6-2001

Neyman's smooth test and its use in econometrics

Anil K. BERA
University of Illinois at Urbana-Champaign

Aurobindo GHOSH
Singapore Management University, AUROBINDO@SMU.EDU.SG

Neyman's Smooth Test and Its Applications in Econometrics

Anil K Bera
Department of Economics
University of Illinois at Urbana-Champaign
1206 S. Sixth Street
Champaign, IL 61801; USA
phone: 217-333-4596 fax: 217-244-6678
email: anil@fisher.econ.uiuc.edu

Aurobindo Ghosh
Department of Economics
University of Illinois at Urbana-Champaign
1206 S. Sixth Street
Champaign, IL 61801; USA.
phone: 217-333-2179 fax: 217-244-6678
email: a-ghosh1@uiuc.edu

Abstract

The following essay is a reappraisal of the role of the smooth test proposed by Neyman (1937) in the context of current applications in econometrics. We revisit the derivation of the smooth test and put it into the perspective of the existing literature on tests based on probability integral transforms suggested by early pioneers such as R.A.Fisher (1930, 1932) and Karl Pearson (1933, 1934) and the other tests for goodness-of-fit. Our discussion touches data-driven and other methods of testing and inference on the order of the smooth test and the motivation and choice of orthogonal polynomials used by Neyman and others. We review other locally most powerful unbiased tests and look at their differential geometric interpretations in terms of Gaussian curvature of the power hypersurface and review some recent advances. Finally, we venture into some applications in econometrics by evaluating density forecast calibrations discussed by Diebold, Gunther and Tay (1998) and others. We discuss the use of smooth tests in survival analysis as done by Peña (1998), Gray and Pierce (1985) and in tests based on p-values and other probability integral transforms suggested in Meng (1994). Uses in diagnostic analysis of stochastic volatility models are also mentioned. Along with our narrative of the smooth test and its various applications, we also provide some historical anecdotes and sidelights that we think interesting and instructive.

1 Introduction

Statistical hypothesis testing has a long history. Neyman and Pearson (1933 [80]) traced its origin to Bayes (1763 [8]). However, the systematic use of hypothesis testing began only after the publication of Pearson's (1900 [86]) goodness-of-fit test. Even after 100 years, this statistic is very much in use in a variety of applications and is regarded as one of the 20 most important scientific breakthroughs in the twentieth century. Simply stated, Pearson's (1900 [86]) test statistic is given by

$$P_{\chi^2} = \sum_{j=1}^{q} \frac{(O_j - E_j)^2}{E_j}, \tag{1}$$

where O_j denotes the observed frequency and E_j is the (expected) frequency that would be obtained under the distribution of the null hypothesis, for the j^{th} class, $j = 1, 2, ..., q$. Although K. Pearson (1900 [86]) was an auspicious beginning to twentieth century statistics, the basic foundation of the theory of hypothesis testing was laid more than three decades later by Neyman and Pearson (1933 [80]). For the first time the concept of "optimal test" was introduced through the analysis of "power functions." A general solution to the problem of maximizing power subject to a size condition was obtained for the single parameter case when both the null and the alternative hypotheses were simple. The result was the celebrated Neyman-Pearson(N-P) lemma, which provides a way to construct an uniformly most powerful (UMP) test. A UMP test, however, rarely exists, and, therefore, it is necessary to restrict optimal tests to a suitable subclass that requires the test to satisfy other criteria such as *local* optimality and *unbiasedness*. Neyman and Pearson (1936 [81]) derived a locally most powerful unbiased (LMPU) test for the one-parameter case and called the corresponding critical region the "type-A region." Neyman and Pearson (1938 [82]) obtained the LMPU test for testing a *multi-parameter* hypothesis and termed the resulting critical region as the "type-C region."

Neyman's (1937 [76]) smooth test is based on the type-C critical region. Neyman suggested the test to rectify some of the drawbacks of the Pearson goodness-of-fit statistic given in (1). He noted that it is not clear how the class intervals should be determined and that the distributions under the alternative hypothesis were not "smooth." By smooth densities, Neyman meant those that are close to and have few intersections with the null density function. In his effort to find a smooth class of alternative distributions, Neyman (1937 [76]) considered the probability integral transformation of the density, say $f(x)$, under the null hypothesis and showed that the probability integral transform is distributed as uniform in $(0, 1)$ irrespective of the specification of $f(x)$. Therefore, in some sense, "all" testing problems can be converted into testing only *one kind of hypothesis*.

Neyman was not the first to use the idea of probability integral transformation to reformulate

the hypothesis testing problem into a problem of testing uniformity. E. Pearson (1938 [84]) discussed how Fisher (1930 [41], 1932 [43]) and K. Pearson (1933 [87], 1934 [88]) also developed the same idea. They did not, however, construct any formal test statistic. What Neyman (1937 [76]) achieved was to integrate the ideas of tests based on the probability integral transforms in a concrete fashion along with designing "smooth" alternative hypotheses based on normalized Legendre polynomials.

The aim of this paper is modest. We put the Neyman (1937 [76]) smooth test in perspective with the existing methods of testing available at that time; evaluate it based on the current state of the literature; derive the test from the widely used Rao (1948 [93]) score principle of testing, and, finally, we discuss some of the applications of the smooth test in econometrics and statistics.

Section 2 discusses the genesis of probability integral transforms as a criterion for hypothesis testing with Subsections 2.1 through 2.3 putting Neyman's smooth test in perspective in the light of current research in probability integral transforms and related areas. Section 2.4 discusses the main theorem of Neyman's smooth test. Section 3 gives a formulation of the relationship of Neyman's smooth test as Rao's score (RS) and other optimal tests. Here, we also bring up the notion of unbiasedness as a criterion for optimality in tests and also puts forward the differential geometric interpretation. In Section 4 we look at different applications of Neyman's smooth tests. In particular, we discuss inference using different orthogonal polynomials, density forecast evaluation and calibration in financial time series data, survival analysis and applications in stochastic volatility models. The paper concludes in Section 5.

2 Background and Motivation

2.1 Probability integral transform and the combination of probabilities from independent tests

In statistical work, sometimes, we have a number of *independent* tests of significance for the same hypothesis, giving different probabilities (like p-values). The problem is to combine results from different tests in a single hypothesis test. Let us suppose that we have carried out n independent tests with p-values, $y_1, y_2, ..., y_n$. Tippett (1931 [112], p. 142) suggested a procedure based on the minimum p-value, i.e., on $y_{(1)} = \min(y_1, y_2, ..., y_n)$. If all n null hypotheses are valid, then $y_{(1)}$ has a standard beta distribution with parameters $(1, n)$. One can also use any smallest p-value, $y_{(r)}$, the r^{th} smallest p-value in place of $y_{(1)}$, as suggested by Wilkinson (1951 [115]).The statistic $y_{(r)}$ will have a beta distribution with parameters $(r, n - r + 1)$. It is apparent that there is some arbitrariness in this approach through the choice of r. Fisher (1932 [43], Section 21.1, pp. 99-100) suggested a simpler and more appealing procedure based on the product of

the p-values, $\lambda = \prod_{i=1}^{n} y_i$. K. Pearson (1933 [87]) also considered the same problem in a more general framework along with his celebrated problem of goodness-of-fit. He came up with the same statistic λ, but suggested a different approach to compute the p-value of the comprehensive test.[1]

In the current context, Pearson's goodness-of-fit problem can be stated as follows. Let us suppose that we have a sample of size n, $x_1, x_2, ..., x_n$. We want to test whether it comes from a population with probability density function (pdf) $f(x)$. Then, the p-values (rather 1−p-values), $y_i\,(i = 1, 2, ..., n)$ can be defined as

$$y_i = \int_{-\infty}^{x_i} f(\omega)\, d\omega. \tag{2}$$

Suppose that we have n tests of significance and the values of our test statistics are T_i, $i = 1, 2, ..., n$, then,

$$y_i = \int_{-\infty}^{T_i} f_{T_i}(t)\, dt, \tag{3}$$

where $f_T(t)$ is the pdf of T. To find the distribution or the p-value of $\lambda = y_1 y_2 ... y_n$ both Fisher and Karl Pearson started in a similar way, though Pearson was more explicit in his derivation. In this exposition, we will follow Karl Pearson's approach.

Let us simply write

$$y = \int_{-\infty}^{x} f(\omega)\, d\omega, \tag{4}$$

and the pdf of y as $g(y)$. Then, from (4) we have

$$dy = f(x)\, dx, \tag{5}$$

and we also have from change of variables

$$g(y)\, dy = f(x)\, dx. \tag{6}$$

[1]To differentiate his methodology from that of Fisher, K. Pearson added the following note at the end of his paper:

> "After this paper had been set up Dr Egon S. Peason drew my attention to Section 21.1 in the Fourth Edition of Professor R.A. Fisher's *Statistical Methods for Research Workers*, 1932. Professor Fisher is brief, but his method is essentially what I had thought to be novel. He uses, however, a χ^2 method, not my incomplete Γ−function solution; ... As my paper was already set up and illustrates, more amply than Professor Fisher's two pages, some of the advantages and some of the difficulties of the new method, which may be helpful to students, I have allowed it to stand."

Hence, combining (5) and (6),

$$g(y) = 1,\ 0 < y < 1, \tag{7}$$

i.e., y has a uniform distribution over $(0,1)$. From this point Pearson's and Fisher's treatments differ.

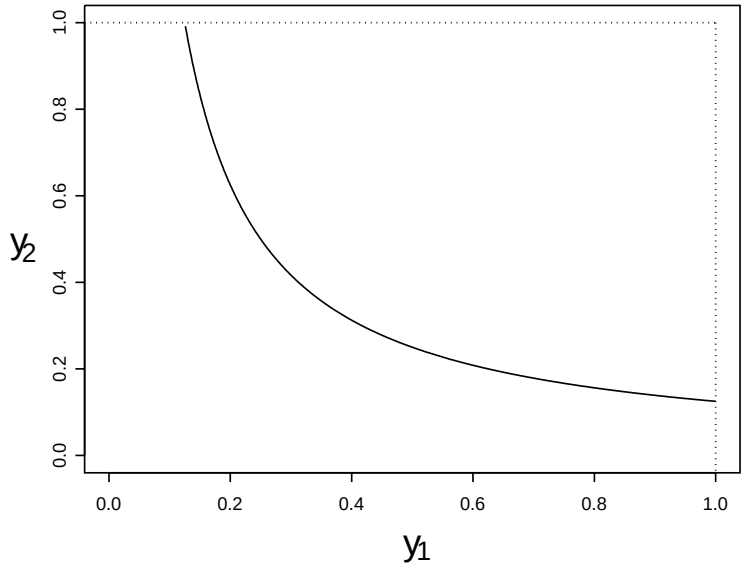

Figure 1. Surface of the equation $y_1 y_2 = \lambda_2$ for $\lambda_2 = 0.125$.

The surface given by the equation

$$\lambda_n = y_1 y_2 ... y_n \tag{8}$$

is termed "$n-$*hyperboloid*" by Pearson, and what is needed is the volume of $n-$cuboid (since $0 < y_i < 1$, $i = 1, 2, ..., n$) cut off by the $n-$hyperboloid. We show the surface λ_n in Figures 1 and 2 for $n = 2$ and $n = 3$, respectively. After considerable algebraic derivation Pearson (1933 [87], p. 382) showed that the p-value for λ_n is given by

$$Q_{\lambda_n} = 1 - P_{\lambda_n} = 1 - I(n-1, -\ln \lambda_n), \tag{9}$$

where $I(.)$ is the incomplete gamma function ratio defined by [Johnson and Kotz (1970 [57], p. 167)]

$$I(n-1, u) = \frac{1}{\Gamma(n)} \int_0^{u\sqrt{n}} t^{n-1} e^{-t} dt. \tag{10}$$

We can use the test statistic Q_{λ_n} both for combining a number of independent tests of significance and for the goodness-of -fit problem.

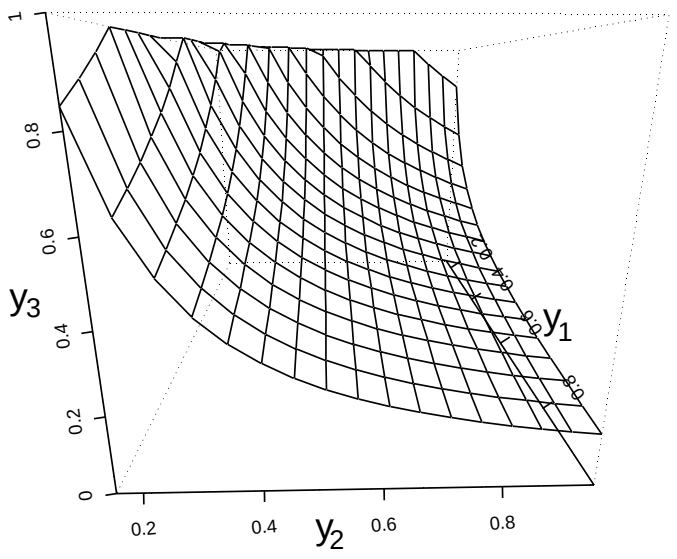

Figure 2. Surface of the equation $y_1 y_2 y_3 = \lambda_3$ for $\lambda_3 = 0.125$.

Pearson (1933 [87], p. 383) stated this very clearly:

> "If Q_{λ_n} be very small, we have obtained an extremely rare sample, and we have then to settle in our minds whether it is more reasonable to suppose that we have drawn a very rare sample at one trial from the supposed parent population, or that our hypothesis as to the character of the parent population is erroneous, i.e., that the sample $x_1, x_2, ..., x_n$ was not drawn from the supposed population."

Pearson (1933 [87], p. 403) even criticized his own celebrated χ^2 statistic, stating that the χ^2 test in equation (1) has the disadvantage of giving the same resulting probability whenever the individuals are in the same class. This criticism has been repeatedly stated in the literature. Bickel and Doksum (1977 [18], p. 378) have put it rather succinctly, "in problems with continuous variables there is a clear loss of information, since the χ^2 test utilizes only the number of observations in intervals rather than the observations themselves." Tests based on P_{λ_n} (or Q_{λ_n}) do not have this problem. Also, when the sample size n is small, grouping the observations in several classes is somewhat hazardous for the inference.

As we mentioned, Fisher's main aim was to combine n p-values from n independent tests to obtain a single probability. By putting $Z = -2 \ln Y$ where $Y \sim U(0,1)$, we see that the pdf of Z is given by

$$f_Z(z) = \frac{1}{2} e^{-\frac{z}{2}}, \tag{11}$$

i.e., Z has a χ^2_2 distribution. Then, if we combine n independent z_i's by

$$\sum_{i=1}^{n} z_i = -2\sum_{i=1}^{n} \ln y_i = -2\ln \lambda_n, \tag{12}$$

this statistic will be distributed as χ^2_{2n}. For quite some time this statistic was known as Pearson's P_λ. Rao (1952 [94], p. 44) called it Pearson's P_λ distribution [see also Maddala (1977 [72], pp. 47-48)]. Rao (1952 [94], pp. 217-219) used it to combine several independent tests of the difference between means and on tests for skewness. In the recent statistics literature this is described as Fisher's procedure [for example, see Becker (1977 [9])].

In summary, to combine several independent tests, both Fisher and K. Pearson arrived at the same problem of testing the uniformity of $y_1, y_2, ..., y_n$. Undoubtedly, Fisher's approach was much simpler, and it is now used more often in practice. We should, however, add that K. Pearson had a much broader problem in mind, including testing goodness-of-fit. In that sense, Pearson's (1933 [87]) paper was more in the spirit of Neyman's (1937 [76]) that came four years later.

As we discussed above, the fundamental basis of Neyman's smooth test is the result that when $x_1, x_2, ..., x_n$ are independent and identically distributed (*IID*) with a common density $f(.)$, then the probability integral transforms $y_1, y_2, ..., y_n$ defined in equation (2) are *IID*, $U(0,1)$ random variables. In econometrics, however, we very often have cases in which $x_1, x_2, ..., x_n$ are not *IID.* In that case we can use Rosenblatt's (1952 [101]) generalization of the above result.

Theorem 1 (Rosenblatt(1952)) *Let $(X_1, X_2, ..., X_n)$ be a random vector with absolutely continuous density function $f(x_1, x_2, ..., x_n)$. Then, the n random variables defined by*

$$\begin{aligned} Y_1 &= P(X_1 \leq x_1), Y_2 = P(X_2 \leq x_2 | X_1 = x_1), \\ ..., Y_n &= P(X_n \leq x_n | X_1 = x_1, X_2 = x_2, ..., X_{n-1} = x_{n-1}) \end{aligned}$$

are IID $U(0,1)$.

The above result can immediately be seen from the following observation that

$$\begin{aligned} P(Y_i \leq y_i, i = 1, 2, ..., n) &= \int_0^{y_1} \int_0^{y_2} ... \int_0^{y_n} f(x_1) dx_1 f(x_2|x_1) dx_2 ... f(x_n|x_1, ..., x_{n-1}) dx_n \\ &= \int_0^{y_1} \int_0^{y_2} ... \int_0^{y_n} dt_1 dt_2 ... dt_n \\ &= y_1 y_2 ... y_n. \end{aligned} \tag{13}$$

Hence, $Y_1, Y_2, ..., Y_n$ are *IID* $U(0,1)$ random variables. Quesenberry (1986 [92], pp. 239-240)

discussed some applications of this result in goodness-of-fit tests. In Section 4.2, we will discuss its use in density forecast evaluation.

2.2 Summary of Neyman (1937)

As we mentioned earlier, Fisher (1932 [43]) and Karl Pearson (1933 [87], 1934 [88]) suggested tests based on the fact that the probability integral transform is uniformly distributed for an *IID* sample under the null hypothesis (or the correct specification of the model). What Neyman (1937 [76]) achieved was to integrate the ideas of tests based on probability integral transforms in a concrete fashion along with the method of designing alternative hypotheses using orthonormal polynomials.[2] Neyman's paper began with a criticism of Karl Pearson's χ^2 test given in (1). First, in Pearson's χ^2 test, it is not clear how the q class intervals should be determined. Secondly, the expression in (1) does not depend on the order of positive and negative differences $(O_j - E_j)$. Neyman (1980 [78], pp. 20-21) gives an extreme example represented by two cases. In the first, the signs of the consecutive differences $(O_j - E_j)$ are not the same, and in the other there is a run of, say, a number of "negative" differences, followed by a sequence of "positive" differences. These two possibilities might lead to similar values of P_{χ^2}, but Neyman (1937 [76], 1980 [78]) argued that in the second case the goodness-of-fit should be more in doubt, even if the value of χ^2 happens to be the small. In the same spirit, the χ^2-test is more suitable for discrete data and the corresponding distributions under the alternative hypothesis are not "smooth." By smooth alternatives Neyman (1937 [76]) meant those densities that have *few* intersections with the null density function and that are *close* to the null.

Suppose we want to test the null hypothesis (H_0) that $f(x)$ is the true density function for the random variable X. The specification of $f(x)$ will be *different* depending on the problem at hand. Neyman (1937 [76], pp. 160-161) first transformed *any* hypothesis testing problem of this type to testing only *one kind of hypothesis*.[3] Let us state the result formally through the

[2]It appears that Jerzy Neyman was not aware of the above papers by Fisher and Karl Pearson. To link Neyman's test to these papers, and possibly since Neyman's paper appeared in a rather recondite journal, Egon Pearson (Pearson 1938 [84]) published a review article in *Biometrika*. At the end of that article Neyman added the following note to express his regret for overlooking, particularly, the Karl Pearson papers:

> "I am grateful to the author of the present paper for giving me the opportunity of expressing my regret for having overlooked the two papers by Karl Pearson quoted above. When writing the paper on the "Smooth test for goodness of fit" and discussing previous work in this direction, I quoted only the results of H.Cramér and R. v. Mises, omitting mention of the papers by K. Pearson. The omission is the more to be regretted since my paper was dedicated to the memory of Karl Pearson."

[3]In the context of testing several different hypotheses, Neyman (1937 [76], p. 160) argued this quite eloquently as follows :

> "If we treat all these hypotheses separately, we should define the set of alternatives for each of them and this would in practice lead to a dissection of a unique problem of a test for goodness of

following simple derivation.

Suppose, that under H_0, $x_1, x_2, ..., x_n$ are independent and identically distributed with a common density function $f(x|H_0)$. Then, the probability integral transform

$$y \equiv y(x) = \int_{-\infty}^{x} f(u|H_0)\, du, \tag{14}$$

has a pdf given by

$$h(y) = f(x|H_0) \frac{\partial x}{\partial y} \quad \text{for } 0 < y < 1. \tag{15}$$

Differentiating (14) with respect to y, we have

$$1 = f(x|H_0) \frac{dx}{dy}. \tag{16}$$

Substituting this into (15), we get

$$h(y) \equiv h(y|H_0) = 1 \quad \text{for } 0 < y < 1. \tag{17}$$

Therefore, testing H_0 is equivalent to testing whether the random variable Y has a uniform distribution in the interval $(0, 1)$, irrespective of the specification of the density $f(.)$.

---

fit into a series of more or less disconnected problems.

However, this difficulty can be easily avoided by substituting for any particular form of the hypothesis H_0, that may be presented for test, another hypothesis, say h_0, which is equivalent to H_0 and which has always the same analytical form. The word equivalent, as used here, means that whenever H_0 is true, h_0 must be true also and inversely, if H_0 is not correct then h_0 must be false."

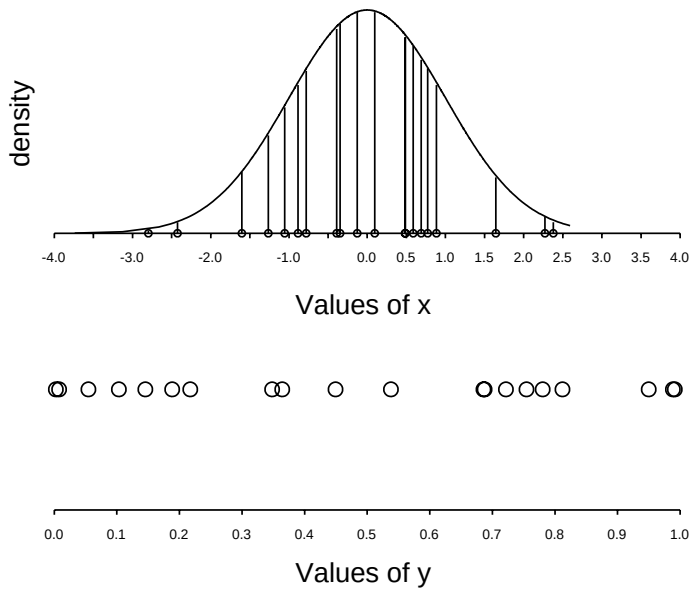

Figure 3. Distribution of the probability integral transform when H_0 is true.

Figure 3 drawn following Pearson (1938 [84], Figure 1) illustrates the relationship between x and y, when $f(.)$ is taken to be $N(0,1)$ and $n = 20$. Let us denote $f(x|H_1)$ as the distribution under the alternative hypothesis H_1. Then, Neyman (1937 [76]) pointed out [see also Pearson (1938 [84], p. 138)] that the distribution of Y under H_1 is given by

$$f(y|H_1) = f(x|H_1).\frac{dx}{dy} = \left.\frac{f(x|H_1)}{f(x|H_0)}\right|_{x=p(y)}, \text{ for } 0 < y < 1, \tag{18}$$

where $x = p(y)$ means a solution to the equation (14). This looks more like a likelihood-ratio and will be different from 1 when H_0 is not true. As an illustration, in Figure 4 we plot values of Y when Xs are drawn from $N(2,1)$ instead of $N(0,1)$, and we can immediately see that these y values [probability integral transforms of values from $N(2,1)$ using the $N(0,1)$ density] are not uniformly distributed.

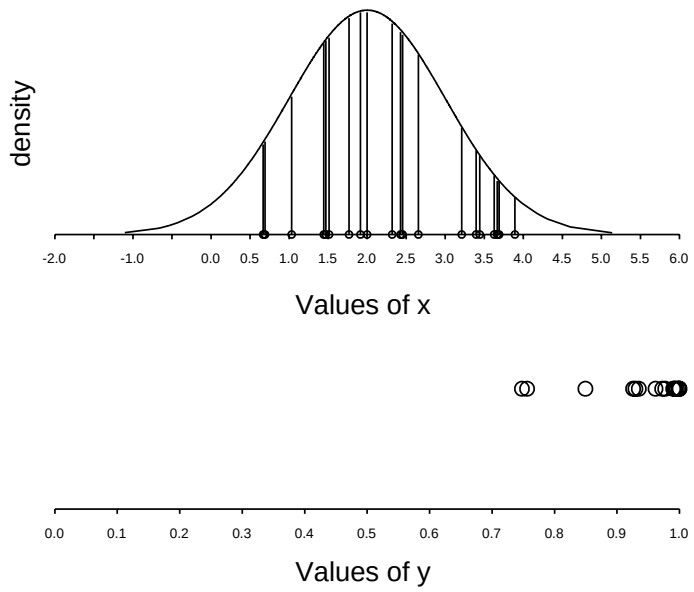

Figure 4. Distribution of the probability integral transform when H_0 is false.

Neyman (1937 [76], p. 164) considered the following smooth alternative to the uniform density:

$$h(y) = c(\theta) \exp\left[\sum_{j=1}^{k} \theta_j \pi_j(y)\right], \tag{19}$$

where $c(\theta)$ is the constant of integration depending only on $(\theta_1, ..., \theta_k)$, $\pi_j(y)$'s are orthonormal polynomials of order j satisfying

$$\int_0^1 \pi_i(y)\,\pi_j(y)\,dy = \delta_{ij}, \quad \text{where } \delta_{ij} = \begin{array}{ll} 1 & if \quad i = j \\ 0 & if \quad i \neq j. \end{array} \tag{20}$$

Under $H_0 : \theta_1 = \theta_2 = ... = \theta_k = 0$, since $c(\theta) = 1$, $h(y)$ in (19) reduces to the uniform density in (17).

Using the generalized Neyman-Pearson (N-P) lemma, Neyman (1937 [76]) derived the locally most powerful symmetric test for $H_0 : \theta_1 = \theta_2 = ... = \theta_k = 0$ against the alternative H_1 : at least one $\theta_i \neq 0$,for small values of $\theta_i's$. The test is symmetric in the sense that the asymptotic power of the test depends only on the distance,

$$\lambda = \left(\theta_1^2 + ... + \theta_k^2\right)^{\frac{1}{2}}, \tag{21}$$

between H_0 and H_1. The test statistic is

$$\Psi_k^2 = \sum_{j=1}^{k} \frac{1}{n} \left[\sum_{i=1}^{n} \pi_j (y_i) \right]^2 , \tag{22}$$

which under H_0 asymptotically follows a central χ_k^2 and under H_1 follows a non-central χ_k^2 with non-centrality parameter λ^2. Neyman's approach requires the computation of the probability integral transform (14) in terms of Y. It is, however, easy to recast the testing problem in terms of the original observations on X and pdf, say, $f(x;\gamma)$. Writing (14) as $y = F(x;\gamma)$ and defining $\pi_i(y) = \pi_i(F(x;\gamma)) = q_i(x;\gamma)$, we can express the orthogonality condition (20) as

$$\int_0^1 \{\pi_i (F(x;\gamma))\} \{\pi_j (F(x;\gamma))\} dF(x;\gamma) = \int_0^1 \{q_i(x;\gamma)\} \{q_j(x;\gamma)\} f(x;\gamma)\, dx = \delta_{ij}. \tag{23}$$

Then, from (19) the alternative density in terms of X takes the form

$$\begin{aligned} g(x;\gamma,\theta) &= h(F(x;\gamma)) \frac{dy}{dx} \\ &= c(\theta;\gamma) \exp\left[\sum_{j=1}^{k} \theta_j q_j (x;\gamma) \right] f(x;\gamma). \end{aligned} \tag{24}$$

Under this formulation the test statistic Ψ_k^2 reduces to

$$\Psi_k^2 = \sum_{j=1}^{k} \frac{1}{n} \left[\sum_{i=1}^{n} q_j (x_i;\gamma) \right]^2 , \tag{25}$$

which has the same asymptotic distribution as before. In order to implement this we need to replace the nuisance parameter γ by an efficient estimate $\hat{\gamma}$, and that will not change the asymptotic distribution of the the test statistic [see Thomas and Pierce (1979 [111])], although there could be some possible change in the variance of the test statistic [see for example, Boulerice and Ducharme (1995 [19])]. Later we will relate this test statistic to a variety of different tests and discuss its properties.

2.3 Interpretation of Neyman's (1937) results and its relation to some later works

Egon Pearson (1938 [84]) provided an excellent account of Neyman's ideas, and he emphasized the need for consideration of the possible alternatives to the hypothesis tested. He discussed both the cases of testing goodness-of-fit and combining results of independent tests of significance.

Another issue that he addressed is whether the upper or the lower tail probabilities (or p-values) should be used for combing different tests. The upper tail probability [see equation (2)]

$$y_i^{'} = \int_{x_i}^{\infty} f(\omega)\, d\omega = 1 - y_i, \tag{26}$$

under H_0, is also uniformly distributed in $(0,1)$, and, hence, $-2\sum_{i=1}^{n} \ln y_i'$ is distributed as χ^2_{2n} following our derivations in equations (11) and (12). Therefore, the tests based on y_i and $y_i^{'}$ will be the same as far as their size is concerned but will, in general, differ in terms of power. Regarding other aspects of the Neyman's smooth test for goodness-of-fit, as Egon Pearson (1938 [84], pp. 140 and 148) pointed out, the greatest benefit that it has over other tests is that it can detect the direction of the alternative when the null hypothesis of correct specification is rejected. The divergence can come from any combination of location, scale, shape etc. By selecting the orthogonal polynomials π_j's in equation (20) judiciously, we can seek power of the smooth test in specific directions. We think that is one of the most important advantages of Neyman's smooth test over Fisher and Karl Pearson's suggestion of using only one function of y_i values, namely $\sum_{i=1}^{n} \ln y_i$. Pearson (1938 [84], p. 139) plotted the function $f(y|H_1)$ [see equation (18)] for various specifications of H_1 when $f(x|H_0)$ is $N(0,1)$ and demonstrated that $f(y|H_1)$ can take a variety of non-linear shapes depending on the nature of the departures, such as the mean being different from zero, the variance being different from 1, and the shape being non-normal. It is easy to see that a single function like $\ln y$ cannot capture all of the non-linearities. However, as Neyman himself argued, a linear combination of orthogonal polynomials might do the job.

Neyman's use of the density function (19) as an alternative to the uniform distribution is also of fundamental importance. Fisher (1922 [40], p. 356) used this type of exponential distribution to demonstrate the equivalence of the method of moments and the maximum likelihood estimator in special cases. We can also derive (19) analytically by *maximizing* the entropy $-E[\ln h(y)]$ subject to the moment conditions $E[\pi_j(y)] = \eta_j$ (say), $j = 1, 2, ..., k$, with parameters θ_j, $j = 1, 2, ..., k$, as the Lagrange multipliers determined by k moment constraints [for more on this see, for example, Bera and Bilias (2001c [13])]. In the information theory literature, such densities are known as *minimum discrimination information* models in the sense that the density $h(y)$ in (19) has the minimum distance from the uniform distributions satisfying the above k moment conditions [see Soofi (2000 [107])].[4] We can say that while testing the density $f(x;\gamma)$, the alternative density function $g(x;\gamma,\theta)$ in equation (24), has a minimum distance from $f(x;\gamma)$ satisfying the moment conditions like $E[q_j(x)] = \eta_j$, $j = 1, ..., k$. From that point

[4]For small values of θ_j $(j = 1, 2, ..., k)$, $h(y)$ will be a smooth density close to uniform when k is moderate, say equal to 3 or 4. However, if k is large then $h(y)$ will present particularities which would not correspond to the intuitive idea of smoothness (Neyman 1937 [76], p. 165). From maximum entropy point of view, each additional moment condition, add some more roughness and possibly some peculiarities of the data to the density.

of view, $g(x;\gamma,\theta)$ is "truely" a *smooth* alternative to the density $f(x;\gamma)$. Looking from another perspective, we can see from (19) that $\ln h(y)$ is essentially a *linear* combination of several polynomials in y. Similar densities have been used in the log-spline model literature [see, for instance, Stone and Koo (1986 [109]) and Stone (1990 [108])].

2.4 Formulation and derivation of the smooth test

Neyman (1937 [76]) derived a locally most powerful symmetric (regular) unbiased test (critical region) for $H_0 : \theta_1 = \theta_2 = ... = \theta_k = 0$ in (19), and he called it an unbiased critical region of type-C. This type-C critical region is an extension of the locally most powerful unbiased (LMPU) test (type-A region) of Neyman and Pearson (1936 [81]) from a single parameter case to a multiparameter situation. We first briefly describe the type-A test for testing $H_0 : \theta = \theta_0$ (where θ is a scalar) for local alternatives of the form $\theta = \theta_0 + \frac{\delta}{\sqrt{n}}$, $0 < \delta < \infty$. Let $\beta(\theta)$ be the power function of the test. Then, assuming differentiability at $\theta = \theta_0$ and expanding $\beta(\theta)$ around $\theta = \theta_0$, we have

$$\begin{aligned}\beta(\theta) &= \beta(\theta_0) + (\theta - \theta_0)\beta'(\theta_0) + \frac{1}{2}(\theta - \theta_0)^2 \beta''(\theta_0) + o\left(n^{-1}\right) \\ &= \alpha + \frac{1}{2}(\theta - \theta_0)^2 \beta''(\theta_0) + o\left(n^{-1}\right), \end{aligned} \tag{27}$$

where α is the size of the test, and unbiasedness requires that the "power" should be minimum at $\theta = \theta_0$, and, hence, $\beta'(\theta_0) = 0$. Therefore, to maximize the local power we need to maximize $\beta''(\theta_0)$. This leads to the well-known LMPU test or the type-A critical region. In other words, we can maximize $\beta''(\theta_0)$ subject to two side conditions, namely, $\beta(\theta_0) = \alpha$ and $\beta'(\theta_0) = 0$. These ideas are illustrated in Figure 5. For a locally optimal test, the power curve should have maximum curvature at the point C (where $\theta = \theta_0$), which is equivalent to minimizing the distance like the chord AB.

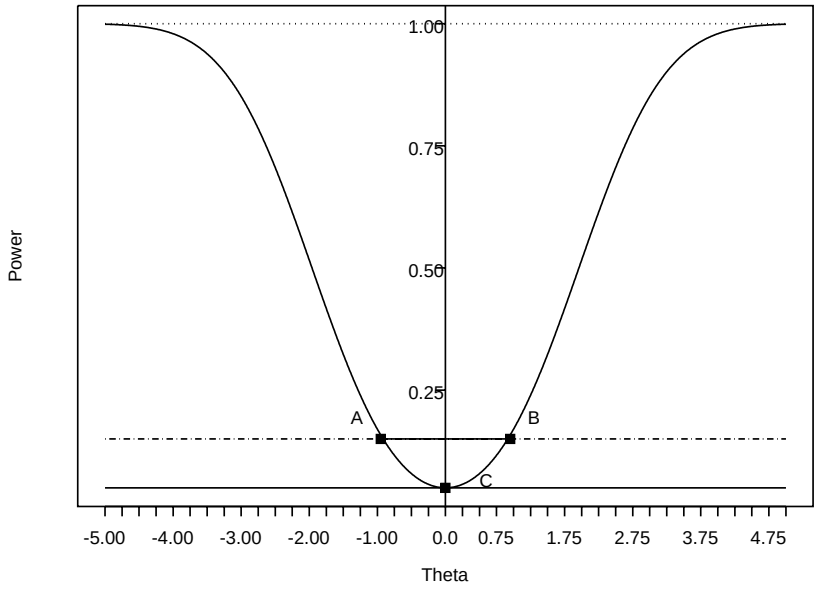

Figure 5. Power curve for one parameter unbiased test.

Using the generalized Neyman-Pearson (N-P) lemma, the optimal (type-A) critical region is given by

$$\frac{d^2L(\theta_0)}{d\theta^2} > k_1\frac{dL(\theta_0)}{d\theta} + k_2L(\theta_0), \tag{28}$$

where $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ is the likelihood function, while the constants k_1 and k_2 are determined through the side conditions of size and local unbiasedness. The critical region in (28) can be expressed in terms of the derivatives of the log-likelihood function $l(\theta) = \ln(L(\theta))$ as

$$\frac{d^2l(\theta_0)}{d\theta^2} + \left[\frac{dl(\theta_0)}{d\theta}\right]^2 > k_1\frac{dl(\theta_0)}{d\theta} + k_2. \tag{29}$$

If we denote the score function as $s(\theta) = \frac{dl(\theta)}{d\theta}$ and its derivative as $s'(\theta)$, then (29) can be written as

$$s'(\theta_0) + [s(\theta_0)]^2 > k_1s(\theta_0) + k_2. \tag{30}$$

Neyman (1937 [76]) faced a difficult problem since his test of $H_0 : \theta_1 = \theta_2 = ... = \theta_k = 0$ in (19) involved testing a parameter vector, namely, $\theta = (\theta_1, \theta_2, ..., \theta_k)'$. Let us now denote the power function as $\beta(\theta_1, \theta_2, ..., \theta_k) = \beta(\theta) \equiv \beta$. Assuming, that the power power function $\beta(\theta)$ is twice differentiable in the neighborhood of $H_0 : \theta = \theta_0$, Neyman (1937 [76], pp. 166-167) formally required that an unbiased critical region of type-C of size α should satisfy the following conditions :

$$1.\ \ \beta\left(0,0,...0\right)=\alpha. \tag{31}$$

$$2.\ \ \beta_j=\left.\frac{\partial\beta}{\partial\theta_j}\right|_{\theta=0}=0,\ j=1,...k. \tag{32}$$

$$3.\ \ \beta_{ij}=\left.\frac{\partial^2\beta}{\partial\theta_i\partial\theta_j}\right|_{\theta=0}=0,\ i,j=1,...k,i\neq j. \tag{33}$$

$$4.\ \ \beta_{jj}=\left.\frac{\partial^2\beta}{\partial\theta_j^2}\right|_{\theta=0}=\left.\frac{\partial^2\beta}{\partial\theta_1^2}\right|_{\theta=0},\ j=2,...k. \tag{34}$$

And finally, over all such critical regions satisfying the conditions (31)-(34), the common value of $\left.\frac{\partial^2\beta}{\partial\theta_j^2}\right|_{\theta=0}$ is the maximum.

To interpret the above conditions it is instructive to look at the $k=2$ case. Here, we will follow the more accessible exposition of Neyman and Pearson (1938 [82]).[5]

By taking the Taylor series expansion of the power function $\beta\left(\theta_1,\theta_2\right)$ around $\theta_1=\theta_2=0$, we have

$$\beta\left(\theta_1,\theta_2\right)=\beta\left(0,0\right)+\theta_1\beta_1+\theta_2\beta_2+\frac{1}{2}\left(\theta_1^2\beta_{11}+2\theta_1\theta_2\beta_{12}+\theta_2^2\beta_{22}\right)+o\left(n^{-1}\right). \tag{35}$$

The type-C *regular* unbiased critical region has the following properties, (i) $\beta_1=\beta_2=0$, which is the condition for any unbiased test; (ii) $\beta_{12}=0$ to ensure that a small positive and a small negative deviations in the θ's should be controlled *equally* by the test; (iii) $\beta_{11}=\beta_{22}$, so that equal departures from $\theta_1=\theta_2=0$ have the same power in all directions; and (iv) the common value of β_{11} (or β_{22}) is maximized over all critical regions satisfying the conditions (i) to (iii). If a critical region satisfies only (i) and (iv), it is called a *non-regular* unbiased critical region of type-C. Therefore, for a type-C regular unbiased critical region, the power function is given by

$$\beta\left(\theta_1,\theta_2\right)=\alpha+\frac{1}{2}\beta_{11}\left(\theta_1^2+\theta_2^2\right). \tag{36}$$

As we can see from Figure 6 , maximization of power is equivalent to the minimization of the exposed top circle in the figure. In order to find out whether we really have a LMPU test, we need to look at the second-order condition, i.e., the Hessian matrix of the power function $\beta\left(\theta_1,\theta_2\right)$ in

[5]After the publication of Neyman (1937 [76]), Neyman in collaboration with Egon Pearson wrote another paper, Neyman and Pearson (1938 [82]) that included a detailed account of the unbiased critical region of type-C. This paper belongs to the famous Neyman-Pearson series on the Contribution to the Theory of Testing Statistical Hypotheses. For historical sidelights on their collaboration see Pearson (1966 [85]).

(35) evaluated at $\theta = 0$,

$$B_2 = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{12} & \beta_{22} \end{bmatrix} \tag{37}$$

should be positive definite, i.e., $\beta_{11}\beta_{22} - \beta_{12}^2 > 0$ should be satisfied.

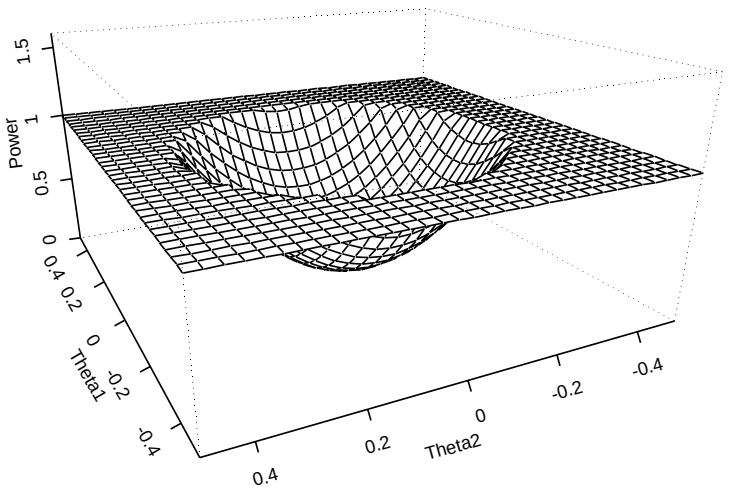

Figure 6. Power surface for two-parameter unbiased test.

We should also note from (35) that, for the unbiased test,

$$\theta_1^2\beta_{11} + 2\theta_1\theta_2\beta_{12} + \theta_2^2\beta_2^2 = \text{constant} \tag{38}$$

represents what Neyman and Pearson (1938 [82], p. 39) termed the ellipse of *equidetectability*. Once we impose the further restriction of "regularity," namely, the conditions (*ii*) and (*iii*) above, the concentric *ellipses* of equidetectability becomes concentric circles of the form (see Figure 6),

$$\beta_{11}\left(\theta_1^2 + \theta_2^2\right) = \text{constant}. \tag{39}$$

Therefore, the resulting power of the test will be a function of the distance measure, $(\theta_1^2 + \theta_2^2)$; Neyman (1937 [76]) called this the symmetry property of the test.

Using generalized N-P lemma, Neyman and Pearson (1938 [82], p. 41) derived the type-C unbiased critical region as

$$L_{11}(\theta_0) \geq k_1\left[L_{11}(\theta_0) - L_{22}(\theta_0)\right] + k_2 L_{12}(\theta_0) + k_3 L_1(\theta_0) + k_4 L_2(\theta_0) + k_5 L(\theta_0), \tag{40}$$

where $L_i(\theta) = \frac{\partial L(\theta)}{\partial \theta_i}$, $i = 1, 2$, $L_{ij}(\theta) = \frac{\partial^2 L(\theta)}{\partial \theta_i \partial \theta_j}$, $i, j = 1, 2$ and $k_i (i = 1, 2, ..., 5)$ are constants determined from the size and the three side conditions $(i) - (iii)$.

The critical region (40) can also be expressed in terms of the derivatives of the log-likelihood function $l(\theta) = \ln L(\theta)$. Let us denote $s_i(\theta) = \frac{\partial l(\theta)}{\partial \theta_i}$, $i = 1, 2$ and $s_{ij}(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j}$, $i, j = 1, 2$. Then, it is easy to see that

$$L_i(\theta) = s_i(\theta) L(\theta), \tag{41}$$

$$L_{ij}(\theta) = [s_{ij}(\theta) + s_i(\theta) s_j(\theta)] L(\theta), \tag{42}$$

where $i, j = 1, 2$. Using these, (40) can be written as

$$\begin{gathered}
(1 - k_1) s_{11}(\theta_0) + k_1 s_{22}(\theta_0) - k_2 s_{12}(\theta_0) + (1 - k_1) s_1^2(\theta_0) + \\
k_1 s_2^2(\theta_0) - k_2 s_1(\theta_0) s_2(\theta_0) - k_3 s_1(\theta_0) - k_4 s_2(\theta_0) + k_5 \geq 0 \\
\Rightarrow \left[s_{11}(\theta_0) - s_1^2(\theta_0)\right] - k_1 \left[s_{11}(\theta_0) - s_{22}(\theta_0) + s_1^2(\theta_0)\right] - \\
k_2 [s_1(\theta_0) s_2(\theta_0) + s_{12}(\theta_0)] - k_3 s_1(\theta_0) - k_4 s_2(\theta_0) + k_5 \geq 0.
\end{gathered} \tag{43}$$

When we move to the general multiple parameter case $(k > 2)$, the analysis remains essentially the same. We will then need to satisfy Neyman's conditions (31)-(34). In the general case, the Hessian matrix of the power function evaluated at $\theta = \theta_0$ in equation (37) has the form

$$B_k = \begin{bmatrix} \beta_{11} & \beta_{12} & ... & \beta_{1k} \\ \beta_{12} & \beta_{22} & ... & \beta_{2k} \\ ... & ... & ... & ... \\ \beta_{1k} & \beta_{2k} & ... & \beta_{kk} \end{bmatrix}. \tag{44}$$

Now for the LMPU test B_k should be positive definite i.e., all the principle cofactors of this matrix should be positive. For this general case, it is hard to express the type-C critical region in a simple way as in (40) or (43). However, as Neyman (1937 [76]) derived, the resulting test procedure takes a very simple form given in the next theorem.

Theorem 2 (Neyman (1937)) *For large n, the type-C regular unbiased test (critical region) is given by,*

$$\Psi_k^2 = \sum_{j=1}^{k} u_j^2 \geq C_\alpha, \tag{45}$$

where $u_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \pi_j(y_i)$ and the critical point C_α is determined from $P[\chi_k^2 \geq C_\alpha] = \alpha$.

Neyman (1937 [76], pp. 186-190) further proved that the limiting form of the power function of this test is given by

$$\left(\frac{1}{\sqrt{2\pi}}\right)^k \int \cdots \int_{\sum u_i^2 \geq C_\alpha} e^{-\frac{1}{2}\sum_{j=1}^k (u_j-\theta_j)^2} du_1 du_2 ... du_k. \tag{46}$$

In other words, under the alternative hypothesis $H_1 : \theta_j \neq 0$, at least for some $j = 1, 2, ...k$, the test statistic Ψ_k^2 approaches a non-central χ_k^2 distribution with the non-centrality parameter $\lambda = \sum_{j=1}^k \theta_j^2$. From (36), we can also see that the power function for this general k case is

$$\beta(\theta_1, \theta_2, ..., \theta_k) = \alpha + \frac{1}{2}\beta_{11} \sum_{j=1}^k \theta_j^2 = \alpha + \beta_{11}\lambda. \tag{47}$$

Since, the power depends only on the "distance" $\sum_{j=1}^k \theta_j^2$ between H_0 and H_1, Neyman called this test *symmetric*.

Unlike Neyman's earlier work with Egon Pearson on general hypothesis testing, the smooth test went unnoticed in the statistics literature for quite some time. It is quite possible that Neyman's idea of explicitly deriving a test *statistic* from the very first principles under a very general framework was well ahead of its time, and its usefulness in practice was not immediately apparent.[6] Isaacson (1951 [56]) was the first notable paper that referred to Neyman's work while proposing the type-D unbiased critical region based on Gaussian or total curvature of the power hypersurface. However, D.E. Barton was probably the first to do a serious analysis of Neyman's smooth test. In a series of papers (1953 [3], 1955 [5], 1956 [6]), he discussed its small sample distribution, applied the test to discrete data and generalized the test to some extent to the composite null hypothesis situation [see also Hamdan (1962 [48], 1964 [49])]. In the next section we demonstrate that the smooth tests are closely related to some of the other more popular tests. For example, the Pearson χ^2 goodness-of-fit statistic can be derived as a special case of the smooth test. We can also derive Neyman's smooth test statistic Ψ^2 in a simple way using Rao's (1948) score test principle.

[6] Reid (1982 [100], p. 149) described an amusing anecdote. In 1937, W. E. Deming was preparing publication of Neyman's lectures by the United States Department of Agriculture. In his lecture notes Neyman misspelled *smooth* when referring to the smooth test. "I don't understand the reference to 'Smouth'," Deming wrote Neyman, " Is that name of a statistician?".

3 The Relationship of Neyman's Smooth Test with Rao's Score and Other Locally Optimal Tests

Rayner and Best (1989 [99]) provided an excellent review of smooth tests of various categorized and uncategorized data and related procedures. They also elaborated on many interesting, little-known results. For example, Pearson's (1900 [86]) P_{χ^2} statistic in (1) can be obtained as a Neyman's smooth test for a categorized hypothesis. To see this result let us write the probability of the j^{th} class in terms of our density (24) under the alternative hypothesis as

$$p_j = c(\theta) \exp\left[\sum_{i=1}^{r} \theta_i h_{ij}\right] \theta_{j0}, \tag{48}$$

where θ_{j0} is the value p_j under the null hypothesis, $j = 1, 2, ..., q$. In (48), h_{ij} are values taken by a random variable H_i with $P(H_i = h_{ij}) = \theta_{j0}$, $j = 1, 2, ..., q$; $i = 1, 2, ..., r$. These h_{ij} are also orthonormal with respect to the probabilities θ_{j0}. Rayner and Best (1989 [99], pp. 57-60) showed that the smooth test for testing $H_0 : \theta_1 = \theta_2 = ... = \theta_r = 0$ is the same as the Pearson's P_{χ^2} in (1) with $r = q - 1$. Smooth-type tests can be viewed as a compromise between an omnibus test procedure such as Pearson's χ^2, which generally has low power in all directions, and more specific tests with power directed only towards certain alternatives.

Rao and Poti (1946 [98]) suggested a locally most powerful (LMP) one-sided test for the *one-parameter* problem. This test criterion is the precursor to Rao's (1948 [93]) celebrated score test in which the basic idea of Rao and Poti (1946 [98]) is generalized to the *multiparameter* and composite hypothesis cases. Suppose the null hypothesis is composite, like $H_0 : \delta(\theta) = 0$, where $\delta(\theta)$ is an $r \times 1$ vector function of $\theta = (\theta_1, \theta_2, ..., \theta_k)$ with $r \leq k$. Then, the general form of Rao's score (RS) statistic is given by

$$RS = s\left(\tilde{\theta}\right)' \mathcal{I}\left(\tilde{\theta}\right)^{-1} s\left(\tilde{\theta}\right), \tag{49}$$

where $s(\theta)$ is the score function $\frac{\partial l(\theta)}{\partial \theta}$, $\mathcal{I}(\theta)$ is the information matrix $E\left[-\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'}\right]$ and $\tilde{\theta}$ is the restricted maximum likelihood estimator (MLE) of θ. Asymptotically, under H_0, RS is distributed as χ^2_r. Let us derive the RS test statistic for testing $H_0 : \theta_1 = \theta_2 = ... = \theta_k = 0$ in (24); so that the number of restrictions are $r = k$ and $\tilde{\theta} = 0$. We can write the log likelihood function as

$$l(\theta) = \sum_{i=1}^{n} \ln g(x_i; \theta) = \sum_{i=1}^{n} \ln c(\theta) + \sum_{i=1}^{n} \sum_{j=1}^{k} \theta_j q_j(x_i) + \sum_{i=1}^{n} \ln f(x). \tag{50}$$

For the time being we ignore the nuisance parameter γ, and later we will adjust the variance of the RS test when γ is replaced by an efficient estimator $\tilde{\gamma}$.

The score vector and the information matrix under H_0, are given by

$$s\left(\tilde{\theta}\right) = n\left.\frac{\partial \ln c\left(\theta\right)}{\partial \theta}\right|_{\theta=0} + \sum_{j=1}^{k}\sum_{i=1}^{n} q_j\left(x_i\right), \tag{51}$$

and

$$\mathcal{I}\left(\tilde{\theta}\right) = -n\left.\frac{\partial^2 \ln c\left(\theta\right)}{\partial \theta \partial \theta'}\right|_{\theta=0}, \tag{52}$$

respectively. Following Rayner and Best (1989 [99], pp. 77-80) and differentiating the identity $\int_{-\infty}^{\infty} g\left(x;\theta\right) dx = 1$ twice, we see that

$$\frac{\partial \ln c\left(\theta\right)}{\partial \theta_j} = -E_g\left[q_j\left(x\right)\right], \tag{53}$$

$$\frac{\partial^2 \ln c\left(\theta\right)}{\partial \theta_j \partial \theta_l} = -\int_{-\infty}^{\infty} q_j\left(x\right)\frac{\partial g\left(x;\theta\right)}{\partial \theta_l} dx,\ j,l = 1,2,...,k, \tag{54}$$

where $E_g\left[.\right]$ is expectation taken with respect to the density under the alternative hypothesis, namely, $g\left(x;\theta\right)$. For the RS test we need to evaluate everything at $\theta = 0$. From (53) it is easy to see that

$$\left.\frac{\partial \ln c\left(\theta\right)}{\partial \theta_j}\right|_{\theta=0} = 0, \tag{55}$$

and, thus, the score vector given in equation (51) simplifies to

$$s\left(\tilde{\theta}\right) = \sum_{j=1}^{k}\sum_{i=1}^{n} q_j\left(x_i\right). \tag{56}$$

From (24) we have

$$\frac{\partial \ln g\left(x;\theta\right)}{\partial \theta_l} = \frac{1}{g\left(x;\theta\right)}\frac{\partial g\left(x;\theta\right)}{\partial \theta_l} = \frac{\partial \ln c\left(\theta\right)}{\partial \theta_l} + q_l\left(x\right), \tag{57}$$

i.e.,

$$\frac{\partial g\left(x;\theta\right)}{\partial \theta_l} = \left[\frac{\partial \ln c\left(\theta\right)}{\partial \theta_l} + q_l\left(x\right)\right] g\left(x;\theta\right). \tag{58}$$

Hence, we can rewrite (54) and evaluate under H_0 as

$$\begin{aligned}\frac{\partial^2 \ln c(\theta)}{\partial\theta_j \partial\theta_l} &= -\int_{-\infty}^{\infty} q_j(x)\left[\frac{\partial \ln c(\theta)}{\partial\theta_l} + q_l(x)\right] g(x;\theta)\,dx \\ &= -\int_{-\infty}^{\infty} q_j(x)\left[-E_g\left[q_l(x)\right] + q_l(x)\right] g(x;\theta)\,dx \\ &= \int_{-\infty}^{\infty} q_j(x)\, q_i(x)\, g(x;\theta)\,dx \\ &= -Cov_g\left(q_j(x), q_l(x)\right) \\ &= \delta_{jl}, \end{aligned} \tag{59}$$

where $\delta_{jl} = 1$ when $j = l$; $= 0$ otherwise. Then, from (52), $\mathcal{I}\left(\tilde{\theta}\right) = nI_k$, where I_k is a $k-$dimensional identity matrix. This also means that the asymptotic variance-covariance matrix of $\frac{1}{\sqrt{n}} s\left(\tilde{\theta}\right)$ will be

$$V\left[\frac{1}{\sqrt{n}} s\left(\tilde{\theta}\right)\right] = I_k. \tag{60}$$

Therefore, using (49) and (56), the RS test can be simply expressed as

$$RS = \sum_{i=1}^{k} \frac{1}{n}\left[\sum_{i=1}^{n} q_j(x_i)\right]^2, \tag{61}$$

which is the "same" as Ψ_k^2 in (25), the test statistic for Neyman's smooth test. To see clearly why this result holds, let us go back to the expression of Neyman's type-C unbiased critical region in equation (40). Consider the case $k = 2$, then, using (56) and (59) we can put $s_j(\theta_0) = \sum_{i=1}^{n} q_j(x_i)$, $s_{jj}(\theta_0) = 1$, $j = 1, 2$ and $s_{12}(\theta_0) = 0$. It is quite evident that the second-order derivatives of the log-likelihood function do not play any role. Therefore, Neyman's test must be based only on score functions $s_1(\theta)$ and $s_2(\theta)$ evaluated at the null hypothesis $\theta = \theta_0 = 0$.

From the above facts, we can possibly assert that Neyman's smooth test is the *first* formally derived RS test. Given this connection between the smooth and the score tests, it is not surprising that Pearson's goodness-of-fit test is nothing but a categorized version of the smooth test as noted earlier. Pearson's test is also a special case of the RS test [see Bera and Bilias (2001a [11])]. To see the impact of estimation of the nuisance parameter γ [see equation (24)] on the RS statistic, let us use the result of Pierce (1982 [91]). Pierce established that for a statistic $U(.)$ depending on parameter vector γ, the asymptotic variances of $U(\gamma)$ and $U(\tilde{\gamma})$, where $\tilde{\gamma}$ is an efficient estimator of γ, is related by

$$Var\left[\sqrt{n}U(\tilde{\gamma})\right] = Var\left[\sqrt{n}U(\gamma)\right] - \lim_{n\longrightarrow\infty} E\left[\frac{\partial U(\gamma)}{\partial\gamma}\right]' Var\left(\sqrt{n}\tilde{\gamma}\right) \lim_{n\longrightarrow\infty} E\left[\frac{\partial U(\gamma)}{\partial\gamma}\right]. \tag{62}$$

Here, $\sqrt{n}U(\tilde{\gamma}) = \frac{1}{\sqrt{n}} s\left(\tilde{\theta}, \tilde{\gamma}\right) = \frac{1}{\sqrt{n}} \sum_{j=1}^{k} \sum_{i=1}^{n} q_j(x_i; \tilde{\gamma})$, $Var\left[\sqrt{n}U(\gamma)\right] = I_k$ as in (60), and, finally, $Var(\sqrt{n}\tilde{\gamma})$ is obtained from maximum likelihood estimation of γ under the null hypothesis. Furthermore, Neyman (1959 [77]) showed that

$$\lim_{n \longrightarrow \infty} E\left[\frac{\partial U(\gamma)}{\partial \gamma}\right] = -Cov\left[\sqrt{n}U(\gamma), \frac{1}{\sqrt{n}} \frac{\partial \ln f(x;\gamma)}{\partial \gamma}\right] = B(\gamma), \text{ say,} \tag{63}$$

and this can be computed for the given density $f(x;\gamma)$ under the null hypothesis. Therefore, from (62), the adjusted formula for the score function is

$$V\left[\frac{1}{\sqrt{n}} s\left(\tilde{\theta}, \tilde{\gamma}\right)\right] = I_k - B'(\gamma) Var\left(\sqrt{n}\tilde{\gamma}\right) B(\gamma) = V(\gamma) \text{ (say),} \tag{64}$$

which can be evaluated simply by replacing γ by $\tilde{\gamma}$. From (60) and (62), we see that in some sense the variance "decreases" when the nuisance parameter is replaced by its efficient estimator. Hence, the final form of the score or the smooth test will be

$$\begin{aligned} RS = \Psi^2 &= \frac{1}{n} s\left(\tilde{\theta}, \tilde{\gamma}\right)' V(\tilde{\gamma})^{-1} s\left(\tilde{\theta}, \tilde{\gamma}\right) \\ &= \frac{1}{n} s(\tilde{\gamma})' V(\tilde{\gamma})^{-1} s(\tilde{\gamma}), \end{aligned} \tag{65}$$

since for our case under the null $\tilde{\theta} = 0$. In practical applications, $V(\tilde{\gamma})$ may not be of full rank. In that case a generalized inverse of $V(\tilde{\gamma})$ could be used, and then the degree of freedom of the RS statistic will be the rank of $V(\tilde{\gamma})$ instead of k. Rayner and Best (1989 [99], pp. 78-80) also derived the same statistic [see also Boulerice and Ducharme (1995 [19])]; however, our use of Pierce (1982 [91]) makes the derivation of the variance formula a lot simpler.

Needless to say, since it is based on the score principle, Neyman's smooth test will share the optimal properties of the RS test procedure and will be asymptotically locally most powerful.[7] However, we should keep in mind all the restrictions that conditions (33) and (34) imposed while deriving the test procedure. The result is not as straightforward as testing the *single parameter* case for which we obtained the LMPU test in (28) by maximizing the power function. In the *multiparameter* case, the problem is that, instead of a power function, we have a power *surface* (or a power *hypersurface*). An ideal test would be one that has a power surface that has a maximum curvature along *every* cross-section at the point $H_0 : \theta = (0, 0, ..., 0)' = \theta_0$, say, subject to the conditions of size and unbiasedness. Such a test, however, rarely exists even for the simple cases. As Isaacson (1951 [56], p. 218) explained, if we maximize the curvature along one cross-section, it

[7] Recent work in higher order asymptotics support [see Chandra and Joshi (1983 [21]), Ghosh (1991 [44]) and Mukerjee (1994 [74])] the validity of Rao's conjecture about the optimality of the score test over its competitors under local alternatives particularly in locally asymptotically unbiased setting.

will generally cause the curvature to diminish along some other cross-section, and, consequently, the curvature cannot be maximized along *all* cross-sections simultaneously. In order to overcome this kind of problem Neyman (1937 [76]) required the type-C critical region to have constant power in the neighborhood of $H_0 : \theta = \theta_0$ along a given family of concentric ellipsoids. Neyman and Pearson (1938 [82]) called these the ellipsoids of equidetectability. However, one can only choose this family of ellipsoids if one knows the relative importance of power in different directions in an infinitesimal neighborhood of θ_0. Isaacson (1951 [56]) overcame this objection to the type-C critical region by developing a natural generalization of the Neyman-Pearson type-A region [see equations (28)-(30)] to the multiparameter case. He maximized the Gaussian (or total) curvature of the power surface at θ_0 subject to the conditions of size and unbiasedness, and called it the type-D region. Gaussian curvature of a function $z = f(x, y)$ at a point (x_0, y_0) is defined as [see Isaacson (1951 [56]), p. 219]

$$G = \frac{\left.\frac{\partial^2 z}{\partial x^2}\right|_{(x_0,y_0)} \left.\frac{\partial^2 z}{\partial y^2}\right|_{(x_0,y_0)} - \left[\left.\frac{\partial^2 z}{\partial x \partial y}\right|_{(x_0,y_0)}\right]^2}{\left[1 + \left[\left.\frac{\partial z}{\partial x}\right|_{(x_0,y_0)}\right]^2 + \left[\left.\frac{\partial z}{\partial y}\right|_{(x_0,y_0)}\right]^2\right]^2}. \tag{66}$$

Hence, for the two-parameter case, from (35) we can write the total curvature of the power hypersurface as

$$G = \frac{\beta_{11}\beta_{22} - {\beta_{12}}^2}{\left[1 + 0 + 0\right]^2} = \det\left(B_2\right), \tag{67}$$

where B_2 is defined by (37). The Type-D unbiased critical region for testing $H_0 : \theta = 0$ against the two sided alternative for a level α test is defined by the following conditions [see Isaacson (1951 [56], p. 220)]:

$$1.\ \ \beta(0,0) = \alpha. \tag{68}$$

$$2.\ \ \beta_i(0,0) = 0,\ i = 1, 2. \tag{69}$$

$$3.\ \ B_2 \text{ is positive definite.} \tag{70}$$

4. And, finally, over all such critical regions satisfying the conditions (68)-(70), $\det\left(\mathrm{B}_2\right)$ is maximized.

Note that for the type-D critical region restrictive conditions like $\beta_{12} = 0$, $\beta_{11} = \beta_{22}$ [see equations (33)-(34)] are not imposed. The type-D critical region maximizes the *total* power,

$$\beta(\theta|\omega) \simeq \alpha + \frac{1}{2}\left[\theta_1^2\beta_{11} + 2\theta_1\theta_2\beta_{12} + \theta_2^2\beta_{22}\right], \tag{71}$$

among all locally unbiased (LU) tests, whereas the type-C test maximizes power only in "limited" directions. Therefore, for finding the type-D unbiased critical region we minimize the area of the ellipse (for $k > 2$ case, it will be the volume of an ellipsoid)

$$\theta_1^2\beta_{11} + 2\theta_1\theta_2\beta_{12} + \theta_2^2\beta_{22} = \delta, \tag{72}$$

which is given by

$$\frac{\pi\delta}{\sqrt{\begin{vmatrix} \beta_{11} & \beta_{12} \\ \beta_{12} & \beta_{22} \end{vmatrix}}} = \frac{\pi\delta}{\sqrt{\det(B_2)}}. \tag{73}$$

Hence, maximizing the determinant of B_2 as in condition 4 above, is same as minimizing the volume of the ellipse shown in equation (73). Denoting ω_0 as the type-D unbiased critical region we can show that inside ω_0 the following is true [see Isaacson (1951 [56])]

$$k_{11}L_{11} + k_{12}L_{12} + k_{21}L_{21} + k_{22}L_{22} \geq k_1L + k_2L_1 + k_3L_2, \tag{74}$$

where $k_{11} = \int_{\omega_0} L_{22}(\theta)\, d\mathbf{x}$, $k_{22} = \int_{\omega_0} L_{11}(\theta)\, d\mathbf{x}$, $k_{12} = k_{21} = -\int_{\omega_0} L_{12}(\theta)\, d\mathbf{x}$, $\mathbf{x} = (x_1, x_2, ..., x_n)'$ denotes the sample, k_1, k_2 and k_3 are constants satisfying the conditions for size and unbiasedness (68) and (69).

However, one major problem with this approach, despite its geometric attractiveness, is that one has to *guess* the critical region and then verify it. As Isaacson (1951 [56], p. 223) himself noted, "...we must know our region ω_0 in advance so that we can calculate k_{11} and k_{22} and thus verify whether ω_0 has the structure required by the lemma or not." The type-E test suggested by Lehmann (1959 [71], p. 342) is same as the type-D test for testing composite hypothesis.

Given the difficulties in finding the type-D and type-E tests in actual applications, SenGupta and Vermeire ([102]) suggested a locally most mean powerful unbiased (LMMPU) test that maximizes the *mean* (instead of total) curvature of the power hypersurface at the null hypothesis among all LU level α tests. This average power criterion maximizes the *trace* of the matrix B_2 in (37) [or B_k in (44) for $k > 2$ case]. If we take an eigenvalue decomposition of the matrix B_k relating to the power function, the eigenvalues, λ_i, give the principal power curvatures while the eigenvectors corresponding to that gives the principal power directions. Isaacson (1951 [56]) used the determinant, which is the *product* of the eigenvalues, while SenGupta and Vermeire (1986

[102]) used their *sum* as a measure of curvature. Thus, LMMPU critical regions are more easily constructed using just the generalized N-P lemma. For testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, an LMMPU critical region for the $K = 2$ case is given by

$$s_{11}(\theta_0) + s_{22}(\theta_0) + s_1^2(\theta_0) + s_2^2(\theta_0) \geq k + k_1 s_1(\theta_0) + k_2 s_2(\theta_0), \tag{75}$$

where k, k_1 and k_2 are constants satisfying the size and unbiasedness conditions (68) and (69). It is easy to see that (75) is very close to Neyman's type-C region given in (43). It would be interesting to derive the LMMPU test and also the type-D and type-E regions (if possible) for testing $H_0 : \theta = 0$ in (24) and to compare that with Neyman's smooth test. We leave that topic for future research. After this long discussion of theoretical developments, we now, turn to possible applications of Neyman's smooth test.

4 Applications

We can probably credit Lawrence Klein (Klein 1991 [62], pp. 325-326) for making the first attempt to introduce Neyman's smooth test to econometrics. He gave a seminar on "Neyman's Smooth Test" at the 1942-43 MIT statistics seminar series.[8] However, Klein's effort failed, and we do not see any direct use of the smooth test in econometrics. This is particularly astonishing as testing for possible misspecification is central to econometrics. The particular property of Neyman's smooth test that makes it remarkable is the fact that it can be used very effectively both as an omnibus test for detecting departures from the null in several directions as well as a more directional test aimed at finding out the exact nature of the departure from H_0 of correct specification of the model.

Neyman (1937 [76], p. 180-185) himself illustrated a practical application of his test using Mahalanobis (1934 [73]) data on normal deviates with $n = 100$. When mentioning this application, Rayner and Best (1989 [99], pp. 46-47) stressed that Neyman also reported the *individual* components of the Ψ_k^2 statistic [see equation (45)]. This shows that Neyman (1937 [76]) believed that more specific directional tests identifying the cause and nature of deviation from H_0 can be obtained from these components.

[8] Klein joined the MIT graduate program in September 1942 after studying with Neyman's group in statistics at Berkeley, and he wanted to draw the attention of econometricians to Neyman's paper since it was published in a rather recondite journal. This may not be out of place to mention that Trygve Haavelmo was also very much influenced by Jerzy Neyman, as he mentioned in his Nobel prize lecture (see Haavelmo 1997 [47])

> "...I was lucky enough to be able to visit the United States in 1939 on a scholarship...I then had the privilege of studying with the world famous statistician Jerzy Neyman in California for a couple of months."

Haavelmo (1944 [46]) contains a seven page account of the Neyman-Pearson theory.

4.1 Orthogonal polynomials and Neyman's smooth test

Orthogonal polynomials have been widely used in estimation problems, but their use in hypothesis testing has been very limited at best. Neyman's smooth test, in that sense, pioneered the use of orthogonal polynomials for specifying the density under the alternative hypothesis. However, there are two very important concerns that need to be addressed before we can start a full-fledged application of Neyman's smooth test. First, Neyman used normalized Legendre polynomials to design the "smooth" alternatives; however, he did not justify the use of those over other orthogonal polynomials such as the truncated Hermite polynomials or the Laguerre polynomials (Barton 1953 [4], Kiefer 1985 [60]) or Charlier Type B polynomials (Lee 1986 [70]). Second, he also did not discuss how to choose the number of orthogonal polynomials to be used.[9] We start by briefly discussing a general model based on orthonormal polynomials and the associated smooth test. This would lead us to the problem of choosing the optimal value of k, and, finally, we discuss a method of choosing an alternate sequence of orthogonal polynomials.

We can design a smooth-type test in the context of regression model (Hart 1997 [52], Ch. 5)

$$Y_i = r(x_i) + \varepsilon_i, \;\; i = 1, 2, ..., n, \tag{76}$$

where Y_i is the dependent variable and $x_i's$ are fixed design points $0 < x_1 < x_2 < \cdots < x_n < 1$, and $\varepsilon_i's$ are *IID* $(0,\, \sigma^2)$. We are interested in testing the "constant regression" or "no-effect" hypothesis, i.e., $r(x) = \theta_0$, where θ_0 is an unknown constant. In analogy with Neyman's test, we consider an alternative of the form (Hart 1997 [52], p. 141)

$$r(x) = \theta_0 + \sum_{j=1}^{k} \theta_j \phi_{j,n}(x), \tag{77}$$

where $\phi_{1,n}(x), ..., \phi_{k,n}(x)$ are orthonormal over the domain of x,

$$\frac{1}{n}\sum_{q=1}^{n} \phi_{i,n}(x_q)\, \phi_{j,n}(x_q) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j, \end{cases} \tag{78}$$

and $\phi_{0,n} \equiv 1$. Hence, a test for $H_0 : \theta_1 = \theta_2 = ... = \theta_k = 0$ against $H_1 : \theta_i \neq 0$, for some $i = 1, 2, ..., k$ can be done by testing the overall significance of the model given in (76). The least

[9] Neyman (1937 [76], p. 194) did not discuss in detail the choice of the value k and simply suggested:

> "My personal feeling is that in most practical cases, there will be no need to go beyond the fourth order test. But this is only an *opinion* and not any mathematical result."

However, from their experience in using the smooth test, Thomas and Pierce (1979 [111], p. 442) thought that for the case of composite hypothesis $k = 2$ would be a better choice.

square estimators of $\theta_j's$ are given by

$$\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_{j,n}(x_i), \ j = 0, ..., k. \tag{79}$$

A test, which is asymptotically true even if the errors are not exactly Gaussian (so long as they have the same distribution and have a constant variance σ^2), is given by

$$\text{Reject } H_0 \text{ if } T_{N,k} = \frac{n\sum_{j=1}^{k}\hat{\theta}_j^2}{\hat{\sigma}^2} \geq c, \tag{80}$$

where $\hat{\sigma}$ is an estimate of σ, the standard deviation of the error terms. We can use any set of orthonormal polynomials in the above estimator including, for example, the normalized Fourier series $\phi_{j,n}(x) = \sqrt{2n}\cos(\pi j x)$ with Fourier coefficients

$$\theta_j = \int_0^1 r(x)\sqrt{2n}\cos(\pi j x)\, dx. \tag{81}$$

Observing the obvious similarity in the hypothesis tested, the test procedure in (80) can be termed as a Neyman smooth test for regression (Hart 1997 [52], p. 142).

The natural question that springs to mind at this point is what the value of k should be. Given a sequence of orthogonal polynomials, we can also test for the number of orthogonal polynomials, say k, that would give a desired level of "goodness-of-fit" for the data. Suppose now, the sample counterpart of θ_j, defined above, is given by $\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i\sqrt{2n}\cos(\pi j x_i)$. If we have an *IID* sample of size n, then, given that $E\left(\hat{\theta}_j\right) = 0$ and $V\left(\hat{\theta}_j\right) = \frac{\sigma^2}{2n}$, let us *normalize* the sample Fourier coefficients using $\hat{\sigma}$, a consistent estimator of σ. Appealing to the central limit theorem for sufficiently large n, we have the test statistic

$$S_k = \sum_{j=1}^{k}\left(\frac{\sqrt{2n}\hat{\theta}_j}{\hat{\sigma}}\right)^2 \overset{a}{\sim} \chi_k^2, \tag{82}$$

for fixed $k \leq n-1$; this is nothing but the Neyman smooth statistic in equation (45) for the Fourier series polynomials.

The optimal choice of k has been studied extensively in the literature of data-driven smooth tests first discussed by Ledwina (1994 [68]) among others. In order to reduce the subjectivity of the choice of k we can use a criterion like the Schwarz information criterion (SIC) or the Bayesian information criterion (BIC). Ledwina (1994 [68]) proposed a test that rejects the null hypothesis that k is equal to 1 for large values of $S_{\tilde{k}} = \max_{1\leq k\leq n-1}\{S_k - k\ln(n)\}$, where S_k is defined in (82). She also showed that the test statistic $S_{\tilde{k}}$ asymptotically converges to a χ_1^2 random variable

[for further insight into data-driven smooth tests see, for example, Hart (1997 [52]), pp. 185-187].

For testing uniformity, Solomon and Stephens (1985 [105]) and Stephens (1986 [31], p. 352) found that $k = 2$ is optimal in most cases where the location-scale family is used; but $k = 4$ might be a better choice when higher-order moments are required. As mentioned earlier, other papers, including Neyman (1937 [76]) and Thomas and Pierce (1979 [111]), suggested using small values of k. It has been suggested in the literature that for heavier-tailed alternative distributions, it is better to have more classes for Pearson's P_{χ^2} test in (1) or equivalently, in the case of Neyman's smooth test, more orthogonal polynomials (see, for example, Kallenberg, Oosterhoff, Schriever 1985 [59]). However, they claimed that too many class intervals, can be a potential problem for lighter-tailed distributions like normal and some other exponential family distributions (Kallenberg et al. 1985 [59], p. 959). Several studies have discussed cases where increasing the order of the test k slowly to ∞ would have better power for alternative densities having heavier tails [Kallenberg et al. (1985 [59]), Inglot, Jurlewicz, Ledwina (1990 [54]) and Eubank and LaRiccia (1992 [38])].

Some other tests like the Cramér-von Mises (CvM) and the Kolmogorov-Smirnov (KS) approaches are examples of omnibus test procedures that have power against various directions, and, hence, those tests will be consistent against many alternatives (see Eubank and LaRiccia 1992 [38]). However, to test against a specific kind of alternative, one would have to use a more directional alternative geared towards detecting specific types of departures from the null. Neyman's smooth test serves as a compromise between the two criteria. The smooth test statistic gives equal weight to all k components in equation (45) unlike the KS and the CvM type statistics, which severely down-weight the terms with the higher-order moments (see Eubank and LaRiccia 1992 [38], p. 2072). The procedure for selecting the truncation point k in Neyman (1937 [76]) smooth test is similar to the choice of the number of classes in the Pearson χ^2 test and has been discussed in Kallenberg et al. (1985 [59]) and Fan (1996 [39]).

Let us now revisit the problem of a choosing an optimal sequence of orthogonal polynomials around the density $f(x;\gamma)$ under H_0. The following discussion closely follows Smith (1989 [103]) and Cameron and Trivedi (1990 [20]). They used the score test after setting up the alternative in terms of orthogonal polynomials with the baseline density $f(x;\gamma)$ under the null hypothesis. Expanding the density $g(x;\gamma,\theta)$ using an orthogonal polynomial sequence with respect to $f(x;\gamma)$, we have

$$g(x;\gamma,\theta) = f(x;\gamma) \sum_{j=0}^{\infty} a_j(\gamma,\theta)\, p_j(x;\gamma), \tag{83}$$

where

$$a_0(\gamma,\theta) \equiv 1,\ p_0(x;\gamma) \equiv 1,\ p_j(x;\gamma) = \sum_{i=0}^{j} \alpha_{ij} x^i. \tag{84}$$

The polynomials $p_j's$ are orthonormal with respect to the density $f(x;\gamma)$.

We can construct orthogonal polynomials through the moments. Suppose we have a sequence of moments $\{\mu_n\}$ of the random variable X, then a necessary and sufficient condition for the existence of a unique orthogonal polynomial sequence is that $\det(M_n) > 0$, where $M_n = [M_{ij}] = [\mu_{i+j-2}]$, for $n = 0, 1, \ldots$. We can write $\det(M_n)$ as

$$|M_n| = \begin{vmatrix} M_{n-1} & \mathbf{m} \\ \mathbf{m}' & \mu_{2n} \end{vmatrix} = \mu_{2n} |M_{n-1}| - \mathbf{m}' Adj(M_{n-1}) \mathbf{m}, \tag{85}$$

where $\mathbf{m}' = (\mu_n, \mu_{n+1}, ..., \mu_{2n-1})$, $|M_{-1}| = |M_0| = 1$ and "*Adj*" means the adjugate of a matrix. The n^{th} order orthogonal polynomial can be constructed from

$$\begin{aligned} P_n(x) &= [|M_{n-1}|]^{-1} |D_n(x)|, \\ \text{where } |D_n(x)| &= \begin{bmatrix} M_{n-1} & \mathbf{m} \\ \mathbf{x}'_{(-n)} & x^n \end{bmatrix} \text{ and } \mathbf{x}'_{(-n)} = \left(1\ x\ x^2 ... x^{n-1}\right). \end{aligned} \tag{86}$$

This gives us a whole system of orthogonal polynomials $P_n(x)$ [see Cramér (1946 [28]), pp. 131-132, Cameron and Trivedi (1990 [20]), pp. 4-5 and Appendix A].

Smith (1989 [103]) performed a test of $H_0 : g(x;\gamma,\theta) = f(x;\gamma)$ (i.e., $a_j(\gamma,\theta) = 0$, $j = 1, 2, ..., k$ or $\mathbf{a}_k = \{a_j(\gamma,\theta)\}_{j=1}^k = \mathbf{0}$) using a truncated version of the expression for the alternative density,

$$g(x;\gamma,\theta) = f(x;\gamma) \left\{ 1 + \sum_{j=1}^{k} a_j(\gamma,\theta) \sum_{i=1}^{j} \alpha_{ij}(\gamma) \left[x^i - \mu_{fi}(\gamma) \right] \right\}, \tag{87}$$

where

$$\mu_{fi}(\gamma) = \int x^i f(x;\gamma)\, dx = E_f\left[x^i\right]. \tag{88}$$

However, the expression $g(x;\gamma,\theta)$ in (87) may not be a proper density function. Because of the truncation, it may not be non-negative for all values of x nor will it integrate to unity. Smith referred to $g(x;\gamma,\theta)$ as a *pseudo-density* function.

If we consider y to be the probability integral transform of the original data in x, then defining $E_h(y^i|H_0) = \mu_{h_0 i}$, we can rewrite the above density in (87), in the absence of any nuisance parameter γ, as

$$h(y;\theta) = 1 + \sum_{j=1}^{k} \theta_j \sum_{i=1}^{j} \alpha_{ij} \left[y^i - \mu_{h_0 i} \right]. \tag{89}$$

From this we can get the Neyman smooth test as proposed by Thomas and Pierce (1979 [111]). Here we test $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ where $\theta = (\theta_1, \theta_2, ..., \theta_k)'$. From the equation (89), we can get the score function as $\frac{\partial \ln h(y;\theta)}{\partial \mathbf{a}} = \mathbf{A}_k \mathbf{m}_k$ where $\mathbf{A}_k = [\alpha_{ij}(\theta)]$ is a $k \times k$ lower triangular matrix (with non-zero diagonal elements) and $\mathbf{m}_k$ is a vector of deviations whose i^{th} component is $(y^i - \mu_{h_0 i})$. The score test statistic will have the form

$$S_n = n \mathbf{m}'_{kn} \left[\mathbf{V}_n \right]^{-} \mathbf{m}_{kn}, \tag{90}$$

where $\mathbf{m}_{kn}$ is the vector of the sample mean of deviations, $\mathbf{V}_n = \mathcal{I}_{mm} - \mathcal{I}_{m\theta} \mathcal{I}_{\theta\theta}^{-1} \mathcal{I}_{\theta m}$, with $\mathcal{I}_{mm} = E_\theta \left[\mathbf{m}_k \mathbf{m}'_k \right], \mathcal{I}_{m\theta} = E_\theta \left[\mathbf{m}_k \mathbf{s}'_\theta \right], \mathcal{I}_{m\theta} = E_\theta \left[\mathbf{s}_\theta \mathbf{s}'_\theta \right]$ and $\mathbf{s}_\theta = E_\theta \left[\frac{\partial \ln h(y;\theta)}{\partial \theta} \right]$, is the conditional variance-covariance matrix and $\left[\mathbf{V}_n \right]^{-}$ is its g-inverse (see Smith 1989 [103], pp. 184-185 for details). Here $E_\theta [.]$ denotes expectation taken with respect to the true distribution of y, but eventually evaluated under $H_0 : \theta = 0$. Test statistic (90) can also be computed using an artificial regression of the vector of 1's on the vector of score functions of the nuisance parameters and the deviations from the moments. It can be shown that S_n follows an asymptotic χ^2 distribution with degrees of freedom $= rank(\mathbf{V}_n)$. Possible uses could be in limited dependent variable models like the binary response model and models for duration such as unemployment spells (Smith 1989 [103]). Cameron and Trivedi (1990 [20]) derived an analogous test using moment conditions of the exponential family. For testing exponentiality in the context of duration models, Lee (1984 [69]) transformed the "exponentially distributed" random variable X by $z = \Phi^{-1}[F(x)]$, where F is the exponential distribution function and Φ^{-1} is the inverse normal probability integral transform. Lee then proposed testing normality of z using the score test under a Pearson family of distributions as the alternative density for z. If we restrict to first four moments in Smith (1989 [103]), then the approaches of Lee and Smith are identical.

4.2 Density forecast evaluation and calibration

The importance of density forecast evaluation in economics has been aptly depicted by Crnkovic and Drachman (1997 [30], p. 47) as follows: "At the heart of market risk measurement is the forecast of the probability density functions (PDFs) of the relevant market variables ... a forecast of a PDF is the central input into any decision model for asset allocation and/or hedging ... therefore, the quality of risk management will be considered synonymous with the quality of PDF forecasts." Suppose that we have time series data (say, the daily returns to the S. & P. 500 Index) given by $\{x_t\}_{t=1}^{m}$. One of the most important questions that we would like to answer

is, what is the sequence of the true density functions $\{g_t(x_t)\}_{t=1}^m$ that generated this particular realization of the data? Since this is time series data, at time t we know all the past values of x_t up to time t or the *information set* at time t, namely, $\Omega_t = \{x_{t-1}, x_{t-2}, ...\}$. Let us denote the one-step-ahead forecast of the sequence of densities as $\{f_t(x_t)\}$ conditional on Ω_t. Our objective is to determine how much the forecast density $\{f_t\}$ depicts the true density $\{g_t\}$. The main problem in performing such a test is that both the actual density $g_t(.)$ and the one-step-ahead predicted density $f_t(.)$ could depend on the time t and, thus, on the information set Ω_t. This problem is unique, since, on one hand, it is a classical goodness-of-fit problem but, on the other, it is also a combination of several different, possibly dependent, goodness-of-fit tests. One approach to handling this particular problem would be to reduce it to a more tractable one in which we have the same, or similar, hypotheses to test, rather than a host of different hypotheses. Following Neyman (1937 [76]) this is achieved using the probability integral transform

$$y_t = \int_{-\infty}^{x_t} f_t(u)\, du. \tag{91}$$

Using equations (3), (6) and (17), the density function of the transformed variable y_t is given by

$$h_t(y_t) = 1,\ 0 < y_t < 1, \tag{92}$$

under the null hypothesis that our forecasted density is the true density for all t, i.e., $H_0 : g_t(.) = f_t(.)$.

If we are only interested in performing a goodness-of-fit test that the variable y_t follows a uniform distribution, we can use a parametric test like Pearson's χ^2 on grouped data or non-parametric tests like the KS or the CvM or a test using the Kuiper statistics (see Crnkovic and Drachman 1997 [30], p. 48). Any of those suggested tests would work as a good *omnibus* test of goodness-of-fit. If we fail to reject the null hypothesis we can conclude that there is not enough evidence that the data is *not* generated from the forecasted density $f_t(.)$; however, a rejection would not throw any light on the possible form of the true density function.

Diebold, Gunther and Tay (1998 [33]) used Theorem 1, discussed in Subsection 2.1, and tested $H_0 : g_t(.) = f_t(.)$ by checking whether the probability integral transform y_t in (91) follows *IID* $U(0,1)$. They employed a graphical (visual) approach to decide on the structure of the alternative density function by a two-step procedure. First, they visually inspected the histogram of y_t to see if it comes from $U(0,1)$ distribution. Then, they looked at the individual correlograms of each of the first four powers of the variable $z_t = y_t - 0.5$ in order to check for any residual effects of bias, variance or higher-order moments. In the absence of a more analytical test of goodness-of-fit, this graphical method has also been used in Diebold, Tay and Wallis (1999 [36]) and Diebold, Hahn and Tay (1999 [34]). For reviews on density forecasting

and forecast evaluation methods, see Tay and Wallis (2000 [110]) and Diebold and Lopez (1996 [35]). The procedure suggested is very attractive due to its simplicity of execution and intuitive justification; however, the resulting size and power of the procedure is unknown. Also, we are not sure about the optimality of such a diagnostic method. Berkowitz (2000 [17], p. 4) commented on the Diebold et al.(1998 [33]) procedure: "Because their interest centers on developing tools for diagnosing *how* models fail, they do not pursue formal testing." Neyman's smooth test (1937 [76]) provides an *analytic* tool to determine the structure of the density under the alternative hypothesis using orthonormal polynomials (normalized Legendre polynomials) $\pi_j(y)$ defined in (20).[10] While, on one hand, the smooth test provides a basis for a classical goodness-of-fit test (based on the generalized N-P lemma), on the other hand, it can also be used to determine the sensitivity of the power of the test to departures from the null hypothesis in different directions, for example, deviations in variance, skewness and kurtosis (see Bera and Ghosh 2001 [14]). We can see that the Ψ_k^2 statistic for Neyman's smooth test defined in equation (22) is comprised of k components of the form $\frac{1}{n}\left(\sum_{i=1}^{n}\pi_j(y_i)\right)^2, j=1,..k$, which are nothing but the squares of the efficient score functions. Using Rao and Poti (1946 [98]), Rao(1948 [93]) and Neyman (1959 [77]) one can risk the "educated speculation" that an *optimal test* should be based on the *score function* [for more on this, see Bera and Billias (2001a [11], 2001b [12])]. From that point of view we achieve *optimality* using the smooth test.

Neyman's smooth-type test can also be used in other areas of macroeconomics such as evaluating the density forecasts of realized inflation rates. Diebold, Tay and Wallis (1999 [36]) used a graphical technique as did Diebold et al. (1998 [33]) on the density forecasts of inflation from the *Survey of Professional Forecasters*. Neyman's smooth test in its original form was intended mainly to provide an *asymptotic test* of significance for testing goodness-of-fit for "smooth" alternatives. So, one can argue that although we have large enough data in the daily returns of the S. & P. 500 Index, we would be hard pressed to find similar size data for macroeconomic series such as GNP, inflation. This might make the test susceptible to significant small-sample fluctuations, and the results of the test might not be strictly valid. In order to correct for size or

[10] Neyman (1937 [76]) used $\pi_j(y)'s$ as the orthogonal polynomials which can be obtained by using the following conditions,

$$\pi_j(y) = a_{j0} + a_{j1}y + ... + a_{jj}y^j, a_{jj} \neq 0,$$

given the restrictions of orthogonality given in Subsection 2.2. Solving these the first five $\pi_j(y)$ are (Neyman 1937 [76], pp. 163-164)

$\pi_0(y) = 1,$
$\pi_1(y) = \sqrt{12}\left(y - \frac{1}{2}\right),$
$\pi_2(y) = \sqrt{5}\left(6\left(y-\frac{1}{2}\right)^2 - \frac{1}{2}\right),$
$\pi_3(y) = \sqrt{7}\left(20\left(y-\frac{1}{2}\right)^3 - 3\left(y-\frac{1}{2}\right)\right),$
$\pi_4(y) = 210\left(y-\frac{1}{2}\right)^4 - 45\left(y-\frac{1}{2}\right)^2 + \frac{9}{8}$

power problems due to small sample size, we can either do a size correction [similar to other score tests, see Harris (1985 [50]), Harris (1987 [51]), Cordeiro and Ferrari (1991 [25]), Cribari-Neto and Ferrari (1995 [29]) and Bera and Ullah (1991 [16]) for applications in econometrics] or use a modified version of the "smooth test" based on Pearson's P_λ test discussed in Subsection 2.1. This promises to be an interesting direction for future research.

We can easily extend Neyman's smooth test to a multivariate setup of dimension N for m time periods, by taking a combination of Nm sequences of univariate densities as discussed by Diebold, Hahn and Tay (1999 [34]). This could be particularly useful in fields like financial risk management to evaluate densities for high-frequency financial data like stock or derivative (options) prices and foreign exchange rates. For example, if we have a sequence of the joint density forecasts of more than one, say 3, daily foreign exchange rates over a period of $1,000$ days, we can evaluate its accuracy using the smooth test for $3,000$ univariate densities. One thing that must be mentioned here, there could be both temporal and contemporaneous dependencies in these observations, we are assuming that taking conditional distribution both with respect to time and across-variables is feasible (see, for example, Diebold, Hahn and Tay 1999 [34], p. 662).

Another important area of the literature on the evaluation of density forecasts is the concept of *calibration*. Let us consider this in the light of our formulation of Neyman's smooth test in the area of density forecasts. Suppose that the actual density of the process generating our data, $g_t(x_t)$, is different from the forecasted density, $f_t(x_t)$, say,

$$g_t(x_t) = f_t(x_t)\, r_t(y_t), \tag{93}$$

where $r_t(y_t)$ is a function depending on the probability integral transforms and can be used to calibrate the forecasted densities, $f_t(x_t)$, recursively. This procedure of calibration might be needed if the forecasts are off in a consistent way, that is to say, the probability integral transforms $\{y_t\}_{t=1}^m$ are not $U(0,1)$ but are independent and identically distributed with some other distribution (see, for example, Diebold, Hahn and Tay 1999 [34]).

If we compare equation (93) with the formulation of the smooth test given by equation (24), where $f_t(x)$, the density under H_0, is embedded in $g_t(x)$ (in the absence of the nuisance parameter γ), the density under H_1, we can see that

$$\begin{aligned} r_t(y_{t+1}) &= c(\theta) \exp\left[\sum_{j=1}^{k} \theta_j \pi_j(y_{t+1})\right] \\ \Leftrightarrow \ln r_t(y_{t+1}) &= \ln c(\theta) + \sum_{j=1}^{k} \theta_j \pi_j(y_{t+1}). \end{aligned} \tag{94}$$

Hence, we can actually estimate the calibrating function from (94). It might be worthwhile to

compare the method of calibration suggested by Diebold, Hahn and Tay (1999 [34]) using non-parametric (kernel) density estimation with the one suggested here coming from a parametric setup [also see Thomas and Pierce (1979 [111]) and Rayner and Best (1989 [99], p. 77) for a formulation of the alternative hypothesis].

So far, we have discussed only one aspect of the use of Neyman's smooth test, namely, how it can be used for evaluating (and calibrating) density forecast estimation in financial risk management and macroeconomic time-series data such as inflation. Let us now discuss another example that recently has received substantial attention, namely the Value-at-Risk (VaR) model in finance. VaR is generally defined as an extreme quantile of the value distribution of a financial portfolio. It measures the maximum allowable value the portfolio can lose over a period of time at, say, the 95% level. This is a widely used measure of portfolio risk or exposure to risk for corporate portfolios or asset holdings [for further discussion see Smithson and Minton (1997 [104])]. A common method of calculating VaR is to find the proportion of times the upper limit of interval forecasts have been exceeded. Although this method is very simple to compute, it requires a large sample size (see Kupiec 1995 [65], p. 83). For smaller sample size, which is common in risk models, it is often advisable to look at the entire probability density function or a map of quantiles. Hypothesis tests on the goodness-of-fit of VaRs could be based on the tail probabilities or tail expected loss of risk models in terms of measures of "exceedence" or the number of times that the total loss has exceeded the forecasted VaR. The tail probabilities are often of more concern than the interiors of the density of the distribution of asset returns.

Berkowitz (2000 [17]) argued that in some applications highly specific testing guidelines are necessary, and, in order to give a more formal test for the graphical procedure suggested by Diebold et al. (1998 [33]), he proposed a formal likelihood ratio test on the VaR model. An advantage of his proposed test is that it gives some indication of the nature of the violation when the goodness-of-fit test is rejected. Berkowitz followed Lee's (1984 [69]) approach but used the likelihood ratio test (instead of the score test) based on the inverse standard normal transformation of the probability integral transforms of the data. The main driving forces behind the proposed test are its tractability and the properties of the normal distribution. Let us define the inverse standard normal transform $z_t = \Phi^{-1}\left(\hat{F}\left(y_t\right)\right)$ and consider the following model

$$z_t - \mu = \rho\left(z_{t-1} - \mu\right) + \varepsilon_t. \tag{95}$$

To test for independence, we can test $H_0 : \rho = 0$ in the presence of nuisance parameters μ and σ^2 (the constant variance of the error term ε_t). We can also perform a joint test for the parameters $\mu = 0, \rho = 0$ and $\sigma^2 = 1$ using the likelihood ratio test statistic

$$LR = -2\left(l\left(0, 1, 0\right) - l\left(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}\right)\right), \tag{96}$$

that is distributed as a χ^2 with three degrees of freedom, where $l(\theta) = \ln L(\theta)$ is the log-likelihood function. The above test can be considered a test based on the tail probabilities. Berkowitz (2000 [17]) reported Monte Carlo simulations for the Black-Scholes model and demonstrated superiority of his test with respect to the KS, CvM and a test based on the Kuiper statistic. It is evident that there is substantial similarity between the test suggested by Berkowitz and the smooth test; the former explicitly puts in the conditions of higher-order moments through the inverse standard Gaussian transform, while the latter looks at a more general exponential family density of the form given by equation (19). Berkowitz exploits the properties of the normal distribution to get a likelihood ratio test, while Neyman's smooth test is a special case of Rao's score test, and, therefore, asymptotically they should give similar results.

To further elaborate, let us point out that finding the distributions of VaR is equivalent to finding the distribution of quantiles of the asset returns. LaRiccia (1991 [67]) proposed a quantile function-based analog of Neyman's smooth test. Suppose, we have a sample $(y_1, y_2, ..., y_n)$ from a fully specified cumulative distribution function (cdf) of a location-scale family $G(.; \mu, \sigma)$ and define the order statistics as $\{y_{1n}, y_{2n}, ..., y_{nn}\}$. We want to test the null hypothesis that $G(.; \mu, \sigma) \equiv F(.)$ is the true data-generating process. Hence, under the null hypothesis, for large sample size n, the expected value of the i^{th} order statistic, Y_{in}, is given by $E(Y_{in}) = \mu + \sigma Q_0\left[\frac{i}{(n+1)}\right]$, where $Q_0(u) = \inf\{y : F(y) \geq u\}$ for $0 < u < 1$. The covariance matrix under the null hypothesis is approximated by

$$\begin{aligned}\sigma_{ij} = Cov(Y_{in}, Y_{jn}) \approx \sigma^{-2}&\left[fQ_0\left(\frac{i}{n+1}\right) fQ_0\left(\frac{j}{n+1}\right)\right] \\ &\times\left[\min\left(\frac{i}{n+1}, \frac{j}{n+1}\right) - \frac{ij}{(n+1)^2}\right],\end{aligned} \tag{97}$$

where $fQ_0(.) \equiv f(Q_0(.))$ is the density of the quantile function under H_0. LaRiccia took the alternative model as

$$E(Y_{in}) \simeq \mu + \sigma Q_0\left[\frac{i}{(n+1)}\right] + \sum_{j=1}^{k} \delta_j p_j\left[\frac{i}{(n+1)}\right], \tag{98}$$

with $Cov(Y_{in}, Y_{jn})$ as given in (97) and $p_1(.), p_2(.), \ldots, p_k(.)$ are functions for some fixed value of k. LaRiccia (1991 [67]) proposed a likelihood ratio test for $H_0 : \delta = (\delta_1, \delta_2, ..., \delta_k)' = 0$, which turns out to be analogous to the Neyman smooth test.

4.3 Smooth tests in survival analysis with censoring and truncation

One of the important questions econometricians often face is whether there are one or more unobserved variables that have a significant influence on the outcome of a trial or experiment. Social scientists like economists have to rely mainly on observational data. Although, in some other disciplines, it is possible to control for unobserved variables to a great extent through experimental design, econometricians are not that fortunate most of the time. This gives rise to misspecification in the model through unobserved heterogeneity (for example, ability, expertise, genetical traits, inherent resistance to diseases), which, in turn, could significantly influence outcomes like income or survival times. In this Subsection we look at the effect of misspecification on distribution of survival times through a random multiplicative heterogeneity in the *hazard* function (Lancaster 1985 [66]) utilizing Neyman's smooth test with generalized residuals.

Suppose now that we observe survival times $t_1, t_2, ..., t_n$, which are independently distributed (for the moment, without any censoring) with a density function $g(t;\gamma,\theta)$ and cdf $G(t;\gamma,\theta)$, where γ are parameters. Let us define the hazard function $\lambda(t;\gamma,\theta)$ by

$$P(t < T < t+dt | T > t) = \lambda(t;\gamma,\theta)\, dt,\ t > 0, \tag{99}$$

which is the conditional probability of death or failure over the next infinitesimal period dt given that the subject has survived till time t. There could be several different specifications of the hazard function $\lambda(t;\gamma,\theta)$ such as the proportional hazards models. If the survival time distribution is Weibull then the hazard function is given by

$$\lambda(t;\alpha,\beta) = \alpha t^{\alpha-1} \exp(x'\beta). \tag{100}$$

It can be shown (for example, see Cox and Oakes 1984 [27], p. 14) that if we define the survival function as $\bar{G}(t;\gamma,\theta) = 1 - G(t;\gamma,\theta)$, then we would have

$$\lambda(t;\gamma,\theta) = \frac{g(t;\gamma,\theta)}{\bar{G}(t;\gamma,\theta)} \Rightarrow g(t;\gamma,\theta) = \lambda(t;\gamma,\theta)\bar{G}(t;\gamma,\theta). \tag{101}$$

We can also obtain the survival function as

$$\bar{G}(t;\gamma,\theta) = \exp\left(-\int_0^t \lambda(s;\gamma,\theta)\, ds\right) = \exp(-H(t;\gamma,\theta)). \tag{102}$$

$H(t;\gamma,\theta)$ is known as the integrated hazard function. Suppose we have the function, $t_i = T_i(\delta,\varepsilon_i)$, where $\delta = (\gamma',\theta')'$, and also let R_i be uniquely defined so that, $\varepsilon_i = R_i(\delta,t_i)$. Then, the functional ε_i is called a generalized error, and we can estimate it by $\hat{\varepsilon}_i = R_i\left(\hat{\delta},t_i\right)$. For example, a generalized residual could be the integrated hazard function such as $\hat{\varepsilon}_i =$

$H\left(t_i;\hat{\gamma},\hat{\theta}\right) = \int_0^{t_i} \lambda\left(s;\hat{\gamma},\hat{\theta}\right) ds$ (Lancaster 1985 [66]), or it could be the distribution function such as $\hat{\varepsilon}_i = G\left(t_i;\hat{\gamma},\hat{\theta}\right) = \int_0^{t_i} g\left(s;\hat{\gamma},\hat{\theta}\right) ds$ (Gray and Pierce 1985 [45]).

Let us consider a model with hazard function given by $\lambda_z(t) = z\lambda(t)$, where $z = e^u$ is the multiplicative heterogeneity and $\lambda(t)$ is the hazard function with no multiplicative heterogeneity (ignoring the dependence on parameters and covariates, for the sake of simplicity). Hence the survivor function given z is

$$\bar{G}_z(t|z) = \exp(-z\varepsilon). \tag{103}$$

Let us further define σ_z^2 as the variance of z, $F(t) = E[\exp(-\varepsilon)]$ is the survival function and ε is the integrated hazard function evaluated at t, under the hypothesis of no unobserved heterogeneity. Then, using the integrated hazard function as the generalized residual, the survival function is given by (see Lancaster 1985 [66], pp. 164-166)

$$\bar{G}_z(t) \simeq \bar{F}(t)\left\{1 + \frac{\sigma_z^2}{2}\varepsilon^2\right\}. \tag{104}$$

Differentiating with respect to t and after some algebraic manipulation of (104), we get for small enough values of σ_z^2

$$g_z(t) \simeq f(t)\left\{1 + \frac{\sigma_z^2}{2}\left(\varepsilon^2 - 2\varepsilon\right)\right\}, \tag{105}$$

where g_z is the density function with multiplicative heterogeneity z, f is the density with $z = 1$. We can immediately see that if we used normalized Legendre polynomials to expand g_z, we would get a setup very similar to that of Neyman's (1937 [76]) smooth test with nuisance parameters γ (see also Thomas and Pierce, 1979 [111]). Further, the score test for the existence of heterogeneity ($H_0 : \theta = 0$ i.e., $H_0 : \sigma_z^2 = 0$) is based on the sample counterpart of the score function, $\frac{1}{2}(\varepsilon^2 - 2\varepsilon)$ for $z = 1$. If s^2 is the estimated variance of the generalized residuals $\hat{\varepsilon}$, then the score test, which is also White's (1982 [114]) information matrix (IM) test of specification, is based on the the expression, $s^2 - 1$, divided by its estimated standard error (Lancaster 1985 [66]). This is a particular case of the result that the IM test is a score test for neglected heterogeneity when the variance of the heterogeneity is small, as pointed out in Cox (1983 [26]) and Chesher (1984 [22]).

Although the procedure outlined by Lancaster (1985 [66]) shows a lot of promise for applying Neyman's smooth test to survival analysis, there are two major drawbacks. First, it is difficult, if not impossible, to obtain real life survival data without the problem of censoring or truncation; second, Lancaster (1985 [66]) worked within the framework of the Weibull model, and the impact of model misspecification needs to be considered. Gray and Pierce (1985 [45]) focused on the second issue of misspecification in the model for survival times and also tried to answer the first question of censoring in some special cases.

Suppose the observed data is of the form

$$\begin{cases} Y_i = & \min\{T_i, V_i\} \\ Z_i = & I\{T_i \geq V_i\}, \end{cases} \tag{106}$$

where $I\{A\}$ is an indicator function for event A and V_i are random censoring times generated independently of the data from cdfs $C_i, i = 1, 2, ..., n$. Gray and Pierce (1985 [45]) wanted to test the validity of the function $\bar{G}$ rather than the effect of the covariates x_i on T_i. We can look at any survival analysis problem (with or without censoring or truncation) in two parts. First, we want to verify the functional form of the cdf G_i i.e., to answer the question whether the survival times are generated from a particular distribution like $G_i(t;\beta) = 1 - \exp(-\exp(x_i'\beta)\, t)$; second, we want to test the effect of the covariates x_i on the survival time T_i. The second problem has been discussed quite extensively in the literature. However, there has been relatively less attention given to the first problem. This is probably because there could be an infinite number of choices of the functional form of the survival function. Techniques based on Neyman's smooth test provide an opportunity to address the problem of misspecification in a more concrete way.[11]

The main problem discussed by Gray and Pierce (1985 [45]) is to test H_0 which states that the generalized error $U_i = G_i(T_i;\gamma,\theta=0) = F_i(T_i;\gamma)$ is *IID* $U(0,1)$ against the alternative H_1, which is characterized by the pdf

$$g_i(t;\gamma,\theta) = f_i(t;\gamma)\exp\left\{\sum_{l=1}^{k}\theta_l\psi_l(F_i(t;\gamma))\right\}\exp\{-K(\theta,\gamma)\}, \tag{107}$$

where $f_i(t;\gamma)$ is the pdf under H_0. Thomas and Pierce (1979 [111]) chose $\psi_l(u) = u^l$, but one could use any system of orthonormal polynomials such as the normalized Legendre polynomials. In order to perform a score test as discussed in Thomas and Pierce (1979 [111]), which is an extension of Neyman's smooth test in presence of nuisance parameters, one must determine the asymptotic distribution of the score statistic. In the case of censored data, the information matrix under the null hypothesis will depend on the covariates, the estimated nuisance parameters and also on the generally unknown censoring distribution, even in the simplest location-scale setup. In order to solve this problem, Gray and Pierce (1979 [45]) used the distribution conditional

[11] We should mention here that a complete separation of the misspecification problem and the problem of the effect of covariates is not always possible to a satisfactory level. In their introduction, Gray and Pierce (1985 [45]), pointed out:

> "Although, it is difficult in practice to separate the issues, our interest is in testing the adequacy of the form of F, rather than in aspects related to the adequacy of the covariables."

This sentiment has also been reflected in Peña (1998 [89]) as he demonstrated that the issue of the effect of covariates is "... highly intertwined with the goodness-of-fit problem concerning $\lambda(.)$."

on observed values in the same spirit as the EM algorithm (Dempster, Laird and Rubin 1977 [32]). When there is censoring, the true cdf or the survival function can be estimated using a method like the Kaplan-Meier or the Nelson-Aalen estimators (Hollander and Peña 1992 [53], p. 99). Gray and Pierce (1985 [45]) reported limited simulation results where they looked at data generated by exponential distribution with Weibull waiting time. They obtained encouraging results using Neyman's smooth test over the standard likelihood ratio test.

In the survival analysis problem, a natural function to use is the hazard function rather than the density function. Peña (1998 [89]) proposed the smooth goodness-of-fit test obtained by embedding the baseline hazard function $\lambda(.)$ in a larger family of hazard functions developed through smooth, and possibly random, transformations of $\lambda_0(.)$ using the Cox proportional hazard model $\lambda(t|X(t)) = \lambda(t)\exp(\beta' X(t))$ where $X(t)$ is a vector of covariates. Peña used an approach based on generalized residuals within a counting process framework as described in Anderson, Borgan, Gill and Keiding (1982 [1], 1991 [2]) and reviewed in Hollander and Peña (1992 [53]).

Suppose now, we consider the same data as given in (106), (Y_i, Z_i). In order to facilitate our discussion on analyzing for censored data for survival analysis, we define:

1. The number of actual failure times observed without censoring before time t : $N(t) = \sum_{i=1}^{n} I(Y_i \leq t, Z_i = 1)$.
2. The number of individuals who are still surviving at time t : $R(t) = \sum_{i=1}^{n} I(Y_i \geq t)$.
3. The indicator function for *any* survivors at time t : $J(t) = I(R(t) > 0)$.
4. The conditional mean number of survivors at risk at any time $s \in (0,t)$, given that they survived till time s : $A(t) = \int_0^t R(s)\lambda(s)\,ds$.
5. The difference between the observed and the expected (conditional) numbers of failure times at time t : $M(t) = N(t) - A(t)$.[12]

Let $F = \{\mathcal{F}_t : t \in T\}$ be the history or the information set (filtration) or the predictable process at time t. Then, for the Cox proportional hazards model the long-run smooth "averages" of N are given by $A = \{A(t) : t \in T\}$, where

$$A(t) = \int_0^t R(s)\lambda(s)\exp\{\beta' X(s)\}\,ds,\ i = 1, ..., n \tag{108}$$

[12]In some sense, we can interpret $M(t)$ to be the residual or error in the number of deaths or failures over the smooth conditional average of the number of individuals who would die given that they survived till time $s \in (0,t)$. Hence, $M(t)$ would typically be a martingale difference process. The series $A(t)$, also known as the compensator process, is absolutely continuous with respect to the Lebesgue measure and is predetermined at time t, since it is the definite integral upto time t of the predetermined *intensity* process given by $R(s)\lambda(s)$ (for details see Hollander and Peña 1992 [53], pp. 101-102).

and β is a $q \times 1$ vector of regression coefficients and $X(s)$ is a $q \times 1$ vector of predictable (or predetermined) covariate processes.

The test developed by Peña (1998 [89]) is for $H_0 : \lambda(t) = \lambda_0(t)$, where $\lambda_0(t)$ is a completely specified baseline hazard rate function associated with the integrated hazard given by $H_0(t) = \int_0^t \lambda_0(s)\, ds$, which is assumed to be strictly increasing. Following Neyman (1937 [76]) and Thomas and Pierce (1979 [111]), the smooth class of alternatives for the hazard function is given by

$$\lambda(t;\theta,\beta) = \lambda_0(t) \exp\{\theta' \psi(t;\beta)\}, \tag{109}$$

where $\theta \in \mathbb{R}^k$,$k = 1, 2...$, and $\psi(t;\beta)$ is a vector of locally bounded predictable (predetermined) processes that are twice continuously differentiable with respect to β. So, as in the case of the traditional smooth test, we can rewrite the null as, $H_0 : \theta = 0$. This gives the score statistic process under H_0 as

$$\begin{aligned} U_\theta^F(t;\theta,\beta)\big|_{\theta=0} &= \int_0^t \left[\frac{\partial}{\partial\theta} \log\lambda(s;\theta,\beta)\right] dM(s;\theta,\beta)\Big|_{\theta=0} \\ &= \int_0^t \psi(s;\beta)\, dM(s;0,\beta), \end{aligned} \tag{110}$$

where $M(t;\theta,\beta) = N(t) - A(t;\theta,\beta)$, $i = 1,\ldots,n$. To obtain a workable score test statistic one has to replace the nuisance parameter β by its MLE under H_0. The efficient score function $\frac{1}{\sqrt{n}} U_\theta^F\left(t;0,\hat{\beta}\right)$ process has an asymptotic normal distribution with 0 mean [see Peña (1998 [89]), p. 676 for the variance-covariance matrix $\Gamma(\cdot,\cdot;\beta)$].

The proposed smooth test statistic is given by

$$s\left(\tau;\hat{\beta}\right) = \frac{1}{n} U_\theta^F\left(\tau;0,\hat{\beta}\right)' \Gamma\left(\tau,\tau;\hat{\beta}\right)^{-} U_\theta^F\left(\tau;0,\hat{\beta}\right), \tag{111}$$

which has an asymptotic $\chi^2_{k^*}$ distribution, $\widehat{k}^* = rank\left[\Gamma\left(\tau,\tau;\hat{\beta}\right)\right]$, where $\Gamma\left(\tau,\tau;\hat{\beta}\right)$ is the asymptotic variance of the score function.[13]

Peña (1998 [89]) also proposed a procedure to combine the different choices of ψ to get an omnibus smooth test that will have power against several possible alternatives. Consistent with the original idea of Neyman (1937 [76]) and as later proposed by Gray and Pierce (1985 [45]) and Thomas and Pierce (1979 [111]), Peña considered the polynomial $\psi(t;\beta) = \left(1,\ H_0(t), ...,\ H_0(t)^{k-1}\right)'$, where, $H_0(t)$ is the integrated hazard function under the null [for details of the test see Peña (1998 [89])]. Finally, Peña (1998a [90]) using a similar counting-process

[13] Peña (1998 [89], p. 676) claimed that we cannot get the same asymptotic results in terms of the nominal size of the test if we replace β by any other consistent estimator under H_0. The test statistic might not even be asymptotically χ^2.

approach suggested a smooth goodness-of-fit test for the composite hypothesis (see Thomas and Pierce 1979 [111], Rayner and Best 1989 [99] and Section 3).

4.4 Posterior predictive p-values and related tests in Bayesian statistics and econometrics

In several areas of research p-value might well be the single most reported statistic. However, it has been widely criticized because of its indiscriminate use and relatively unsatisfactory interpretation in the empirical literature. Fisher (1945 [42], pp. 130-131), while criticizing the axiomatic approach to the test, pointed out that setting up fixed probabilities of Type I error *a priori* could yield misleading conclusions about the data or the problem at hand. Recently, this issue gained attention in some fields of medical research. Donahue (1999 [37]) discussed the information content in the p-value of a test. If we consider $F(t|H_0) = F(t)$ to be the cdf of a test statistic T under H_0 and $F(t|H_1) = G(t)$ be the cdf of T under the alternative, the p-value defined as $P(t) = P\{T > t\} = 1 - F(t)$ is a sample statistic. Under H_0, the p-value has a cdf given by

$$F_p(p|H_0) = 1 - F\left[F^{-1}(1-p)|H_0\right] = p, \tag{112}$$

while under the alternative H_1 we have

$$F_p(p|H_1) = \Pr\{P \leq p|H_1\} = 1 - G\left(\left(F^{-1}(1-p)|H_0\right)\right). \tag{113}$$

Hence, the density function of the p-value (if it exists) is given by

$$\begin{aligned} f_p(p|H_1) &= \frac{\partial}{\partial p} F_p(p|H_1) \\ &= -g\left(F^{-1}(1-p)\right) \cdot \frac{-1}{f\left(F^{-1}(1-p)\right)} \\ &= \frac{g\left(F^{-1}(1-p)\right)}{f\left(F^{-1}(1-p)\right)}. \end{aligned} \tag{114}$$

This is nothing but the "likelihood ratio" as discussed by Egon Pearson (1938 [84], p. 138) and given in equation (18). If we reject H_0 if the sample statistic $T > k$, then the probability of Type I error is given by $\alpha = \Pr\{T > k|H_0\} = 1 - F(k)$ while the power of the test is given by

$$\beta = \Pr\{T > k|H_1\} = 1 - G(k) = 1 - G\left(F^{-1}(1-\alpha)\right). \tag{115}$$

Hence, the main point of Donahue (1999 [37]) is that, if we have a small p-value, we can say that the test is significant, and we can also refer to the strength of the significance of the test. This,

however, is usually not the case when we fail to reject the null hypothesis. In that case, we do not have any indication about the probability of Type II error that is being committed. This is reflected by the power and size relationship given in (115).

The p-value and its generalization, however, are firmly embedded in Bayes theory as the tail probability of a predictive density. In order to calculate the p-value, Meng (1994 [74]) also considered having a nuisance parameter in the likelihood function or predictive density. We can see that the classical p-value is given by $p = P\left\{T\left(X\right) \geq T\left(x\right)|H_0\right\}$, where $T\left(.\right)$ is a sample statistic and x is a realization of the random sample X that is assumed to follow a density function $f\left(X|\xi\right)$, where $\xi = \left(\delta', \gamma'\right)' \in \Xi$. Suppose now, we have to test $H_0 : \delta = \delta_0$ against $H_1 : \delta \neq \delta_0$. In Bayesian terms, we can replace X by a future replication of x, call it x^{rep}, which is like a "future observation." Hence, we define the predictive p-value as $p_B = P\left\{T\left(x^{rep}\right) \geq T\left(x\right)|x, H_0\right\}$ calculated under the posterior predictive density

$$f\left(x^{rep}|x, H_0\right) = \int_{\Xi} f\left(x^{rep}|\xi\right) \Pi_0\left(d\xi|x\right) = \int_{\Gamma_0} f\left(x^{rep}|\delta_0, \gamma\right) \pi_0\left(\xi|x\right) d\xi, \tag{116}$$

where $\Pi_0\left(\xi|x\right)$ and $\pi_0\left(\xi|x\right)$ are respectively the posterior predictive distribution and density functions of ξ, given x, and under H_0. Simplification in (116) is obtained by assuming $\Gamma_0 = \left\{\xi : H_0 \text{ is true}\right\} = \left\{\left(\delta_0, \gamma\right) : \gamma \in A,\ A \subset \mathbb{R}^d, d \geq 1\right\}$ and defining $\bar{\pi}\left(\gamma|\delta_0\right) = \pi\left(\delta, \gamma|\delta = \delta_0\right)$, which gives

$$\begin{aligned} \pi_0\left(\xi|x\right) &= \frac{f\left(x|\xi,H_0\right) \pi\left(\xi|H_0\right)}{\int_{\Gamma_0} f\left(x|\xi,H_0\right) \pi\left(\xi|H_0\right) d\xi},\ \xi \in \Gamma_0, \\ &= \frac{f\left(x|\delta = \delta_0, \gamma\right) \pi\left(\delta, \gamma|\delta = \delta_0\right)}{\int_{\Gamma_0} f\left(x|\delta = \delta_0, \gamma\right) \pi\left(\delta, \gamma|\delta = \delta_0\right) d\xi}, \\ &= \frac{f\left(x|\delta_0, \gamma\right) \bar{\pi}\left(\gamma|\delta_0\right)}{\int_A f\left(x|\delta_0, \gamma\right) \bar{\pi}\left(\gamma|\delta_0\right) d\gamma},\ \gamma \in A. \end{aligned} \tag{117}$$

This can also be generalized to the case of a composite hypothesis by taking the integral over all possible values of $\delta \in \Delta_0$, the parameter space under H_0. An alternative formulation of the p-value, which makes it clearer that the distribution of the p-value depends on the nuisance parameter γ, is given by $p\left(\gamma\right) \equiv P\left\{D\left(X, \xi\right) \geq D\left(x, \xi\right)|\delta_0, \gamma\right\}$, where the probability is taken over the sampling distribution $f\left(X|\delta_0, \gamma\right)$, and $D\left(X, \xi\right)$ is a test statistic in the classical sense that can be taken as a measure of discrepancy. In order to estimate the p-value $p\left(\gamma\right)$ given that γ is unknown, the obvious Bayesian approach is to take the mean of $p\left(\gamma\right)$ over the posterior distribution of γ under H_0, i.e., $E\left[p\left(\gamma\right)|x, H_0\right] = p_B$.

The above procedure of finding the distribution of the p-value can be used in diagnostic procedures in a Markov chain Monte Carlo setting discussed by Kim, Shephard and Chib (1998

[61]). Following Kim et al. (1998 [61], pp. 361-362), let us consider the simple stochastic volatility model

$$\begin{aligned} y_t &= \beta e^{h_t/2}\varepsilon_t,\ t \geq 1, \\ h_{t+1} &= \mu + \phi\left(h_t - \mu\right) + \sigma_\eta \eta_t, \\ h_t &\sim N\left(\mu, \frac{\sigma_\eta^2}{1-\phi^2}\right), \end{aligned} \tag{118}$$

where y_t is the mean corrected return on holding an asset at time t, h_t is the log volatility which is assumed to be stationary (i.e., $|\phi| < 1$) and h_1 is drawn from a stationary distribution and, finally, ε_t and η_t are uncorrelated standard normal white noise terms. Here, β can be interpreted as the modal instantaneous volatility and ϕ is a measure of the persistence of volatility and σ_η is the volatility of log volatility h_t.[14]

Our main interest is handling of model diagnostics under the Markov chain Monte Carlo method. Defining $\xi = \left(\mu, \phi, \sigma_\eta^2\right)'$, the problem is to sample from the distribution of $h_t|Y_t, \xi$, given a sample of draws $h_{t-1}^1, h_{t-1}^2, ..., h_{t-1}^M$ from $h_{t-1}|Y_{t-1}, \xi$, where we can assume ξ to be fixed. Using the Bayes rule discussed in equations (116) and (117), the one-step-ahead prediction density is given by

$$f\left(y_{t+1}|Y_t, \xi\right) = \int f\left(y_{t+1}|Y_t, h_{t+1}, \xi\right) f\left(h_{t+1}|Y_t, h_t, \xi\right) f\left(h_t|Y_t, \xi\right) dh_{t+1} dh_t, \tag{119}$$

and for each value of h_t^j $(j = 1, 2, \ldots, M)$, we sample h_{t+1}^j from the conditional distribution h_{t+1}^j given h_t^j . Based on M such draws, we can estimate the probability that y_{t+1}^2 would be less than the observed y_{t+1}^{o2} is given by

$$P\left(y_{t+1}^2 \leq y_{t+1}^{o2}|Y_t, \theta\right) \cong u_{t+1}^M = \frac{1}{M}\sum_{j=1}^{M} P\left(y_{t+1}^2 \leq y_{t+1}^{o2}|h_{t+1}^j, \xi\right), \tag{120}$$

which is the sample equivalent of the posterior mean of the probabilities discussed in Meng (1994 [74]). Hence, u_{t+1}^M under the correctly specified model will be *IID* $U\left(0,1\right)$ distribution as $M \to \infty$. This result is an extension of Karl Pearson (1933 [87], 1934 [88]), Egon Pearson (1938 [84]) and Rosenblatt (1952 [101]) discussed earlier and is very much in the spirit of the goodness-of-fit test suggested by Neyman (1937 [76]). Kim et al. (1998 [61]) also discussed a procedure similar to the one followed by Berkowitz (2000 [17]), where instead of looking at the just u_{t+1}^M, they look at the inverse Gaussian transformation, then carry out tests on normality, autocorrelation and

[14] As Kim, Shepherd and Chib (1998 [61], p. 362) noted that the parameters β and μ are related in the true model by $\beta = \exp\left(\mu/2\right)$, however when estimating the model they set $\beta = 1$ and left μ unrestricted. Finally, they reported the estimated value of β from the estimated model as $\exp\left(\mu/2\right)$.

heteroscedasticity. A more comprehensive test could be performed on the validity of forecasted density based on Neyman's smooth test techniques that we discussed in Subsection 4.2 in connection to the forecast density evaluation literature (Diebold et al. 1998 [33]). We believe that the smooth test provide a more constructive procedure instead of just checking uniformity of an average empirical distribution function u_{t+1}^{M} on the square of the observed values y_{t+1}^{o2} given in (120) and other graphical techniques like the Q-Q plots and correlograms as suggested by Kim et al. (1998 [61], pp. 380-382).

5 Epilogue

Once in a great while a paper is written that is truly fundamental. Neyman's (1937 [76]) is one that seem impossible to compare with anything but oneself given the statistical scene in the 1930s. Starting from the very first principles of testing, Neyman derived an *optimal* test statistic and discussed its applications along with its possible drawbacks. Earlier tests, such as Karl Pearson's (1900 [86]) goodness-of-fit and Jerzy Neyman and Egon Pearson's (1928 [79]) likelihood ratio tests are also fundamental, but those test statistics were mainly based on intuitive grounds and had no claim for optimality when they were proposed. In terms of its significance in the history of hypothesis testing, Neyman (1937 [76]) is comparable only to the later papers by the likes of Wald (1943 [113]), Rao (1948 [93]) and Neyman (1959 [77]), each of which also proposed fundamental test principles that satisfied certain optimality criteria.

Although econometrics is a separate discipline, it is safe to say that the main fulcrum of advances in econometrics is, as it always has been, the statistical theory. From that point of view, there is much to gain from borrowing suitable statistical techniques and adapting them for econometric applications. Given the fundamental nature of Neyman's (1937 [76]) contribution, we are surprised that the smooth test has not been formally used in econometrics, to the best of our knowledge. And this paper is our modest attempt to bring Neyman's smooth test to mainstream econometric research.

Acknowledgements:

We would like to thank Aman Ullah and Alan Wan without whose encouragement and prodding, this paper would not have been completed. We are also grateful to an anonymous referee and Zhijie Xiao for many helpful suggestions that have considerably improved the paper. However, we retain the responsibility for any remaining errors.

References

[1] P.K. Anderson, O. Borgan, R.D. Gill, N. Keiding. Linear nonparametric tests for comparison of counting processes, with applications to censored survival data. International Statistical Review 50:219-258, 1982.

[2] P. K. Anderson, O. Borgan, R.D. Gill, N. Keiding. Statistical Models Based on Counting Processes. New York: Springer-Verlag, 1991.

[3] D.E. Barton. On Neyman's smooth test of goodness of fit and its power with respect to a particular system of alternatives. Skandinaviske Aktuarietidskrift 36:24-63, 1953.

[4] D.E. Barton. The probability distribution function of a sum of squares. Trabajos de Estadistica 4:199-207, 1953.

[5] D.E. Barton. A form of Neyman's χ^2 test of goodness of fit applicable to grouped and discrete data. Skandinaviske Aktuarietidskrift 38:1-16, 1955.

[6] D.E. Barton. Neyman's ψ_k^2 test of goodness of fit when the null hypothesis is composite. Skandinaviske Aktuarietidskrift 39:216-246, 1956.

[7] D.E. Barton. Neyman's and other smooth goodness-of-fit tests. In: S. Kotz and N.L. Johnson , eds. Encylcopedia of Statistical Sciences, Vol. 6. New York: Wiley, 1985, pp. 230-232.

[8] Rev. T. Bayes. An essay toward solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society 53:370-418, 1763.

[9] B.J. Becker. Combination of p-values. In: S. Kotz, C.B. Read and D.L. Banks, eds. Encyclopedia of Statistical Sciences, Update Vol. I. New York: Wiley, 1977, pp. 448-453.

[10] A.K. Bera. Hypothesis testing in the 20^{th} century with special reference to testing with misspecified models. In: C.R. Rao and G. Szekely, eds. Statistics in the 21st Century. New York: Marcel Dekker, 2000, pp. 33-92.

[11] A.K. Bera, Y. Bilias. Rao's score, Neyman's C(α) and Silvey's LM test: An essay on historical developments and some new results. Journal of Statistical Planning and Inference, Volume 97, 2001, forthcoming (2001a).

[12] A.K. Bera, Y. Bilias. On some optimality properties of Fisher-Rao score function in testing and estimation. Communications in Statistics, Theory and Methods, 2001, forthcoming (2001b).

[13] A.K. Bera, Y. Bilias. The MM, ME, MLE, EL, EF, and GMM approaches to estimation: A synthesis. Manuscript, 2001 (2001c).

[14] A.K. Bera, A. Ghosh. Evaluation of density forecasts using Neyman's smooth test. Work in Progress, 2001.

[15] A.K. Bera, G. Premaratne. General hypothesis testing. In: Badi Baltagi, Basil Blackwell, eds. Companion in Econometric Theory. Oxford: Blackwell Publishers, 2001, pp. 38-61.

[16] A.K. Bera, A. Ullah. Rao's score test in econometrics. Journal of Quantitative Economics 7:189-220, 1991.

[17] J. Berkowitz. The accuracy of density forecasts in risk management. Manuscript, 2000.

[18] P.J. Bickel, K.A. Doksum. Mathematical Statistics: Basic Ideas and Selected Topics. Oakland, California: Holden-Day, 1977.

[19] B. Boulerice, G.R. Ducharme. A note on smooth tests of goodness of fit for location-scale families. Biometrika 82:437-438, 1995.

[20] A.C. Cameron, P.K. Trivedi. Conditional moment tests and orthogonal polynomials, Working paper in Economics, Number 90-051, Indiana University, 1990.

[21] T.K. Chandra, S.N. Joshi. Comparison of the likelihood ratio, Rao's and Wald's tests and a conjecture by C.R. Rao. Sankhyā Ser. A 45:226-246, 1983.

[22] A.D. Chesher. Testing for neglected heterogeneity. Econometrica 52:865-872, 1984.

[23] S. Choi, W.J. Hall, A. Schick. Asymptotically uniformly most powerful tests in parametric and semiparametric models. Annals of Statistics 24:841-861, 1996.

[24] P.F. Christoffersen. Evaluating interval forecasts. International Economic Review 39:841-862, 1998.

[25] G.M. Cordeiro, S.L.P. Ferrari. A modified score test statistic having chi-squared distribution to order n^{-1}. Biometrika 78:573-582, 1991.

[26] D.R. Cox. Some remarks on over-dispersion. Biometrika 70:269-274, 1983.

[27] D.R. Cox, D. Oakes. Analysis of Survival Data. New York: Chapman and Hall, 1984.

[28] H. Cramér . Mathematical methods of Statistics. New Jersey: Princeton University Press, 1946.

[29] F. Cribari-Neto, S.L.P. Ferrari. An improved Lagrange multiplier test for heteroscedasticity. Communications in Statistics–Simulation and Computation 24:31-44, 1995.

[30] C. Crnkovic, J. Drachman. Quality control. In: VAR: Understanding and Applying Value-at-Risk. London: Risk Publication. 1997.

[31] R.B. D'Agostino, M.A. Stephens. Goodness-of-Fit Techniques. New York: Marcel Dekker, 1986.

[32] A.P. Dempster, N.M Laird, D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B 39:1-38, 1977.

[33] F.X. Diebold, T.A. Gunther, A.S. Tay. Evaluating density forecasts with applications to financial risk management. International Economic Review 39:863-883, 1998.

[34] F.X. Diebold, J. Hahn, A.S. Tay. Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns in foreign exchange. Review of Economics and Statistics 81:661-673, 1999.

[35] F.X. Diebold, J.A. Lopez. Forecast evaluation and combination. In: G.S. Maddala and C.R. Rao, eds. Handbook of Statistics, Vol. 14. Amsterdam: North-Holland, 1996, pp. 241-268.

[36] F.X. Diebold, A.S. Tay, K.F. Wallis. Evaluating density forecasts of inflation: the survey of professional forecasters. In: R.F Engle, H. White, eds. Cointegration, Causality and Forecasting: Festschrift in Honour of Clive W. Granger. New York: Oxford University Press, 1999, pp. 76-90.

[37] R.M.J. Donahue. A note on information seldom reported via the p value. American Statistician 53:303-306, 1999.

[38] R.L. Eubank, V.N. LaRiccia. Asymptotic comparison of Cramér-von Mises and nonparametric function estimation techniques for testing goodness-of-fit. Annals of Statistics 20:1412-1425, 1992.

[39] J. Fan. Test of significance based on wavelet thresholding and Neyman's truncation. Journal of the American Statistical Association 91:674-688, 1996.

[40] R.A. Fisher. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society A222:309-368, 1922.

[41] R.A. Fisher. Inverse Probability. Proceedings of the Cambridge Philosophical Society 36:528-535, 1930.

[42] R.A. Fisher. The logical inversion of the notion of a random variable. Sankhyā 7: 129-132, 1945.

[43] R.A. Fisher. Statistical Methods for Research Workers. 13th ed. New York: Hafner Publishing Company Inc, 1958.

[44] J.K. Ghosh. Higher order asymptotics for the likelihood ratio, Rao's and Wald's tests. Statistics & Probability Letters 12:505-509, 1991.

[45] R.J. Gray, D.A. Pierce. Goodness-of-fit tests for censored survival data. Annals of Statistics 13: 552-563, 1985.

[46] T. Haavelmo. The probability approach in econometrics. Supplements to Econometrica 12, 1944.

[47] T. Haavelmo. Econometrics and the welfare state: Nobel lecture, December 1989. American Economic Review 87:13-15, 1997.

[48] M.A. Hamdan. The power of certain smooth tests of goodness of fit. Australian Journal of Statistics 4:25-40, 1962.

[49] M.A. Hamdan. A smooth tests of goodness of fit based on Walsh functions. Australian Journal of Statistics 6:130-136, 1964.

[50] P. Harris. An asymptotic expansion for the null distribution of the efficient score statistics. Biometrika 72:653-659, 1985.

[51] P. Harris. Correction to 'An asymptotic expansion for the null distribution of the efficient score statistic.' Biometrika 74:667, 1987.

[52] J. D. Hart. Nonparametric Smoothing and Lack of Fit Tests. New York: Springer-Verlag, 1997.

[53] M. Hollander, E.A. Peña. Classes of nonparametric goodness-of-fit tests for censored data. In: A.K. Md. E. Saleh, ed. A New Approach in Nonparametric Statistics and Related Topics, Amsterdam: Elsevier, 1992, pp. 97-118.

[54] T. Inglot, T. Jurlewicz, T. Ledwina. On Neyman-type smooth tests of fit. Statistics 21:549-568, 1990.

[55] T. Inglot, W.C.M. Kallenberg, T. Ledwina. Power approximations to and power comparison of smooth goodness-of-fit tests. Scandinavian Journal of Statistics 21:131-145, 1994.

[56] S.L. Isaacson. On the theory of unbiased tests of simple statistical hypothesis specifying the values of two or more parameters. Annals of Mathematical Statistics 22:217-234, 1951.

[57] N. L. Johnson, S. Kotz. Continuous Univariate Distributions-1. New York: John Wiley & Sons, 1970.

[58] N. L. Johnson, S. Kotz. Continuous Univariate Distributions-2. New York: John Wiley & Sons, 1970.

[59] W.C.M. Kallenberg, J. Oosterhoff, B.F. Schriever. The number of classes in χ^2 goodness of fit test. Journal of the American Statistical Association 80:959-968, 1985 .

[60] N.M. Kiefer. Specification diagnostics based on Laguerre alternatives for econometric models of duration. Journal of Econometrics 28:135-154, 1985.

[61] S. Kim, N. Shephard, S.Chib. Stochastic volatility: Likelihood inference and comparison with ARCH models. Review of Ecomonic Studies 65:361-393, 1998.

[62] L. Klein. The Statistics Seminar, MIT, 1942-43. Statistical Science 6: 320-330, 1991.

[63] K. J. Kopecky, D.A. Pierce. Efficiency of smooth goodness-of-fit tests. Journal of the American Statistical Association 74: 393-397, 1979.

[64] J.A. Koziol. An alternative formulation of Neyman's smooth goodness of fit tests under composite alternatives. Metrika 34:17-24, 1987.

[65] P.H. Kupiec. Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives.Winter: 73-84, 1995.

[66] T. Lancaster. Generalized residuals and heterogeneous duration models. Journal of Econometrics 28: 155-169, 1985.

[67] V.N. LaRiccia. Smooth goodness-of-fit tests: A quantile function approach. Journal of the American Statistical Association 86:427-431, 1991.

[68] T. Ledwina. Data-driven version of Neyman's smooth test of fit. Journal of the American Statistical Association 89:1000-1005, 1994.

[69] L.-F. Lee. Maximum likelihood estimation and a specification test for non-normal distributional assumption for the accelerated failure time models. Journal of Econometrics 24:159-179, 1984.

[70] L.-F. Lee. Specification test for Poisson regression models. International Economic Review 27:689-706, 1986.

[71] E.L. Lehmann. Testing Statistical Hypothesis. New York: John Wiley & Sons, 1959.

[72] G.S. Maddala. Econometrics. New York: McGraw-Hill, 1977.

[73] P.C. Mahalanobis. A revision of Risley's anthropometric data relating to Chitagong hill tribes. Sankhyā B 1:267-276, 1934.

[74] X.-L. Meng. Posterior predictive p-values. Annals of Statistics. 22: 1142-1160, 1994.

[75] R. Mukerjee. Rao's score test: recent asymptotic results. In: G.S. Madala, C.R. Rao, H.D. Vinod, eds. Handbook of Statistics, Vol. 11. Amsterdam: North-Holland, 1993, pp. 363-379.

[76] J. Neyman. "Smooth test" for goodness of fit. Skandinaviske Aktuarietidskrift 20:150-199, 1937.

[77] J. Neyman. Optimal asymptotic test of composite statistical hypothesis. In: U. Grenander, ed. Probability and Statistics, the Harold Cramér Volume. Uppsala: Almqvist and Wiksell. 1959, pp. 213-234.

[78] J. Neyman. Some memorable incidents in probabilistic/statistical studies. In: I. M. Chakrabarti, ed. Asymptotic Theory of Statistical Tests and Estimation. New York: Academic Press. 1980, pp. 1-32.

[79] J. Neyman, E.S. Pearson. On the use and interpretation of certain test criteria for purpose of statistical inference. Biometrika 20:175-240, 1928.

[80] J. Neyman, E.S. Pearson. On the problem of the most efficient tests of statistical hypothesis. Philosophical Transactions of the Royal Society Series A 231:289-337, 1933.

[81] J. Neyman, E.S. Pearson. Contributions to the theory of testing statistical hypothesis I: Unbiased critical regions of Type A and A_1. Statistical Research Memoirs 1:1-37, 1936.

[82] J. Neyman, E.S. Pearson. Contributions to the theory of testing statistical hypothesis. Statistical Research Memoirs 2:25-57, 1938.

[83] F. O'Reilly, C.P. Quesenberry. The conditional probability integral transformation and applications to obtain composite chi-square goodness of fit tests. Annals of Statistics 1: 74-83, 1973.

[84] E.S. Pearson. The probability integral transformation for testing goodness of fit and combining independent tests of significance. Biometrika 30:134-148, 1938.

[85] E.S. Pearson. The Neyman-Pearson story: 1926-34, historical sidelights on an episode in Anglo-Polish collaboration. In: F.N. David, ed. Reasearch Papers in Statistics, Festscrift for J. Neyman. New York: John Wiley and Sons, 1966, pp. 1-23.

[86] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. Philosophical Magazine 5^{th} Series 50:157-175, 1900.

[87] K. Pearson. On a method of determining whether a sample of size n is supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. Biometrika 25:379-410, 1933.

[88] K. Pearson. On a new method of determining "goodness of fit." Biometrika 26:425-442, 1934.

[89] E. Peña. Smooth goodness-of-fit tests for the baseline hazard in Cox's proportional hazards model. Journal of the American Statistical Association 93:673-692, 1998.

[90] E. Peña. Smooth goodness-of-fit tests for composite hypothesis in hazard based models. Annals of Statistics 28:1935-1971, 1998.

[91] D.A. Pierce. The asymptotic effect of substituting estimators for parameters in certain types of statistics. Annals of Statistics 10:475-478, 1982.

[92] C.P. Queesenberry. Some transformation methods in goodness-of-fit. In: R.B. D'Agostino, M.A. Stephens, eds. Goodness-of-Fit Techniques. New York: Marcel Dekker, 1986, pp. 235-277.

[93] C. R. Rao. Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. Proceedings of the Cambridge Philosophical Society 44:50-57, 1948.

[94] C.R. Rao. Advanced Statistical Methods in Biometric Research. New York: John Wiley and Sons, 1952.

[95] C.R. Rao. Linear Statistical Inference and Its Applications. New York: John Wiley & Sons, 1973.

[96] C. R. Rao, R. Mukerjee. Tests based on score statistics: power properties and related results. Mathematical Methods of Statistics 3:46-61, 1994.

[97] C. R. Rao, R. Mukerjee. Comparison of LR, score and Wald tests in a non-IID setting. Journal of Multivariate Analysis 60:99-110, 1997.

[98] C.R. Rao, S.J. Poti. On locally most powerful tests when the alternatives are one-sided. Sankhyā 7:439, 1946.

[99] J.C.W. Rayner, D.J. Best. Smooth Tests of Goodness of Fit. New York: Oxford University Press, 1989.

[100] C. Reid. Neyman–From Life. New York: Springer-Verlag, 1982.

[101] M. Rosenblatt. Remarks on a multivariate transformation. Annals of Mathematical Statistics 23:470-472, 1952.

[102] A. SenGupta, L. Vermeire. Locally optimal tests for multiparameter hypotheses. Journal of the American Statistical Association 81:819-825, 1986.

[103] R. J. Smith. On the use of distributional misspecification checks in limited dependent variable models. The Economic Journal 99 (Supplement Conference Papers):178-192, 1989.

[104] C. Smithson, L. Minton. How to calculate VAR. In: VAR: understanding and applying value-at risk. London: Risk Publications, 1997, pp. 27-30.

[105] H. Solomon, M.A. Stephens. Neyman's test for uniformity. In: S. Kotz and N.L. Johnson , eds. Encylcopedia of Statistical Sciences Vol. 6. New York: Wiley, 1985, pp. 232-235.

[106] E.S. Soofi. Information theoretic regression methods. In: T.M. Fomby, R.C. Hill, eds. Advances in Econometrics Vol 12. Greenwich: Jai Press, 1997, pp. 25-83.

[107] E.S. Soofi. Principal information theoretic approaches. Journal of the American Statistical Association 95:1349-1353, 2000.

[108] C.J. Stone. Large-sample inference for log-spline models. Annals of Statistics 18:717-741, 1990.

[109] C.J. Stone, C-Y. Koo. Logspline density estimation. Contemporary Mathematics 59:1-15, 1986.

[110] A.S. Tay, K.F. Wallis. Density forecasting: A Survey. Journal of Forecasting 19:235-254, 2000.

[111] D.R. Thomas, D.A. Pierce. Neyman's smooth goodness-of-fit test when the hypothesis is composite. Journal of the American Statistical Association 74:441-445, 1979.

[112] L.M.C. Tippett. The Methods of Statistics. 1^{st} Edition. London: Williams and Norgate, 1931.

[113] A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society 54:426-482, 1943.

[114] H. White. Maximum likelihood estimation of misspecified models. Econometrica 50: 1-25, 1982.

[115] B. Wilkinson. A statistical consideration in psychological research. Psychology Bulletin 48: 156-158, 1951.