Journal of Econometrics 188 (2015) 490–501

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

A data-driven smooth test of symmetry

Ying Fang [a], Qi Li [b,c,*], Ximing Wu [d,e], Daiqiang Zhang [b]

[a] Wang Yanan Institute for Studies in Economics, MOE Key Laboratory of Econometrics, Fujian Key Laboratory of Statistical Science, Xiamen University, Xiamen, Fujian 361005, China

[b] Department of Economics, Texas A&M University, College Station, TX 77843, USA

[c] ISEM, Capital University of Economics and Business, Beijing, China

[d] Department of Agricultural Economics, Texas A&M University, College Station, TX 77843, USA

[e] School of Economics, Fujian Agricultural and Forestry University, Fuzhou, China

ARTICLE INFO

Article history:
Available online 25 March 2015

ABSTRACT

In this paper we propose a data driven smooth test of symmetry. We first transform the raw data via the probability integral transformation according to a symmetrized empirical distribution, and show that under the null hypothesis of symmetry, the transformed data has a limiting uniform distribution, reducing testing for symmetry to testing for uniformity. Employing Neyman's smooth test of uniformity, we show that only odd-ordered orthogonal moments of the transformed data are required in constructing the test statistic. We present a standardized smooth test that is distribution-free asymptotically and derive the asymptotic behavior of the test and establish its consistency. Extension to dependent data case is discussed. We investigate the finite sample performance of the proposed tests on both homogeneous and mixed distributions (with unobserved heterogeneity). An empirical application on testing symmetry of wage adjustment process, based on heterogeneous wage contracts with different durations, is provided.

1. Introduction

Many economic variables are known to behave differently under contrasting situations. For instance, Bacon (1991) made the famous observation that the retail price of gasoline, relative to its wholesale price, rises like a rocket but falls like a feather. Whether a quantity of interest exhibits asymmetric behavior plays an important role in the investigation of many economic questions. For example, in macroeconomics, researchers want to know if an economic variable behaves similarly during expansions and recessions; in finance, investors are interested in knowing whether return distributions are skewed. Answers to these important questions provide valuable input in people's decision making.

The knowledge of symmetry may have important consequence in statistical and econometric analyses. For example under symmetry, convergence rate of bootstrap confidence interval coverage probability is of order n^{-1} rather than the typical $n^{-1/2}$ (Hall, 1995). The intercept of a linear regression is identified in adaptive estimations when the error distribution is symmetric (Bickel, 1982). The efficiency of estimators can sometimes be improved by incorporating symmetry (see, e.g., Horowitz and Markatou (1996) for panel model estimation via deconvolution, and Chen (2000) for binary choice models). Therefore, it is important to test whether the data one analyze is symmetrically distributed or not. Testing for symmetry has a long history in statistics and econometrics. One can characterize symmetry tests into two broad categories. One approach is based on some characteristics of symmetric distributions. For instance, there are tests based on moments (see, e.g., Bai and Ng (2005) and Premaratne and Bera (2005)), tests based on ranks (e.g., Hájek and Sidak (1967) and Shorack and Wellner (1986)), and tests based on the empirical characteristic functions (e.g., Csörgő and Heathcote (1987)). The second approach employs a symmetrization of the sample distribution and constructs a proper distance between the sample distribution/density and its symmetrization as the test statistic. For a random sample $\{Y_i\}_{i=1}^n$ from a continuous distribution, denote the empirical distribution by $\hat{F}$ and its symmetrization about a center tendency parameter θ by $s\hat{F}_\theta$, which is defined below. Butler (1969) and Schuster and Barker (1987) designed a Kolmogorov–Smirnov type test based on $\sup_y |\hat{F}(y) - s\hat{F}_\theta(y)|$. Rothman and Woodroofe (1972) proposed a test based on the integrated squared difference in distributions $\int(\hat{F}(y) - s\hat{F}_\theta(y))^2 dy$. There also exist tests based

* Corresponding author at: Department of Economics, Texas A&M University, College Station, TX 77843, USA.
E-mail address: qi@econmail.tamu.edu (Q. Li).

http://dx.doi.org/10.1016/j.jeconom.2015.03.013

on distance in densities. Let $\hat{f}$ and $s\hat{f}_\theta$ be densities, if they exist, associated with F_n and $s\hat{F}_\theta$ respectively. Ahmad and Li (1997) suggested a test based on the integrated squared difference $\int \left(\hat{f}(y) - s\hat{f}_\theta(y)\right)^2 dy$, and Maasoumi and Racine (2009) presented a test based on the Hellinger distance $\int \left(\sqrt{\hat{f}(y)} - \sqrt{s\hat{f}_\theta(y)}\right)^2 dy$.

In this paper we follow the second approach and propose an alternative test that explores the discrepancy between the sample distribution and its symmetrization. Our approach is based on the probability integral transformation $X_i(\theta) = s\hat{F}_\theta(Y_i), i = 1, \ldots, n$. Under the null hypothesis of symmetry, $\hat{F}$ should be close to $s\hat{F}_\theta$ and $\{X_i(\theta)\}_{i=1}^n$ is distributed according to the standard uniform distribution asymptotically. Thus testing for symmetry is reduced to testing for uniformity. There exists a myriad of uniformity tests in the literature. Among them, Neyman's smooth test is particularly popular; thanks to its appealing theoretical optimality, ease of construction and good small sample performance. Its finite sample performance, however, can be sensitive to the number of terms contained in the test statistic. We therefore adopt a data driven version of Neyman's test proposed by Ledwina (1994). This test chooses the number of terms according to an information criterion and is shown to be consistent and adapts to the underlying distribution.

We show that when orthogonal moments of the transformed data are used, only odd-ordered moments are required. This results hold true for arbitrary (symmetric or asymmetric) distributions and arbitrary location parameters. Because of the presence of $s\hat{F}_{\hat{\theta}}$, an effectively infinite-dimensional nuisance parameter, the test is composite and thus not free from nuisance parameters. We derive the asymptotic distribution of the test statistic and propose a standardized version that is distribution free asymptotically and easy to calculate. We also extend the proposed test to dependent data case.

We demonstrate that our proposed test offers several advantages. First, it is invariant to monotone transformation of data — thanks to the probability integral transformation of raw data. Second, it has power against essentially any asymmetric alternatives that satisfy some mild conditions and at the same time adapts to the unknown distribution. Third, the proposed test is more robust than tests based on moments of raw data because high order moments of raw data can be sensitive to outliers. Fourth, compared with tests based on nonparametrically estimated distributions/densities on raw data, the proposed test, which reduces to a test of uniformity of the transformed data, is based on an "easier" approximation problem because the asymptotic bias of approximating nonparametrically a uniform distribution is zero. Lastly, our test only requires straightforward calculations of a number of sample moments and is therefore computationally easy and performs well for small sample sizes.

The remaining part of the paper is organized as follows. Section 2 briefly reviews the data driven Neyman's test. Section 3 presents the proposed symmetry test based on Neyman's approach, its asymptotic properties and a bootstrap method of inference. Section 4 extends the proposed test to dependent data case. Section 5 reports Monte Carlo simulations on homogeneous distributions and mixed distributions. Section 6 provides an empirical example on testing symmetry of wage adjustment process based on wage contracts of varying durations. The last section concludes. Mathematical proofs are gathered in an Appendix.

2. Smooth test of distributions

We are interested in testing the hypothesis that an iid random sample $\{Y_i\}_{i=1}^n$ are generated from a known continuous distribution F_Y. Denote the probability integral transformation of Y_i by $X_i = F_Y(Y_i), i = 1, \ldots, n$. It follows that $\{X_i\}_{i=1}^n$ are distributed according to the standard uniform distribution under the null hypothesis. Thus all tests of distributions can be formulated as a test of uniformity following the probability integral transformation.

There exist many tests of uniformity. In this study we focus on Neyman's (1937) smooth test, which enjoys appealing theoretical properties and good finite sample performance (c.f., a review of goodness-of-fit tests by Rayner and Best (1990)). Let $X = \{X_i\}_{i=1}^n$ be an iid sample from a continuous distribution p_0 defined on [0, 1]. The null hypothesis is that p_0 is the uniform distribution. Neyman (1937) considered the following family of alternative distributions

$$p(x; \gamma^{(k)}) = c(\gamma^{(k)}) \exp\left\{\sum_{j=1}^k \gamma_j b_j(x)\right\}, \quad x \in [0, 1], \ k \geq 1, \quad (1)$$

where $b_1, \ldots, b_k$ are normalized Legendre polynomials on [0, 1], $\gamma^{(k)} = (\gamma_1, \ldots, \gamma_k)' \in R^k$, and $c(\gamma^{(k)}) = \left(\int_{[0,1]} \exp\left\{\sum_{j=1}^k \gamma_j b_j(x)\right\} dx\right)^{-1}$ is a normalization factor.

When $\gamma_j = 0$ for all $j = 1, \ldots, k$ (i.e., $\gamma^{(k)}$ is a zero vector), we have $p(x; 0) = c(0) = 1$, corresponding to the uniform distribution. Thus the test of uniformity amounts to the test on $H_0 : \gamma^{(k)} = 0$ versus $H_1 : \gamma^{(k)} \neq 0$ under this framework. Note that the normalized Legendre polynomials are orthonormal with respect to the Lebesgue measure on [0, 1] (the null distribution) such that $\int b_i(x)b_j(x)dx = 0$ for $i \neq j$ and $\int b_j^2(x)dx = 1$ for $1 \leq i, j \leq k$. A score version of Neyman's test is particularly convenient. Define

$$N_k = \sum_{j=1}^k \left(n^{-1/2} \sum_{i=1}^n b_j(X_i)\right)^2. \quad (2)$$

The test statistic N_k is an asymptotically locally optimal solution to the original testing problem. In practice we reject the null hypothesis is rejected for large values of N_k.

The term *smooth test* comes from the fact that the alternative distribution (1) is a smooth deviation from the uniform distribution in directions suggested by b_j's. Despite its well-documented good performance, one limitation of the smooth test is that there is no clear guidance on how to select the number of terms k in (1), which affects the power of the test. Traditionally, many authors restrict their attentions to studying the power of Neyman's test with a small k. The past two decades have seen some studies that consider data driven methods in the selection of k. Usually a two step procedure is employed. First, a proposed selection rule is applied to find a suitable distribution within the family of (1). In the second step, Neyman's test is applied to the selected model. The test is thus adaptive in the sense that Neyman's test is applied in the 'right' direction.

Ledwina (1994) pioneered the data driven smooth tests and proposed a test using a Schwartz's Bayesian Information Criteria (BIC) selection rule. We now describe this testing procedure. Let K_n be the maximum number of terms considered for a sample of size n. Define, for $1 \leq s \leq K_n$,

$$L_s = \log \prod_{i=1}^n p(X_i; \hat{\gamma}^{(s)}), \quad (3)$$

where $\hat{\gamma}^{(s)}$ is the MLE solution to model (1). We then choose the best-fitting density according to the BIC

$$S = \min\left\{s : L_s - \frac{1}{2} s \log n \geq L_r - \frac{1}{2} r \log n, \ 1 \leq r, s \leq K_n\right\}. \quad (4)$$

When there are ties in log likelihoods, the model with the smallest dimension is selected. Having chosen a model of dimension S, we use N_S as the data-driven Neyman's test.