Tatra Mt. Math. Publ. **39** (2008), 235–243

TATRA MOUNTAINS
Mathematical Publications

TESTING PROCEDURES BASED ON THE EMPIRICAL CHARACTERISTIC FUNCTIONS II: k-SAMPLE PROBLEM, CHANGE POINT PROBLEM

Marie Hušková — Simos G. Meintanis

ABSTRACT. This is the second part of a partial survey of test procedures based on empirical characteristic functions. We focus on the k sample problem and on detection of a change in the distribution of a sequence independent observations.

1. Introduction

Tests based on empirical characteristic functions (ecf's) for k-sample problem and for the detection of changes are constructed along the line of the procedures described in Part I. Let us recall that they are based on a weighted distance of empirical characteristic function and null hypothesis corresponds to equality of their theoretical counterparts while under alternatives it is positive. We describe the procedures and their properties with relevant references.

2. k sample problem

Let $\boldsymbol{Y}_j = (Y_{j,1}, \ldots, Y_{j,n_j})^T$, $j = 1, \ldots, k$ be k independent random samples, the distribution function related to the jth sample is F_j. For k-sample the testing problem

$$H_0 : F_1 = \ldots = F_k \quad \text{against} \quad H_1 : H_0 \quad \text{is not true} \tag{2.1}$$

2000 Mathematics Subject Classification: Primary: 62G20, Secondary: 62E20, 60F17.
Keywords: empirical characteristic function, k sample problem, change point analysis.
The first author's work was supported by grants GAČR 201/06/0186 and MSM 0021620839.
The second author's research was funded through the Operational Programme for Education and Initial Vocational Training (O.P. "Education") in the framework of the project "Reformation of undergraduate programme of the Department of Economics/University of Athens" with 75% from European Social Funds and 25% from National Funds.

we consider test procedures based on empirical characteristic functions

$$\widehat{\phi}_j(t) = \frac{1}{n_j}\sum_{j=1}^{n_j}\exp\{itY_{js}\}, \qquad t \in R^1, \qquad j = 1,\dots,k, \tag{2.2}$$

$$\widehat{\overline{\phi}}(t) = \sum_{j=1}^{k}\frac{n_j}{n}\widehat{\phi}_j(t), \qquad t \in R^1. \tag{2.3}$$

We assume that the sample sizes satisfy

$$\lim_{n\to\infty} n_j/n = p_j \in (0,1), \qquad j = 1,\dots,k \tag{2.4}$$

and that the weight function $\beta(.)$ fulfills:

$$\beta(t) = \beta(-t) \geq 0, \qquad t \in R^1, \qquad \text{and} \quad 0 < \int_{-\infty}^{\infty}\beta(t)\,\mathrm{d}t < \infty. \tag{2.5}$$

The test statistics for the testing problem (2.1) is defined as

$$T_n(\beta,k) = \int_{-\infty}^{\infty}\sum_{j=1}^{k} n_j|\widehat{\phi}_j(t) - \widehat{\overline{\phi}}(t)|^2\beta(t)\,\mathrm{d}t. \tag{2.6}$$

Clearly, large values indicate that H_0 is violated. The test statistics $T_n(\beta,k)$ can be expressed as

$$T_n(\beta,k) = \int_{-\infty}^{\infty}\sum_{j=1}^{k}\frac{n}{n_j}\big(Z_{jn}(t) - n_j\overline{Z}_n(t)\big)^2\beta(t)\,\mathrm{d}t,$$

where

$$Z_{jn}(t) = \frac{1}{\sqrt{n}}\sum_{s=1}^{n_j}\Big(\big(\cos(tY_{js}) + \sin(tY_{js})\big) - E_{H_0}\big(\cos(tY_{js}) + \sin(tY_{js})\big)\Big),$$

$$\overline{Z}_n(t) = \frac{1}{n}\sum_{j=1}^{k} Z_{jn}(t), \qquad \boldsymbol{Z}_n(t) = \big(Z_{1n}(t),\dots,Z_{kn}(t)\big)^T, \qquad t \in R^1$$

$$\boldsymbol{Z}_n = \big\{\boldsymbol{Z}_n(t),\, t \in R^1\big\}.$$

Next we formulate the assertion on the limit behavior under H_0. We use the notation: $\mathcal{H} = L^2(R^1,\mathcal{B},\beta)$ is a space of measurable functions $f\colon R^1 \to R^1$ satisfying $\int_{-\infty}^{\infty} f^2(t)\beta(t)\,\mathrm{d}t < \infty$ with the inner product and the norm in $\mathcal{H}$ denoted by

$$< f,g >= \int_{-\infty}^{\infty} f(t)g(t)\beta(t)\,\mathrm{d}t \quad \text{and} \quad ||f||^2 = \int_{-\infty}^{\infty} f^2(t)\beta(t)\,\mathrm{d}t,$$

respectively. The notation $\to^D$ means the convergence in distribution of random elements and random variables , $\to^P$ stands for the convergence in probability.

Theorem 2.1. *Let* $\boldsymbol{Y}_j = (Y_{j,1}, \dots, Y_{j,n_j})^T$, $j = 1, \dots, k$ *be independent identically distributed (i.i.d.) random variables with common distribution* F *and let* (2.4) *and* (2.5) *be satisfied. Then, as* $n \to \infty$,

$$\boldsymbol{Z}_n \to^D \boldsymbol{Z}$$

and

$$T_n(\beta, k) \to^D \int_{-\infty}^{\infty} \sum_{j=1}^{k} \frac{1}{p_j} \big(Z_j(t) - p_j \overline{Z}(t)\big)^2 \beta(t)\, \mathrm{d}t,$$

where $\boldsymbol{Z} = \big\{(Z_1(t), \dots, Z_k(t))^T,\, t \in R^1\big\}$ *is a* k*-dimensional Gaussian process with zero mean, with independent components and the covariance structures*

$$\begin{aligned} &\operatorname{cov}\big(Z_j(t_1), Z_j(t_2)\big) \\ &\qquad = p_j \operatorname{cov}\big(\cos(t_1 Y) + \sin(t_1 Y), \cos(t_2 Y) + \sin(t_2 Y)\big), \qquad t_1, t_2 \in R^1, \end{aligned}$$

where Y *has the distribution function* F*. Here* $\overline{Z}(t) = \sum_{j=1}^{k} Z_j(t)$.

Clearly, the limit distribution of $T_n(\beta, k)$ is not asymptotically distribution free even under H_0 (see covariance structure of the process).

Remark 2.1. The proof goes along the line of those in M e i n t a n i s [8], where $k = 2$ is treated in detail. Therefore the proof is omitted. Both theoretical results and simulation study are included there. E p p s and S i n g l e t o n [2] employed Mahalanobis-type distance between the ecf's, whereas A l b a *et al.* [1] used an L_2 distance.

Remark 2.2. The Efron bootstrap with or without replacement can be applied in order to get an approximation for critical values. It provides asymptotically correct approximations only when data follow H_0 or some local alternatives. Generally, the bootstrap version $T_n^*(\beta, k)$ of $T_n(\beta, k)$ has the same limit distribution as $T_n(\beta, k)$ corresponding to the distribution $H(x) = \sum_{j=1}^{k} p_j F_j(x)$, $x \in R^1$. At any case $T_n^*(\beta, k) = O_P(1)$ while under large spectrum of alternatives $T_n(\beta, k) \to^P \infty$. Therefore the approximation of the critical values through the bootstrap leads to the consistent test, i.e., using the bootstrap approximation of the critical values we reject the null hypothesis for fixed alternatives with probability tending to 1, as $n \to \infty$.

Remark 2.3. Under fixed alternatives H_1 the test statistics $T_n(\beta, k)$ are tending to ∞ in probability. Concerning the behavior of $T_n(\beta, k)$ under local alternatives

we restrict ourselves to the case when the null distribution F_0 is symmetric and absolutely continuous with the density f_0 and we consider the following class of alternative hypotheses:

$$H_n : Y_{j,1}, \dots, Y_{j,n_j} \quad \text{are i.i.d. with common density } g_{n,j}$$

with

$$g_{n,j}(x) = \left(1 + \frac{\kappa}{\sqrt{n}} u_j(x)\right) f_0(x), \quad j = 1, \dots, k, \qquad x \in R^1, \tag{2.7}$$

where $\kappa \neq 0$ and u_j, $j = 1, \dots, k$ are measurable functions such that

$$\int u_j(x) f_0(x)\, \mathrm{d}x = 0, \quad 0 < \int u_j^2(x) f_0(x)\, \mathrm{d}x < \infty, \qquad j = 1, \dots, k. \tag{2.8}$$

By the third LeCam's lemma the sequence of distributions with densities $\{\Pi_{j=1}^k \Pi_{s=1}^{n_j} g_{n,j}(y_{js})\}$ is contiguous w.r.t. the sequence of $\{\Pi_{j=1}^k \Pi_{s=1}^{n_j} f_0(y_{js})\}$ and this in combination with the limit behavior under H_0 implies that also under H_n (2.7) Theorem 2.1 remains true with $\boldsymbol{Z}$ replaced by its shifted version.

Theorem 2.2. *Let $Y_{j,1}, \dots, Y_{j,n_j}$, $j = 1, \dots, k$ follow the model* (2.1) *with Y_{js} being i.i.d. with common density g_{jn} defined in* (2.7) *satisfying* (2.8). *Let β satisfy* (2.5). *Then there is a zero-mean Gaussian process $\boldsymbol{Z} = \{\boldsymbol{Z}(t);\ t \in R^1\}$ such that, as $n \to \infty$,*

$$\{\boldsymbol{Z}_n(t);\ t \in R^1\} \to^D \{\boldsymbol{Z}(t) + \kappa \boldsymbol{\mu}(t);\ t \in R^1\}$$

and

$$||T_n(\beta, k)||^2 \to^D \int_{-\infty}^{\infty} \sum_{j=1}^{k} \frac{1}{p_j} \big(Z_j(t) - p_j \overline{Z}(t) - \kappa p_j \mu_j^0(t)\big)^2 \beta(t)\, \mathrm{d}t.$$

The process $\{\boldsymbol{Z}(t);\ t \in R^1\}$ is from Theorem 2.1 *and $\boldsymbol{\mu}(t)$ and $\boldsymbol{\mu}^0(t)$ have the components*

$$\mu_j(t) = \int \big(\cos(tx) + \sin(tx)\big) u_j(x) f_0(x)\, \mathrm{d}x, \qquad j = 1, \dots, k,$$

$$\mu_j^0(t) = \mu_j(t) - p_j \sum_{s=1}^{k} p_s \mu_s(t), \quad j = 1, \dots, k, \qquad t \in R^1.$$

The proof of this theorem will be done in a separate paper.

Since the assertion holds true for any $\kappa \neq 0$ we find that with increasing κ and $||\boldsymbol{\mu}^0||^2 \neq 0$ the nonrandom part of $T_n(\beta, k)$, i.e., $\kappa ||\boldsymbol{\mu}^0||^2$ dominates the random one and approaches to ∞. Therefore our test procedure is consistent as soon as $\kappa \to \infty$.

Remark 2.4. A number of modifications can be introduced, for instance, rank based procedures. Additionally we assume that the random variables $Y_{j,s}$, $j = 1, \ldots, k$, $s = 1, \ldots, n_j$ have continuous distribution functions. Denote $R_{r,q}$ the rank of $Y_{r,q}$ among $Y_{j,s}$, $j = 1, \ldots, k$, $s = 1, \ldots, n_j$ for $r = 1, \ldots, k$, $q = 1, \ldots, n_r$. Then for the testing problem H_0 versus H_1 one can develop along the above line the test procedures based on empirical characteristic functions of these ranks. Particularly, replacing the empirical characteristic functions $\widehat{\phi}_j(.)$, $j = 1, \ldots, k$, in (2.6) by their counterparts based on ranks, i.e., by

$$\widehat{\phi}_j(t, \boldsymbol{R}) = \frac{1}{n_j} \sum_{j=1}^{n_j} \exp\{itR_{js}\}, \quad t \in R^1, \qquad j = 1, \ldots, k \tag{2.9}$$

with $\boldsymbol{R} = (R_{r,q};\, r = 1, \ldots, k,\, q = 1, \ldots, n_r)$ we get $T_n(\beta, k, \boldsymbol{R})$. Then the limit distribution of $T_n(\beta, k, \boldsymbol{R})$ under H_0 is the same as that of $T_n(\beta, k)$ when the observations $Y_{j,s}$, $j = 1, \ldots, k$, $s = 1, \ldots, n_j$ are i.i.d. with $(0, 1)$- uniform distribution. This test procedure is distribution free under H_0. Nevertheless, the explicit form of the limit distribution is unknown. However, it can be easily simulated.

Remark 2.5. Next we discuss relation to U-statistics, for simplicity we focus on $k = 2$. Replacing empirical characteristic functions in $T_n(\beta, 2)$ by theoretical ones and n_j/n by $q_j, j = 1, 2$ we get its theoretical counterpart:

$$T(\beta, \phi_1, \phi_2) = \int q_1 q_2^2 |\phi_1(t) - \phi_2(t)|^2 \beta(t)\, \mathrm{d}t,$$

where $\phi_j(.)$ is the characteristics function of the jth sample, $j = 1, 2$. This can be further rewritten as a functional of the distribution functions F_j, $j = 1, 2$:

$$T(\beta, \phi_1, \phi_2) = q_1 q_2^2 \iiiint h_\beta(x_1, x_2; y_1, y_2) \mathrm{d}F_1(x_1)\, \mathrm{d}F_1(x_2)\, \mathrm{d}F_2(y_1)\, \mathrm{d}F_2(y_2),$$

where

$$\begin{aligned} h_\beta(x_1, x_2; y_1, y_2) = {} & h_\beta(x_1 - x_2) + h_\beta(y_1 - y_2) \\ & - \frac{1}{2}\big(h_\beta(x_1 - y_1) \\ & + h_\beta(x_1 - y_2) + h_\beta(x_2 - y - 1) + h_\beta(x_2 - y_2)\big), \end{aligned}$$

$$h_\beta(z) = \int \cos(tz)\beta(t)\, \mathrm{d}t. \tag{2.10}$$

This functional can be estimated by U-statistics

$$U_{n_1,n_2}(\beta) = \frac{1}{n_1(n_1-1)}\frac{1}{n_2(n_2-1)}\sum_{i_1=1}^{n_1}\sum_{i_2=1,i_2\neq i_1}^{n_1}\sum_{j_1=1}^{n_2}\sum_{j_2=1,i_2\neq j_1}^{n_2} h_\beta(X_{i_1},X_{i_2};Y_{j_1},Y_{j_2}).$$

This is a two-sample U-statistic with degenerate kernel under H_0. This can be also used as the test statistics for our problem. Limit behavior is a little bit simple but again a bootstrap has to be used in order to get approximation for critical values. See , e.g., L e e (1990) and M a d u r k a y o v á [7] for more information.

3. Change point problem

This section concerns test procedures for detection of changes based on empirical characteristic functions. We assume that $Y_1,\dots,Y_n$ are independent random variables, Y_j has a distribution function $F_j, j=1,\dots,n$ and we consider the testing problem

$$H_0 : F_1 = \dots = F_n \tag{3.1}$$

against

$$H_1 : F_1 = \dots = F_m \neq F_{m+1} = \dots = F_n \qquad \text{for} \quad m<n, \tag{3.2}$$

where m, F_1 and F_n are unknown. Motivated by the two-sample tests based on empirical characteristic functions H u š k o v á a n d M e i n t a n i s [5] have introduced the following class of test statistics:

$$T_{n,\gamma}(\beta) = \max_{1\le k<n}\left(\frac{k(n-k)}{n^2}\right)^\gamma \frac{k(n-k)}{n}\int_{-\infty}^{\infty}|\widehat{\phi}_k(t)-\widehat{\phi}_k^0(t)|^2\beta(t)\,\mathrm{d}t, \tag{3.3}$$

where $\beta(\cdot)$ is a nonnegative weight function satisfying (2.5), $\widehat{\phi}_k(.)$ and $\widehat{\phi}_k^0(.)$ are empirical characteristic functions based on $Y_1,\dots,Y_k$ and $Y_{k+1},\dots,Y_n$, respectively, i.e.,

$$\widehat{\phi}_k(t)=\frac{1}{k}\sum_{j=1}^{k}\exp\{itY_j\}, \quad \widehat{\phi}_k^0(t)=\frac{1}{n-k}\sum_{j=k+1}^{n}\exp\{itY_j\}, \qquad k=1,\dots,n, \tag{3.4}$$

and γ is a nonnegative constant.

Concerning the choice of the tuning parameter γ, it would be natural, in accordance with other test procedures for detection of changes, to choose $\gamma=0$, then our test statistic is maximum of the standardized test statistics for two sample problem, where the observations are split into two groups with k and

$n - k$ observations. However, the disadvantage of this test statistic is that it tends to infinity even under the null hypothesis. More precisely, under H_0 and mild assumptions on w, as $n \to \infty$,

$$T_{n,\gamma}(\beta) \to \infty, \qquad \gamma = 0$$

in probability, while

$$T_{n,\gamma}(\beta) = O_P(1), \qquad \gamma > 0.$$

Large values of the tests statistics indicate that the null hypothesis is violated.

Next, we present assertion on the limit behavior of $T_{n,\gamma}(\beta)$ under H_0. It can be formulated similarly as for the k-sample problem through functionals of Gaussian processes, but here we present it through a functional of Brownian bridges. Toward this we need following notation. Put

$$\begin{aligned} \tilde{h}_\beta(x,y) = h_\beta(x-y) &- E_{H_0} h_\beta(x - Y_s) \\ &- E_{H_0} h_\beta(Y_r - y) \\ &- E_{H_0} h_\beta(Y_r, Y_s), \qquad r \neq s, \end{aligned} \tag{3.5}$$

where $h_\beta(\cdot)$ is defined in (2.10). Since the properties of the function $\tilde{h}_\beta(x,y)$ and since $E\,\tilde{h}_\beta^2(Y_1,Y_2) = \int \tilde{h}_\beta^2(x,y)\,\mathrm{d}F(x)\,\mathrm{d}F(y) < \infty$ there exist orthogonal eigenfunctions $\{\psi_j(t),\, j = 1,2,\ldots\}$ and eigenvalues $\{\lambda_j,\, j = 1,2,\ldots\}$ such that (see, e.g., S e r f l i n g [9])

$$\lim_{K\to\infty} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\widetilde{h}_\beta(x,y) - \sum_{j=1}^{K} \lambda_j \psi_j(x)\psi_j(y) \right)^2 \mathrm{d}F(x)\,\mathrm{d}F(y) = 0, \tag{3.6}$$

$$\int_{-\infty}^{\infty} \psi_j^2(x)\,\mathrm{d}F(x) = 1, \qquad j = 1,2,\ldots, \tag{3.7}$$

$$\int_{-\infty}^{\infty} \psi_j(x)\psi_i(x)\,\mathrm{d}F(x) = 0, \qquad i \neq j = 1,2,\ldots \tag{3.8}$$

and

$$E\,\widetilde{h}_\beta^2(Y_1,Y_2) = \int \widetilde{h}_\beta^2(x,y)\,\mathrm{d}F(x)\,\mathrm{d}F(y) = \sum_{j=1}^{\infty} \lambda_j^2. \tag{3.9}$$

Here is the assertions on the limit behavior of the test statistic $T_{n,\gamma}(\beta)$ under H_0.

Theorem 3.1. *Let $Y_1, \dots, Y_n$ be i.i.d. random variables with common distribution function F. Let β satisfy* (5). *Then* (i) *for $\gamma > 0$ the limit behavior of $T_{n,\gamma}(\beta)$ is the same as that of*

$$\sup_{0<z<1} \big(z(1-z)\big)^{\gamma} \left| \left(\int \beta(t)\, \mathrm{d}t - E\, h_\beta(Y_1, Y_2) \right) + \sum_{j=1}^{\infty} \lambda_j \left\{ \frac{B_j^2(z)}{z(1-z)} - 1 \right\} \right|, \quad (3.10)$$

where $\big\{B_j(t),\, t \in (0,1)\big\}$, $j = 1, 2 \dots$, are independent Brownian bridges, (ii) *for $\gamma = 0$, as $n \to \infty$*

$$T_{n,0}(\beta)(\log\log n)^{-1} = O_P(1), \qquad T_{n,0}(\beta) \to^P \infty.$$

Remark 3.1. The explicit distribution of (3.10) is unknown. By properties of Brownian bridges this random variable (3.10) is $O_P(1)$.

Remark 3.2. Since the eigenvalues $\{\lambda_j\}$ and eigenfunctions $\{\psi_j\}_j$ depend on the underlying distribution function F which is unknown, the limit distribution of (3.10) depends on the unknown parameters and unknown functions so that the limit distribution does not provide a useful approximation for the critical values.

Remark 3.3. These procedures, as well as related procedures based on empirical characteristic function of ranks, were developed and studied in Hušková and Meintanis [5] and [6]. Bayesian like type procedures were studied in Hušková and Meintanis [4].

Remark 3.4. Similarly as in the previous section, bootstrap with and without replacement provide reasonable approximations for critical values. Such approximations are asymptotically valid when the data follow the null hypothesis or local alternatives. Particularly, bootstrap without replacement leads to a test with level α, while bootstrap without replacement leads to tests with asymptotic level α. In case of the so called fixed alternatives resampling methods do not lead to an asymptotically valid approximation to the critical values, however the resulting tests are consistent.

REFERENCES

[1] ALBA, M. V.—BARRERA, D.—JIMÉNEZ, M. D.: *SA homogeneity test based on empirical characteristic functions*, Comput. Statist. **16** (2001), 255–270.

[2] EPPS, T. W.—SINGLETON, K. J.: *An omnibus test for the two-sample problem using the empirical characteristic functions*, J. Stat. Comput. Simul. **26** (2002), 177–203.

[3] HUŠKOVÁ, M.: *Permutation principle and bootstrap in change point analysis.* In: Asymptotic methods in stochastics. Festschrift for Miklós Csörgö. Proceedings of the international conference, held in honour of the work of Miklós Csörgö on the occasion of his 70th birthday, Ottawa, Canada, May 23–25, 2002, Fields Institute Communications, Vol. 44, AMS, Providence, RI, 2004, pp. 273–291.

[4] HUŠKOVÁ, M.—MEINTANIS, S.: *Bayesian like procedures for detection of changes.* In: Proceedings of COMPSTAT '04 (Antoch, J. ed.), Physica-Verlag, Heidelberg, 2004, pp. 1221–1228.

[5] HUŠKOVÁ, M.—MEINTANIS, S.: *Change point analysis based on empirical characteristic functions.* Metrika **63**, (2006), 145–168.

[6] HUŠKOVÁ, M.—MEINTANIS, S.: *Change-point analysis based on empirical characteristic functions of ranks.* Sequential Analysis **25**, (2006), 421–436.

[7] MADURKAYOVÁ, B.: *Tests Based on U-Statistics.* Master Thesis, Charles University in Prague, Czech Republic, 2006.

[8] MEINTANIS, S.: *Permutation tests for homogeneity based on the empirical characteristic function.* J. Nonparametr. Stat. **17**, (2005), 583–592.

[9] SERFLING, R.: *Approximation Theorems of Mathematical Statistics.* J. Wiley, New York, 1980.

Received September 29, 2006

Marie Hušková
Department of Statistics
Charles University of Prague
Sokolovská 83
CZ–186 75 Prague 8
CZECH REPUBLIC

Institute of Information Theory
and Automation
Czech Academy of Sciences
Pod Vodárenskou věží 4
CZ–182 08 Prague 8
CZECH REPUBLIC
E-mail: marie.huskova@karlin.mff.cuni.cz

Simos G. Meintanis
Department of Economics
National and Kapodistrian University of Athens
8 Pesmazoglou Street
105 59 Athens
GREECE
E-mail: simosmei@econ.uoa.gr