods in this paper mean GSP methods restricted to the class being considered in this paper. Furthermore, we do not touch upon the question of power analysis which might need extensive simulations studies with many models chosen for the alternative hypothesis as we do not have enough computing facilities and resources for such large scale simulation studies, see Cheng and Stephens [3] (p 386) on power of the Moran's statistic with the MSP method which is also called maximum product of spacings method; also see Zhang [9] for simulation studies for assessing the power of some EDF tests.

The paper is organized as follows.

In Section 2, a class of pseudo distances which generate the related GSP methods for estimation and model testing is introduced and the inference methods are based on spacings or equivalently on transformed data. The elementary density estimate introduced by Kale [10] is presented in Section 3 and a pseudo distance between the elementary density estimate and the standard uniform density is used to construct goodness-of-fit statistic for testing the simple null hypothesis, the statistic is shown to be expressible as a simple function of spacings which follows an asymptotic normal distribution without unknown parameters under the simple null hypothesis. In Section 4, for testing the composite hypothesis we can choose a GSP method within the class being considered by minimizing a corresponding pseudo distance which implicitly define a GSP method for estimation then use the obtained estimators to replace the unknown parameters in the pseudo distance to construct a goodness-of-fit statistic and after simplification the statistic is expressible as a simple function of spacings with an asymptotic normal distribution which does not depend on the parameters as in the simple null hypothesis case. In Section 5, tied-observations are discussed and it might be more practical to use a GSP method instead of the MSP method as tied observations do not cause numerical difficulties for GSP methods but extra cares are needed if MSP method is used, see Cheng and Stephens [3] (p 391) for tied observation treatments for MSP method. Section 6 gives some discussions on power analysis using theoretical works highlighting that theoretical power analysis might not give a complete picture of power of the statistics due to functions are involved under the null and alternative hypothesis comparing to the classical set up which only involve scalars.

2. Discrepancy Measures or Pseudo Distances

We shall see that pseudo-distances can be created using a convex function $\phi(x), x \geq 0$ with $\phi'(x)$ and $\phi''(x)$ being respectively its first and second derivatives with $\phi''(x) \geq 0$. We focus on pseudo distances defined by using as $\phi(x) = 1 - x^{\delta}, x \geq 0, 0 < \delta < 1$ and let $\alpha = 1 - \delta$. The GSP estimators given by the vector $\hat{\boldsymbol{\theta}}$ can be seen are based on this class as they are obtained by minimizing the following objective function with respect to $\boldsymbol{\theta}$ and by choosing a value for α,

$$T_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \left(nD_i(\boldsymbol{\theta})\right)^{\alpha}, 0 < \alpha < 1,$$

i.e., specifying $h(x) = -x^{\alpha}, 0 < \alpha < 1, x > 0$.

We shall see that using this class of $h(x)$ using spacings is equivalent to use a class of pseudo distances for densities defined using $\phi(x)$. It has been shown in our previous paper that GSP methods can attain high efficiency for estimation using values for α being positive and near 0.

Note that by letting $\alpha \to 0_+$ we obtain full efficiency and with $\alpha = 0.05$, the asymptotic relative efficiency is around 0.98 for all parametric families comparing to fully efficient methods such as the MSP method or ML method or Hellinger method based on density estimate using the original data introduced by Beran [11]. The elementary density estimate which makes use of spacings is based on transformed data and it is easily obtainable without requiring a kernel. The elementary density estimate is due to Kale [10]. We shall introduce it subsequently after the definition of pseudo distance and give an interpretation to GSP methods as minimum distance methods based on pseudo distances which are density based measures of discrepancy. Presenting from this point of view, it parallels the Hellinger methods introduced by Beran [11] with the use of Hellinger distance and the original data. It might be more complicated for practitioners to implement Beran's minimum Hellinger distance methods which require a kernel density estimate with a choice of window than implementing these GSP methods.

This will also make the GSP methods parallel to EDF methods such as the Cramér Von Mises methods or weighted Cramér-Von Mises distances such as the Anderson-Darling distance methods which also make use of the original data. For Anderson-Darling (AD) distance, see Anderson and Darling [12]. The Anderson-Darling distance is also a pseudo distance which is always nonnegative and measures the discrepancy between two distribution functions and it needs not obey the triangle inequality. Minimizing the discrepancy between the usual empirical distribution and the distribution of the parametric family will give the minimum Anderson-Darling estimators (MAD), see Boos [5].

In general, the MAD estimators are robust and have high efficiencies but for some parametric families, the overall relative efficiency when compared to maximum likelihood (ML) estimators can fall below 0.80, see Boos [5] (p 2754). Once the MAD estimators given by the vector $\bar{\boldsymbol{\theta}}$ is obtained, the Anderson-Darling distance can be used to form the AD statistic which is given by

$$AD\left(\bar{\boldsymbol{\theta}}\right)=n\int_{-\infty}^{\infty}\left(F_n-F_{\bar{\boldsymbol{\theta}}}\right)^2\mathrm{d}F_{\bar{\boldsymbol{\theta}}}\left(x\right) \tag{4}$$

to test the validity of the model specified by the composite $H_0: F\in\left\{F_{\boldsymbol{\theta}}\right\}$, *i.e.*, the data is drawn from a distribution F which belongs to the family $\left\{F_{\boldsymbol{\theta}}\right\}$ and F_n is the usual empirical distribution function using the original data. The expression (4) can also be reexpressed so that it is more suitable for calculations see Boos [5] (p 2748). It has been shown that the null distribution of statistics which is based on empirical distribution function (EDF) such as the AD statistic defined by expression (4) does not have a unique null distribution asymptotically as it will depend on $\left\{F_{\boldsymbol{\theta}}\right\}$ and possibly also on $\boldsymbol{\theta}_0\in\theta$, see Boos [5] (p 2759-2766), Pollard [12] (p 61). Even in the case where the null hypothesis is simple, it is still quite complicated and often extensive simulations are needed to calculate the p-value of such tests or extensive tables are needed for the use of these EDF tests. We shall see that it is not the case for the GSP methods based on the ϕ functions as we have defined earlier. We focus on this class of ϕ functions as it can give high efficiency for estimation and the pseudo distances used for estimation can also be used to construct goodness-of-fit statistics. Unlike the EDF test statistics, for statistics using these pseudo distances we have an asymptotic normal distribution as null distribution regardless of the value of the vector $\boldsymbol{\theta}_0$ and regardless of $\left\{F_{\boldsymbol{\theta}}\right\}$ for goodness-of-fit the parametric model. The goodness-of-fit statistics are easily obtainable as they are based on the same pseudo distances used for estimation and the statistics can be expressed in an equivalent form as simple functions of the spacings. In this paper, we relate estimation and goodness of fit by considering them as inference methods based on pseudo distances; the approach might provide more insights on the methods using spacings which have appeared in the literature as methods for estimation and testing using spacings are usually presented separately.

Before introducing these goodness-of-fit statistics, first we shall define a ϕ-discrepancy measure which induces a ϕ-pseudo-distance. The definitions have been given by Ali and Silvey [13], Pardo [14] (p 5-7) and reproduced below.

Definition (ϕ-pseudo-distance)

The ϕ-pseudo-distance or ϕ-divergence measure between two densities f_1 and f_2 is defined by $d_\phi(f_1,f_2)=E_{f_2}\left(\phi\left(\frac{f_1}{f_2}\right)\right)$, $E_{f_2}(.)$ is the expectation using f_2, ϕ is a convex function with $\phi(x)$ defined for $x\geq 0$ and the second derivative $\phi''(x)$ exists and $\phi''(x)\geq 0,\phi(1)=0$.

We have $d_\phi(f_1,f_2)\geq 0$, $d_\phi(f_1,f_2)=0$ if and only if $f_1=f_2$ except on a set of measure 0. The discrepancy measure needs not be symmetric as $d_\phi(f_1,f_2)\neq d_\phi(f_2,f_1)$ and it does not need to obey the triangle inequality and unless otherwise stated, we focus on the class of $\phi(x)=1-x^\delta, x\geq 0, 0<\delta<1$ and let $\alpha=1-\delta$.

Using the above function, the pseudo distance can be expressed as

$$d_\phi(f_1,f_2)=1-\int_{-\infty}^{\infty}\left(\frac{f_1(x)}{f_2(x)}\right)^\delta f_2(x)\mathrm{d}x=1-\int_{-\infty}^{\infty}f_1^\delta(x)f_2^{1-\delta}(x)\mathrm{d}x$$

and we shall use these pseudo distances to construct goodness-of-fit test statistics using transformed data or equivalently spacings and related them with results which already obtained using spacings which have appeared in the literature. The advantage of this approach is an unified treatment can be given to estimation and model testing and it can reveal tests based on statistics which make use of spacings which might not be powerful for large samples when used for testing of goodness-of-fit.

Note that Hellinger distance (HD) which is a true distance as used by Beran [11] can be expressed in a similar form with

$$d_{\mathrm{HD}}(f_1,f_2)=\int_{-\infty}^{\infty}\left(f_1(x)-f_2(x)\right)^2\mathrm{d}x=2-2\int_{-\infty}^{\infty}f_1^{\frac{1}{2}}(x)f_2^{\frac{1}{2}}(x)\mathrm{d}x.$$

In the next section, we shall present an elementary density estimate using transformed data and we aim to test the following simple H_0 which specifies that the random sample of observation is drawn from a distribution function $F_0(x)=F_{\boldsymbol{\theta}_0}(x)$, $\boldsymbol{\theta}_0$ is specified and $F_0(x)$ has a closed form expression.

We assume to have a random sample of size $n-1$ which consists of $X_1,\cdots,X_{n-1}$ and these observations are independent and identically distributed(iid) as X which follows a distribution $F\in\{F_{\boldsymbol{\theta}}\},\{F_{\boldsymbol{\theta}}\}$ is the parametric model used and let the order statistics be denoted by $X_1\leq X_2<\cdots\leq X_{n-1}$.The vector of parameters is denoted by $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_m)'$, $\boldsymbol{\theta}_0$ is the true vector of parameters.

If we want to test the simple null hypothesis which specifies that data come from $F=F_0(x)=F_{\boldsymbol{\theta}_0}$, let $U_i=F_0(X_i)$ be the transformed data and the order statistics based on transformed data are $U_{(1)}\leq U_{(2)}\leq\cdots\leq U_{(n-1)}$ and the spacings be defined as $D_i=U_{(i)}-U_{(i-1)}, i=1,\cdots,n$ with $U_{(0)}=0$ and $U_{(n)}=1$ and it is clear that the transformed data will follow a uniform distribution under the null hypothesis. Now using the transformed data and instead of constructing the usual empirical distribution function which is a step function, we use the line segments to join the points where there are jumps so that it becomes a piecewise linear function, *i.e.*, define the following smoothed empirical distribution as given by Kale [10] (p 44),

$$\hat{F}_n(x)=\frac{i}{n}+\frac{x-U_{(i)}}{nD_i},U_{(i-1)}\le x\le U_{(i)},i=1,\cdots,n \tag{5}$$

The density function of $F_U(x)$ is

$$f_U(x)=\frac{\mathrm{d}F_U(x)}{\mathrm{d}x}=1,0\le x\le 1 \text{ and } f_U(x)=0\text{, elsewhere.} \tag{6}$$

The procedure to smooth the empirical distribution using transformed data is similar to the procedure of constructing an ogive function when data have been grouped into intervals and we need to smooth the empirical distribution function, see Klugman *et al.* [15] (p 212) for the ogive function.

The smoothed empirical distribution function admits the following elementary density estimate as density,

$$\hat{f}_n(x)=\frac{1}{nD_i},U_{(i-1)}\le x\le U_{(i)},i=1,\cdots,n \tag{7}$$

and it can be obtained easily without requiring a kernel and specifying a window.

3. Density Based Statistics for Simple Null Hypothesis

It is not difficult to see that under the simple null hypothesis the transformed data follow the uniform distribution with density function given by $f_U(x)=1,0\le x\le 1$ and $f_U(x)=0$ elsewhere and an appropriate goodness-of-fit statistic can be based on

$$d_\Phi\left(\hat{f}_n,f_U\right)=1-\int_{-\infty}^{\infty}\hat{f}_n^{\delta}(x)f_U^{1-\delta}(x)\mathrm{d}x=1-\int_0^1\hat{f}_n^{\delta}(x)\mathrm{d}x$$

since $f_U(x)=1,0\le x\le 1$ and $f_U(x)=0$ elsewhere.

Therefore, if we can find a real number $k>0$ so that

$$n^k d_\Phi\left(\hat{f}_n,f_U\right) \tag{8}$$

has an asymptotic distribution which no longer depends on the functional form of F_0, the statistic for testing goodness-of-fit can be based on the statistic $V=n^k d_\Phi\left(\hat{f}_n,f_U\right)$ and the test will have power since with $n\to\infty$, this will imply $n^k\to\infty$ and with $d_\Phi\left(\hat{f}_n,f_U\right)$ being a measure which can detect whether the sample is drawn from an assumed distribution, the statistic given by expression (8) will be able to detect departure from the null hypothesis in probability as $n\to\infty$.

In fact, we do not need to require $d_\Phi\left(\hat{f}_n,f_U\right)\ge 0$, $d_\Phi\left(\hat{f}_n,f_U\right)$ only needs to be defined up to a positive multiplicative constant and an additive constant. In fact, all we need is to have the following main property with the following situations:

1) If the sample is drawn from a distribution F and $\hat{f}_n$ is constructed based on the transformation $d_\Phi\left(\hat{f}_n,f_U\right)\Big|_2$ applied on the data and we compute

$d_\Phi\left(\hat{f}_n, f_U\right)$, we shall use the notation $d_\Phi\left(\hat{f}_n, f_U\right)\Big|_1$ for $d_\Phi\left(\hat{f}_n, f_U\right)$ computed under situation 1.

2) If the sample is drawn from a distribution G and $\hat{f}_n$ is constructed based on the transformation F applied on the data and we compute $d_\Phi\left(\hat{f}_n, f_U\right)$ and we shall use the notation $d_\Phi\left(\hat{f}_n, f_U\right)\Big|_2$ when it is computed under situation 2.

Then, we should have $d_\Phi\left(\hat{f}_n, f_U\right)\Big|_2 > d_\Phi\left(\hat{f}_n, f_U\right)\Big|_1$ in probability and in general this property holds using Theorem 1 as given by Kirmani and Alam [16] (p 200). Consequently, by having this property, eventually we can detect departure from the null hypothesis as $n \to \infty$ using a statistic based on $d_\Phi\left(\hat{f}_n, f_U\right)$.

Furthermore, if we can simplify the expression of V so that we can have an equivalent statistic which serves the same purpose and it is simpler to compute then it is interesting to use its equivalent form. It turns out that this is the case as the statistic can be expressed as a simple function of spacings. However, by relating to the discrepancy measure, the test based on such a statistic can be seen to be consistent. This statistic parallels the one proposed by Beran [11] (p 458) which uses the Hellinger distance with the original data and a kernel density estimate. It is simpler to obtain this statistic than the one given by Beran.

Now we shall examine the component $\int_0^1 \hat{f}_n^\delta(x)\,\mathrm{d}x$ of $d_\Phi\left(\hat{f}_n, f_U\right)$. We observe that

$$\int_0^1 \hat{f}_n^\delta(x)\,\mathrm{d}x = \sum_{i=1}^n \int_{U_{(i-1)}}^{U_{(i)}} \left(\frac{1}{nD_i}\right)^\delta \mathrm{d}x = \sum_{i=1}^n D_i\left(nD_i\right)^{-\delta}$$

and it can be re-expressed as

$$\int_0^1 \hat{f}_n^\delta(x)\,\mathrm{d}x = \sum_{i=1}^n (n)^{-\delta}\left(D_i\right)^{1-\delta} = \frac{n^{-\delta}}{n^\alpha}\sum_{i=1}^n \left(nD_i\right)^\alpha, \alpha = 1-\delta, \qquad (9)$$

see Kirmani and Alam [17] for goodness of fit test using statistic of the form $sgn(r)\sum_{i=1}^n D_i^{1+r}, r > -1$. Our approach here is slightly different as we relate the statistic with the pseudo distance and by doing so we can examine the test statistic from two angles and see whether the test statistic might have good power or not for large samples.

Using results as given in section 2 by Luong [2], we can conclude that $\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(nD_i\right)^\alpha$ has an asymptotic Normal distribution as a Central limit Theorem can be applied to the expression. By letting the mean and variance of W^α with W which follows a standard exponential distribution and using the moment generating function of the log-gamma distribution which states that

$$E\left(W^c\right) = \Gamma(1+c) \quad \text{for} \quad c > -1,$$

we have $\mu = E\left(W^\alpha\right) = \Gamma(1+\alpha)$ with $\Gamma(.)$ being the usual gamma function,

$$\sigma^2 = E\left(W^{2\alpha}\right) - \left(E\left(W^\alpha\right)\right)^2 = \Gamma(1+2\alpha) - \left(\Gamma(1+\alpha)\right)^2,$$

so that we have an asymptotic normal distribution for the test statistic P defined below and if we need to emphasize the dependence on $\boldsymbol{\theta}_0$, we also use the notation

$P = P(\boldsymbol{\theta}_0)$ and it is given by $P = \dfrac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(nD_i)^{\alpha} - \sqrt{n}\mu}{\sigma} \xrightarrow{L} N(0,1)$ with $\xrightarrow{L}$ to denote convergence in distribution or in law.

Therefore, if we look for the scaling factor k using expression (9) we should consider

$$k = \delta + \alpha - \frac{1}{2} = 1 - \frac{1}{2} = \frac{1}{2} \quad \text{since} \quad \alpha = 1 - \delta \tag{10}$$

and with $k > 0$, this will imply $n^k \to \infty$ with $n \to \infty$ and the test will have power for large samples. This magnified scaling factor n^k will make the statistic sensitive to departure from the null hypothesis when the sample sizes become large. This property of consistency of the tests did not seem to have received attention in the literature.

The asymptotic distribution of the statistic

$$V = n^k d_{\Phi}\left(\hat{f}_n, f_U\right) = n^k - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(nD_i)^{\alpha}$$

which can also be represented as $V = n^k - Z$ with $Z = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(nD_i)^{\alpha}$ which follows a Normal distribution with mean $\mu_Z = \sqrt{n}\mu$ and variance $\sigma_Z^2 = \sigma^2$. We should reject H_0 if $V \geq b$ and b is chosen so that the approximate probability (*aP*) given by the asymptotic distribution with $aP(V \geq b) = p$ for a test of size *p*. From the following equalities,

$$p = aP(V \geq b) = aP\left(V - n^k \geq b - n^k\right) = aP\left(-Z \geq b - n^k\right)$$

and hence, $aP\left(Z \leq n^k - b\right) = p$, we reject H_0 if

$$P = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(nD_i)^{\alpha} - \sqrt{n}\mu}{\sigma} \leq z_p \tag{11}$$

with z_p being the p-th percentile of the standard normal distribution. Note that we need to restrict α to be positive and near 0 as within this range for α, GSP methods are efficient for estimation. The test based on Matusita's distance or Hellinger distance with $\alpha = \frac{1}{2}$ using transformed data with the statistic as given by expression (11) might make the test having low power for large samples when the null hypothesis is composite; see Kirmani and Alam [16], Kirmani [17] for the statistic using $\alpha = \frac{1}{2}$ but with $\alpha = \frac{1}{2}$ the GSP method is not efficient for estimation. Testing for the null hypothesis which is composite will be considered subsequently.

4. Density Based Statistics for Null Composite Hypothesis

For testing the null composite hypothesis which specifies that data come from the parametric family $\{F_{\boldsymbol{\theta}}\}$ and since $\boldsymbol{\theta}_0$ is unknown, we proceed to estimate $\boldsymbol{\theta}$ by minimizing a chosen pseudo distance based on a value fixed for α and subsequently use the same pseudo distance to construct the statistic. The second step is similar to the construction of the statistic for simple null hypothesis as it consists of replacing the unknown parameters with their estimates and once they are replaced, the statistic will have a similar form as in the simple null hypothesis case.

Parallel to the simple null hypothesis case, we transform the data and let $U_i = F_{\boldsymbol{\theta}}(X_i), i = 1, \cdots, n-1$ and since in this case the U_i's depend on $\boldsymbol{\theta}$, we also use the notation $U_i = U_i(\boldsymbol{\theta})$ and define as before the spacings but they will depend on $\boldsymbol{\theta}$ and given by

$$D_i(\boldsymbol{\theta}) = U_{(i)}(\boldsymbol{\theta}) - U_{(i-1)}(\boldsymbol{\theta}), i = 1, \cdots, n$$

with $U_{(0)}(\boldsymbol{\theta}) = 0, U_{(n)}(\boldsymbol{\theta}) = 1$ for all $\boldsymbol{\theta} \in \theta$, θ is the parameter space assumed to be compact.

Since the transformed data $U_i = F_{\boldsymbol{\theta}}(X_i), i = 1, \cdots, n-1$ depends on $\boldsymbol{\theta}$, we might want to call them pseudo transformed data which leads to define the following pseudo elementary density estimate as

$$\hat{f}_n^{\boldsymbol{\theta}}(x) = \frac{1}{nD_i(\boldsymbol{\theta})}, U_{(i-1)} \le x \le U_{(i)}, i = 1, \cdots, n, \tag{12}$$

using the notations

$$U_{(i-1)} = U_{(i-1)}(\boldsymbol{\theta}), U_{(i)} = U_{(i)}(\boldsymbol{\theta}).$$

We estimate first $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$ which minimizes $d_{\Phi}\left(\hat{f}_n^{\boldsymbol{\theta}}, f_U\right)$ which is equivalent to minimize $-\sum_{i=1}^{n}\left(nD_i(\boldsymbol{\theta})\right)^{\alpha}$ or maximizes $\sum_{i=1}^{n}\left(nD_i(\boldsymbol{\theta})\right)^{\alpha}$.

The estimators given by the vector $\hat{\boldsymbol{\theta}}$ are Generalized spacing (GSP) estimators with a GSP method using $h(x) = -x^{\alpha}, 0 < \alpha < 1$ with α which is specified.

The goodness-of-fit test statistic can be based on $V = V\left(\hat{\boldsymbol{\theta}}\right)$ which depends on $\hat{\boldsymbol{\theta}}$ this time and can be similarly constructed as for the simple hypothesis case with

$$V\left(\hat{\boldsymbol{\theta}}\right) = n^{k(\alpha)} d_{\Phi}\left(\hat{f}_n^{\hat{\boldsymbol{\theta}}}, f_U\right), \Phi(x) = 1 - x^{\alpha}.$$

Argue as in the case of simple null hypothesis it leads to consider the equivalent statistic

$$P\left(\hat{\boldsymbol{\theta}}\right) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(nD_i\left(\hat{\boldsymbol{\theta}}\right)\right)^{\alpha} - \sqrt{n}\mu}{\sigma} \tag{13}$$

We shall show subsequently that we have the equality in distribution

$$P\left(\hat{\boldsymbol{\theta}}\right) =^{d} P\left(\boldsymbol{\theta}_0\right), \tag{14}$$

$\boldsymbol{\theta}_0$ is the true vector of parameters, by showing

$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(nD_i\left(\hat{\boldsymbol{\theta}}\right)\right)^{\alpha} =^{d} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(nD_i\left(\boldsymbol{\theta}_0\right)\right)^{\alpha}$, $\boldsymbol{\theta}_0$ is the vector of true parameters, so that we reject the composite null hypothesis if

$$P\left(\hat{\boldsymbol{\theta}}\right) \le z_p . \tag{15}$$

The statistic is similar to the one used for the simple null hypothesis case. All we need is to replace $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}$ in the expression of the statistic used for the simple hypothesis.

Observe that we can expand the expression

$$H\left(\boldsymbol{\theta}_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(nD_i\left(\boldsymbol{\theta}_0\right)\right)^{\alpha} \tag{16}$$

around $\hat{\boldsymbol{\theta}}$ using a Taylor's type of expansion technique, a technique which is also used in the proof for asymptotic normality of $\hat{\boldsymbol{\theta}}$ as given in section (3.2) by Luong [2] and let $\boldsymbol{H}'\left(\boldsymbol{\theta}\right)$ and $\boldsymbol{H}''\left(\boldsymbol{\theta}\right)$ be respectively the first derivative vector and second derivative matrix and we have $\boldsymbol{H}'\left(\hat{\boldsymbol{\theta}}\right) = 0$ since $\hat{\boldsymbol{\theta}}$ maximizes $H\left(\boldsymbol{\theta}\right)$ or minimizes $-H\left(\boldsymbol{\theta}\right)$. Therefore, we have the following equality in probability

$$\boldsymbol{H}\left(\boldsymbol{\theta}_0\right) =^{p} \boldsymbol{H}\left(\hat{\boldsymbol{\theta}}\right) + \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \left(\frac{1}{2\sqrt{n}} \boldsymbol{H}''\left(\boldsymbol{\theta}_0\right)\right) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \tag{17}$$

with $\left(\frac{1}{\sqrt{n}} \boldsymbol{H}''\left(\boldsymbol{\theta}_0\right)\right)$ and $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)'$ being bounded in probability as $\left(\frac{1}{2\sqrt{n}} \boldsymbol{H}''\left(\boldsymbol{\theta}_0\right)\right)$ is up to a constant equivalent to the matrix $\boldsymbol{A}_0$ which is given by Luong [2] (p 629-630) and $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$ has an asymptotic distribution. Now with $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, we then have $H\left(\hat{\boldsymbol{\theta}}\right) = H\left(\boldsymbol{\theta}_0\right) + o_p(1)$ with $o_p(1)$ being an expression which converges to 0 in probability. Therefore,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(nD_i\left(\hat{\boldsymbol{\theta}}\right)\right)^{\alpha} =^{d} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(nD_i\left(\boldsymbol{\theta}_0\right)\right)^{\alpha} \tag{18}$$

which justifies the use of expression (15).

The same type of property has been shown to hold for the asymptotic distribution of the Moran's statistic with Maximum spacing estimators for testing goodness of fit for parametric models, see Cheng and Stephens [3] (p 390).

The GSP methods with $\alpha \in [0.01, 0.05]$ might be recommended as it induces a minimum loss of efficiency across parametric models, say with $\alpha = 0.01$, the loss of efficiency comparing to the MSP method is around two percents only for all the parametric model and the GSP method has some robustness properties like the minimum Hellinger distance estimator proposed by Beran [11] (see Remark 2 given by Luong [2]). With $\alpha = 0.05$, the loss of efficiency is around ten per cents. It is simpler to implement GSP methods for estimation, parameter hypothesis testing and goodness of fit testing than implementing Hellinger distance method as proposed by Beran and practitioners might want to use the GSP methods based on this class for their applied works. The equivalent statistic for testing the composite using Hellinger distance which is also density based as proposed Beran [11] (p 459) might be more difficult to implement for practitioners. The same can be said for the test based on Genalized Method of Moments (GMM) using a continuum moment of conditions as proposed by Carrasco and Florens [18] (p 813). Of course, these conclusions are based on the distribution of the parametric family has a closed form expression so that data can be easily transformed which leads to consider spacings for inferences. These density based tests share with chi-square tests with parameters estimated with the minimum chi-square methods to have an asymptotic distribution which does not depend on $\{F_\theta\}$ and since there is no need to choose intervals to perform these tests; this might be viewed as an advantage over the corresponding chi-square tests for testing simple null hypothesis and for composite null hypothesis.

5. Tied Observations

In this section, we would like to make the following remark by pointing out that in a data set which is not large and there are many tied observations, it might be preferred to use a GSP method instead of the MSP method as the MSP method is based on minimizing $-\sum_{i=1}^{n}\log\left(D_i(\boldsymbol{\theta})\right)$ or maximizing $\sum_{i=1}^{n}\log\left(D_i(\boldsymbol{\theta})\right)$ and for two ordered observations which are tied, this implies a spacing is equal to 0 and log of this spacing is undefined, see table 3 in Cheng and Stephens [3] (p 391) which gives a real life data set where tied measurements are recorded.

Cheng and Stephens [3] (p 391) also proposed methods to handle tied observations for the use of MSP method but tied measurements do not cause numerical difficulties for the GSP method as discussed and there is little loss of efficiency using a GSP method and a GSP method might be more robust than the MSP method, see Remark 2 as given by Luong [2].

6. Discussions

In this section, we touch upon the question of power analysis for these density based tests. Power analysis for null hypothesis which specifies functions is more complicated than Pitman efficiency analysis when parameters are scalars, see Lehmann [19] (p 158-187) for the classical set up with scalars as parameters instead of functions.

Here, under the null hypothesis a function or functions are specified, this makes the study of power more complicated even for the simplest case when the null hypothesis H_0 is simple which specifies the data comes from F_0 or equivalently the transformed data comes from a standard uniform density with density function $f_U(x)=1$ for $0 \le x \le 1$ and $f_U(x)=0$, elsewhere.

For power study, often a sequence of tests based on a sequence of functions which belongs to the alternative hypothesis H_a is considered. Sethuraman and Rao [8], used the following sequence of functions $\{h_n(x)\}$ with $h_n(x)=1+\frac{l(x)}{n^{1/4}}$, $l(x)$ is twice differentiable with bounded second derivative and

$$\int_0^1 l(x)\,\mathrm{d}x = 0, h_n(x) \to f_U(x), 0 \le x \le 1 .$$

For theoretical works and Pitman efficiencies, the focus is on best tests based on a chosen sequence of functions but it might not provide a complete answer for applications as an optimum statistic might no longer be optimum if another sequence of functions are chosen. In applications, the distributions belonging to the alternative hypothesis which are useful and commonly used might not have been included in the analysis for theoretical works. This makes the assessment of power difficult using theoretical analysis especially when parameters are functions instead of scalars, see Lehman [19] for the classical set up on Pitman efficiency analysis using scalars and parameters belong to the real line. The functional space is more complicated than the real line.

Cheng and Stephens [3] (p 386) recognized this problem and pointed out that power depends on the alternative hypothesis and to get some ideas on the power of these tests often large scale simulations seem to be needed and many parametric families should be considered as given by the alternative hypothesis which are techniques that Zhang [9] has used to conduct power studies for some EDF tests. We do not have resources for these large scale simulation studies. These tests have not been not used extensively and in the future if they are used more frequently and concomitantly with GSP methods for estimation in applications, we will have better ideas on power of these tests.

7. Conclusion

In a previous paper, we have studied estimation, asymptotic properties, robustness and parameter hypothesis testing using GSP methods. In this paper we have adopted the view that GSP methods are minimum density based distance methods using transformed data or equivalently spacings so that estimation and model testing can be treated in a unified way. Model validation via goodness-of-fit tests and construction of density based tests are treated in this paper. We have shown that these statistics for testing come at no extra cost once a GSP method is used for fitting a parametric model and might be useful for assessment of the model in practice. These tests are simple to perform and practitioners might want to use these tests concomitantly with GSP estimation especially when sample sizes are relatively large. For some real life data sets, GSP methods might be preferred over MSP method for estimation and chosen for their robustness property, efficiency and the flexibility to handle tied observations and finally tests statistics for goodness-of-fit can be constructed at no extra cost. The last feature is not shared by maximum likelihood (ML) method.

Acknowledgements

The helpful and constructive comments of a referee which lead to an improvement of the presentation of the paper and support from the editorial staffs of Open Journal of Statistics to process the paper are all gratefully acknowledged.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Ghosh, K. and Jammalamadaka, S.R. (2001) A General Estimation Method Using Spacings. *Journal of Statistical Planning Inference*, **93**, 71-82. https://doi.org/10.1016/S0378-3758(00)00160-9

[2] Luong, A. (2018) Unified Asymptotic Results for Maximum Spacing and Generalized Spacing Method for Continuous Model. *Open Journal of Statistics*, **8**, 614-639. https://doi.org/10.4236/ojs.2018.83040

[3] Cheng, R.C.H. and Stephens, M.A. (1989) A Goodness-of-Fit Test Using Moran's Statistic with Estimated Parameters. *Biometrika*, **76**, 385-392. https://doi.org/10.1093/biomet/76.2.385

[4] Anderson, T.W. and Darling, D.A. (1954) A Test of Goodness of Fit. *Journal of American Statistical Association*, **49**, 765-769. https://doi.org/10.1080/01621459.1954.10501232

[5] Boos, D. (1982) Minimum Anderson Darling Estimation. *Communications in Statistics, Theory and Methods*, **11**, 2747-2774. https://doi.org/10.1080/03610928208828420

[6] Stephens, M.A. (1986) Tests Based on EDF Statistics in Goodness-of-Fit Techniques. d'Agostino, R.B. and Stephens, M.A., Eds., Marcel Dekker, New York.

[7] Greenwood, P.E. and Nikulin, S.M. (1996) A Guide to Chi-Squared Testing. Wiley, New York.

[8] Sethuraman and Rao, J.S. (1970) Pitman Efficiencies of Tests Based on Spacings in Nonparametric Techniques in Statistical Inference. Puri, M.L., Ed. Cambridge University Press, Cambridge.

[9] Zhang, J. (2002) Powerful Goodness-of-Fit Tests Based on the Likelihood Ratio. *Journal of the Royal Statistical Society, Series B*, **62**, 281-294. https://doi.org/10.1111/1467-9868.00337

[10] Kale, B.K. (1969) Unified Derivation of Tests of Goodness of Fit Based on Spacings. *Sankhya, Series A*, **31**, 43-48.

[11] Beran, R. (1977) Minimum Hellinger Distance Estimates for Parametric Models. *Annals of Statistics*, **5**, 445-463. https://doi.org/10.1214/aos/1176343842

[12] Pollard, D. (1980) The Minimum Distance Method of Testing. *Metrika*, **27**, 43-70. https://doi.org/10.1007/BF01893576

[13] Ali, S.M. and Silvey, S.D. (1966) A Generalized Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society*, **28**, 813-828.

[14] Pardo, L. (2006) Statistical Inference Based on Divergence Measures. Chapman and Hall, Boca Raton.

[15] Klugman, S.A., Panjer, H.H. and Willmot, G.E. (2012) Loss Models: From Data to Decisions. 4th Edition, Wiley, New York.

[16] Kirmani, S.N.U.A. and Alam, S.N. (1974) On Goodness of Fit Tests Based on Spacings. *Sankhya, Series A*, **36**,197-203.

[17] Kirmani, S.N.U.A. (1973) On a Goodness of Fit Test Based on Matusita's Distance. *Annals of the Institue of Mathematical Statistics*, **24**, 493-500.

https://doi.org/10.1007/BF02479394

[18] Carrasco, M. and Florens, J.-P. (2000) Generalization of GMM to a Continuum of Moment Conditions. *Econometric Theory*, **16**, 797-834. https://doi.org/10.1017/S0266466600166010

[19] Lehmann, E.L. (1999) Elements of Large Sample Theory. Springer, New York. https://doi.org/10.1007/b98855