AUSTRIAN JOURNAL OF STATISTICS
Volume 37 (2008), Number 1, 101–108

A Simple Method for Testing Independence of High-Dimensional Random Vectors

Gintautas Jakimauskas, Marijus Radavičius and Jurgis Sušinskas
Institute of Mathematics and Informatics, Vilnius

Abstract: A simple, data-driven and computationally efficient procedure for testing independence of high-dimensional random vectors is proposed. The procedure is based on interpretation of testing goodness-of-fit as the classification problem, a special sequential partition procedure, elements of sequential testing, resampling and randomization. Monte Carlo simulations are carried out to assess the performance of the procedure.

Keywords: Nonparametric Test, Sequential Testing, Classification, Bootstrap, Monte Carlo Simulation, Data-Driven Partition, Dyadic Splitting.

1 Introduction

Let $\mathbf{X} := (X(1), \dots, X(N))$ be a sample of the size N of i.i.d. observations of a random vector X having distribution function (d.f.) F on $\mathbf{R}^d$. We are interested in testing some properties of F. Let $\mathcal{F}_H$ and $\mathcal{F}_A$ be two disjoint classes of d-dimensional distributions. Consider a nonparametric hypothesis testing problem:

$$H: \ F \in \mathcal{F}_H \qquad \text{versus} \qquad A: \ F \in \mathcal{F}_A. \tag{1}$$

Testing the independence of two components $X_1 \in \mathbf{R}^{d_1}$ and $X_2 \in \mathbf{R}^{d_2}$, $d_1 + d_2 = d$, of $X = (X_1', X_2')'$ corresponds to

$$\mathcal{F}_H = \{G : G(x) = G_1(x_1) \cdot G_2(x_2), \ x = (x_1', x_2')', \ x_1 \in \mathbf{R}^{d_1}, x_2 \in \mathbf{R}^{d_2}\}, \tag{2}$$

where G_1 and G_2 denote the marginal distributions of G corresponding to the components X_1 and X_2, respectively.

Our goal is to propose a relatively simple, data-driven and computationally efficient procedure for testing problem (1), with key example (2), in case the dimension d of X is *large*. The procedure is based on Vapnik and Chervonenkis (1981) idea of bounding a discrepancy between empirical and true distribution by that of two independent empirical distributions (Vapnik and Chervonenkis, 1981) and a well-known interpretation of testing goodness-of-fit as the classification problem (see, e.g. Hastie, Tibshirani, and Friedman, 2001, pp. 447-449), a special sequential data partition procedure, randomization and resampling (bootstrap), elements of sequential testing. Monte Carlo simulations are used to assess the performance of the procedure.

Thus far, there is no generally accepted methodology for the multivariate nonparametric hypothesis testing. Traditional approaches to multivariate nonparametric hypothesis testing are based on empirical characteristic function Baringhaus and Henze (1988), nonparametric distribution density estimators and smoothing Bowman and Foster (1993), Huang (1997), and classical univariate nonparametric statistics calculated for data projected onto the directions found via the projection pursuit L. Zhu, Fang, and Bhatti (1997), Szekely and Rizzo (2005).

More advanced technique is based on Vapnik-Chervonenkis theory, the uniform functional central limit theorem and inequalities for large deviation probabilities Vapnik (1998), Bousquet, Boucheron, and Lugosi (2004). Recently, especially in applications, the Bayes approach and Markov chain Monte Carlo methods are widely used (see, e.g. Verdinelli and Wasserman, 1998 and references therein). Multidimensional copulas are a convenient way to represent the statistical dependence between components of random vectors. Therefore asymptotic behavior and power of independence testing criteria based on empirical copula processes are extensively studied (see, e.g. Genest and Remillard, 2004). However, these results are not directly applicable in our setting since the components X_1 and X_2 themselves have a large dimensionality.

To identify dependence-independence structure of high-dimensional data the independent component analysis (ICA), a recent extension of principal component analysis and projection, is employed. We refer to the monograph by Hyvärinen, Karhunen, and Oja (2001). An efficient method for testing of (conditional) independence is essential here. Related references to our approach are Szekely and Rizzo (2006), Polonik (1999), L.-X. Zhu and Neuhaus (2000).

In Section 2 the procedure of nonparametric hypothesis testing is introduced. The Monte Carlo simulation results and concluding remarks are presented in the last section.

2 Statistical Test

2.1 Test Statistic

Let $\mathcal{F} := \mathcal{F}_H \bigcup \mathcal{F}_A$. Suppose that the mapping $\Psi : \mathcal{F} \to \mathcal{F}_H$ is such that $\mathcal{F}_H = \{G \in \mathcal{F} : \Psi(G) = G\}$. Given $F \in \mathcal{F}$, denote $F_H = \Psi(F)$. For the independence hypothesis $F_H = F_1 \cdot F_2$.

Consider a mixture model

$$F_{(p)} := (1-p)F_H + pF, \qquad p \in (0,1),$$

of two populations Ω_H and Ω with d.f. F_H and F, respectively. Fix p and let $Y = Y_{(p)} \sim F_{(p)}$ denote a random vector (r.v.) with the mixture distribution $F_{(p)}$. Let $\pi(Y)$ denote the posterior probability of the population Ω given Y, i.e.

$$\pi(Y) := \mathbf{P}\big[\Omega|Y\big] = \frac{pf(Y)}{pf(Y) + (1-p)f_H(Y)}.$$

Here f and f_H denote distribution densities (with respect to a σ-finite measure μ) of F and F_H, respectively.

Let us introduce a *loss function* $\ell(F, F_0) := \mathrm{E}(\pi(Y) - p)^2$. It is clear that

$$\ell(F, F_H) = 0 \quad \text{if and only if} \quad F = F_H,$$

since the posterior probability $\pi(Y)$ is equal to the prior probability p if and only if $F = F_H$.

Let $\mathbf{X}^{(H)} := (X^{(H)}(1), \dots, X^{(H)}(M))$ be a sample of size M of i.i.d. random vectors from Ω_H independent of $\mathbf{X}$. The joint sample is denoted by

$$\mathbf{Y} := \mathbf{X} \mid\mid \mathbf{X}^{(H)} = (X(1), \dots, X(N), X^{(H)}(1), \dots, X^{(H)}(M))$$

and $Z(t) = \mathbf{1}\{t \le N\}$, $t = 1, \dots, N+M$, is the corresponding sequence of indicators of the population Ω. Let $\mathcal{P} := \{P_k,\ k = 0, 1, \dots, K\}$, $P_0 := \{\mathbf{R}^d\}$, $P_{k-1} \subset P_k,\ k = 1, \dots, K$, be a sequence of partitions of $\mathbf{R}^d$, possibly dependent on $\mathbf{Y}$, and let $\{\mathcal{A}_k,\ k = 0, 1, \dots, K\}$ be the corresponding sequence of σ-algebras generated by these partitions.
Remark 1. A computationally efficient choice of $\mathcal{P}$ is the sequential dyadic coordinatewise partition minimizing at each step the mean square error with some restrictions (bounds from below) on the number of the sample $\mathbf{Y}$ elements in the partition sets. An alternative might be a partition into sets with the approximately equal number of the sample $\mathbf{Y}$ elements.

In view of the definition of the loss function $\ell(F, F_0)$ a natural choice of the test statistics would be χ^2-type statistics

$$T_k := \widehat{\mathrm{E}}(Z_k - p)^2, \qquad p := \frac{N}{N+M}, \tag{3}$$

where $\widehat{\mathrm{E}}$ stands for the expectation with respect to the empirical distribution $\hat{F}$ of $\mathbf{Y}$ and

$$Z_k := \widehat{\mathrm{E}}[Z|\mathcal{A}_k]$$

for some $k \in \{1, \dots, K\}$. The integer k can be treated as a *"smoothing" parameter*. It characterizes how small is the partition. We also consider a weighted version of (3)

$$T_k := \widehat{\mathrm{E}}\big((Z_k - p)^2 W_k\big), \tag{4}$$

where W_k is some $\mathcal{A}_k$-measurable weight function. The choice $W_k = |S \bigcap \mathbf{Y}|/(p(1-p))$ on the partition set $S \in P_k$, yields the L_2 distance between the observed and the expected frequencies for the true hypothesis H.

Since the optimal value of k is unknown, we prefer the following definition of the *test statistic*

$$T := \max_{k_0 \le k \le K} (T_k - a_k)/b_k, \tag{5}$$

where $k_0 \ge 1$, a_k and b_k are centering and scaling parameters, respectively, to be specified.
Remark 2. Since the critical region of the criterion is of the form $\mathcal{C}_\alpha := \{T > c_\alpha\}$, where c_α is the critical value corresponding to the significance level α, it is natural to express $\mathcal{C}_\alpha$ as the *sequential testing procedure*:

Step 1: Set $k = k_0 - 1$.

Step 2: $k + 1 \to k$; if $k > K$, then STOP, otherwise calculate T_k.

Step 3: If $T_k > a_k + c_\alpha b_k$, reject H_0 and STOP, otherwise go to Step 2.

2.2 The Null Distribution of the Test Statistic

Let $\tau : I \to I$ be a random permutation of $I := \{1, \dots, N + M\}$ with equal probabilities and $\mathbf{Y}^\tau$ denote the corresponding permutation of $\mathbf{Y}$. For any statistic ξ, let ξ^τ indicate that this statistic is calculated for the randomized sample $\mathbf{Y}^\tau$. In particular, $\mathbf{X}^\tau = (Y^\tau(1), \dots, Y^\tau(N))$.

Under the hypothesis H, $\mathbf{Y}^\tau = \mathbf{Y}$ in distribution. Therefore one can deal with the conditional distribution of the randomized test statistic T_k^τ given the sample $\mathbf{Y}$ in order to assess the properties of the initial test statistic T_k.

Fix the sample $\mathbf{Y}$. For the partition $P_k = \{S_{k,1}, \dots, S_{k,J_k}\}$, let

$$\begin{aligned} n(k) &= (n_1(k), \dots, n_{J_k}(k)) = (|S_{k,j} \bigcap \mathbf{Y}|,\ j = 1, \dots, J_k), \\ \nu(k) &= (\nu_1(k), \dots, \nu_{J_k}(k)) = (|S_{k,j} \bigcap \mathbf{X}|,\ j = 1, \dots, J_k), \quad k = 1, \dots, K. \end{aligned}$$

be the J_k-dimensional vectors of the observed in P_k frequencies of the elements of $\mathbf{Y}$ and $\mathbf{X}$, respectively. Then $Z_k^\tau = \nu_j^\tau(k)/n_j(k)$ on the partition set $S_{k,j}$. Since the discrete random vector $\nu^\tau(k)$ has the *multivariate hypergeometric* distribution with the parameters $N + M$, $n(k)$, and N, the conditional distribution of T^τ, given $\mathbf{Y}$ and the partition $\mathcal{P}$, depends on $\mathbf{Y}$ only through the *"sizes" of the partition sets*, $n(k)$, $k = 1, \dots, K$. This provides a basis for determining a_k and b_k in (5), the analysis of asymptotic distribution of the statistic T, and exponential inequalities for probabilities of large deviations of T. In this study, however, we prefer to perform a simulation experiment.

3 Testing the Independence: A Simulation Experiment

To generate a sample from $F_H = F_1 \cdot F_2$ we apply *bootstrap* method and resample from the distribution $\hat{F}_H := \Psi(\hat{F}) = \hat{F}_1 \cdot \hat{F}_2$ where $\hat{F}_i$ denotes the empirical distribution of F_i, $i = 1, 2$.

Let $\mathbf{X}$ be the repeated independent observations of X having standard multivariate Student distribution with m degrees of freedom. Although the components of X are uncorrelated they are dependent. Since X converges in distribution to a standard normal random vector as $m \to \infty$, the dependence of the components vanishes for large m.

The centering and scaling parameters for the statistics T_k are calculated using approximations by the normal distribution. In what follows it is assumed that $M = N$ and $J_k \equiv k + 1$. Thus, the standardized versions $\hat{T}_k$ and $\hat{T}_k^{(2)}$ of χ^2-type statistic (3) and L_2 distance defined by (4) with the weight function $W_k = |S \bigcap \mathbf{Y}|$, respectively, are given by

$$\hat{T}_k := \frac{T_k - k}{\sqrt{2k}}, \quad T_k = \sum_{j=0}^{k} \frac{(n_j(k) - 2\nu_j(k))^2}{n_j(k)}, \tag{6}$$

and

$$\hat{T}_k^{(2)} := \frac{T_k^{(2)} - 2N}{2\sqrt{N}}, \quad T_k^{(2)} = \sum_{j=0}^{k} (n_j(k) - 2\nu_j(k))^2. \tag{7}$$

In the sequel, the test statistic T (5) based on (6) is considered. As compared with (7) it places greater weights on the partition sets with smaller number of the sample elements. Our experience suggests that $k_0 = 10$ and $K = (M + N)/5$ is an appropriate choice for the maximization interval $[k_0, K]$ of T_k. The critical value c_α for the test is to be chosen in such a way that a portion of samples for which the valid null hypothesis is rejected does not exceed a given significance level α, say $\alpha = 0.05$. Monte Carlo method is used to find c_α. Preliminary results of Monte Carlo simulations show that for a wide range of dimensions, sample sizes and null distributions the behavior of the test statistics for samples from the null distribution (control data) is quite similar (see Figure 1 and Figure 2). In particular, we have also applied the procedure to testing goodness-of-fit for a mixture of multivariate Gaussian distributions. The choice $c_{0.05} = 2.7$ is admissible for most cases.

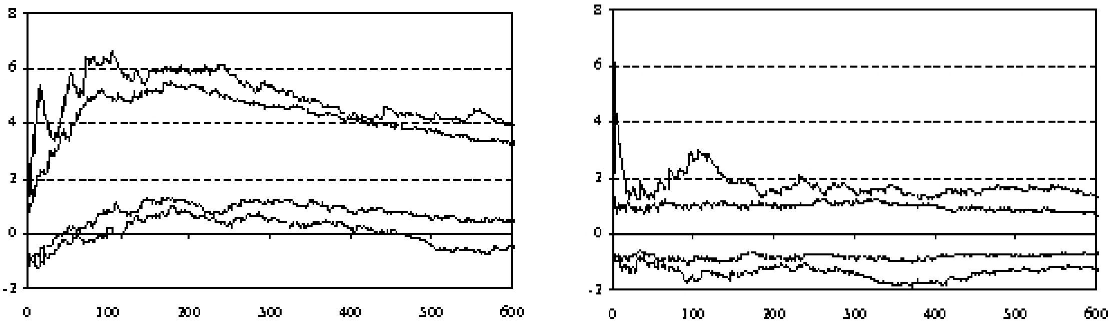

Figure 1: Maximum, minimum and two-side 0.9 confidence limits of T_k for a sample from the Cauchy distribution ($m = 1$) and for the corresponding control data; $d = 20$, $d_1 = d_2 = 10$, $N = 1000$.

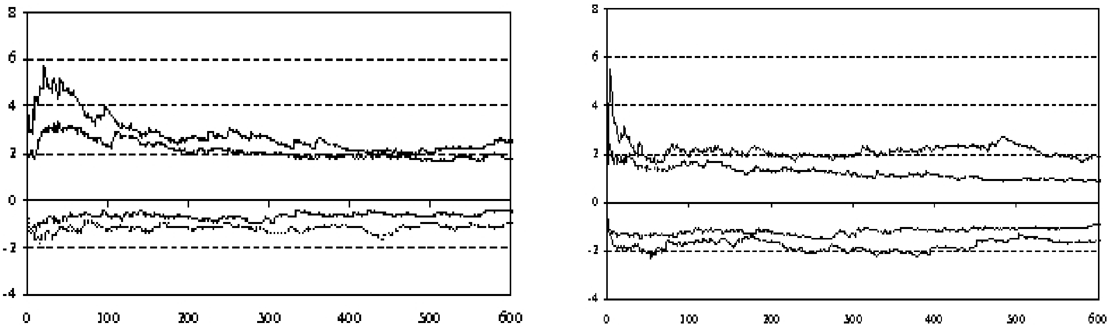

Figure 2: Maximum, minimum and two-side 0.9 confidence limits of T_k for a sample from the Student distribution with d.f. $m = 3$ and for the corresponding control data; $d = 10$, $d_1 = 1$, $d_2 = 9$, $N = 1000$.

The computer simulations are performed for $d \leq 20$, $200 \leq N \leq M \leq 1000$, and $m = 1, \ldots, 7, 25, 100, \infty$. The dimensions d_1 and d_2 of the independent components X_1 and X_2, respectively, are chosen in two ways. In the first case $d_1 = d_2 = d/2$, and in the second case $d_1 = 1$, $d_2 = d - 1$. The typical number of simulations $R = 1000$. Below only results for the $d = 2, 10$ and $N = M = 1000$ are presented.

In the sequel, the test procedure based on T (5) is referred to as JRS test for brevity. The performance of the procedure is compared with the classical criterion of Blum, Kiefer, and Rozenblatt (1961) (BKR test) based on Cramér-Von Mises-type test statistics for testing independence:

$$\omega^2_{BKR} = N \int_{\mathbf{R}^{d_1}} \int_{\mathbf{R}^{d_2}} \left(\hat{F}(u,v) - \hat{F}_1(u)\hat{F}_2(v)\right)^2 d\hat{F}(u,v). \tag{8}$$

Here $\hat{F}_i$ is the empirical distribution function of the component X_i based on the sample $\mathbf{X}$ ($i = 1, 2$).

The power of the JRS test is compared with that of the BKR test. To evaluate the power functions of the independence tests Monte Carlo simulations with $R = 1000$ realizations have been performed. The results are presented in Figures 3 and 4 and Table 1 for the significance levels $\alpha = 0.02, 0.05, 0.1$ and dimensions $d = 2$ and $d = 10$ with $d_1 = d_2 = d/2$. The power of the JRS test slightly decreases for growing dimension d, and for $d = 10$ it is close to the power of the BKR test for $d = 2$. The power of the BKR test for $d = 10$ is very low.

The computational complexity of the BKR (respectively, JRS) test is $O(d \cdot N^2)$ (respectively, $O\big(d^2 \cdot (N+M)\log(N+M)\big)$).

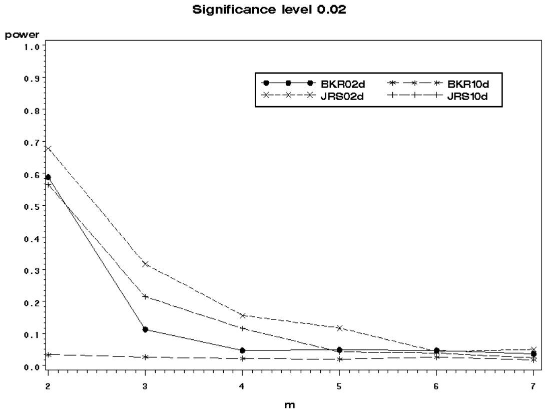

Figure 3: Power functions of the BKR test for the dimensions $d = 2$ ('BKR02d') and $d = 10$ ('BKR10d') and the corresponding power functions for the JRS test ('JRS02d' and 'JRS10d'); the significance level $\alpha = 0.02$.

4 Concluding Remarks

Preliminary results of Monte Carlo simulations show that the procedure proposed is promising. It outperforms the classical BKR test even for low-dimensional data. The dependence of the critical value c_α on the dimensionality d and the partition procedure is weak and can be reduced by imposing appropriate additional requirements on it.

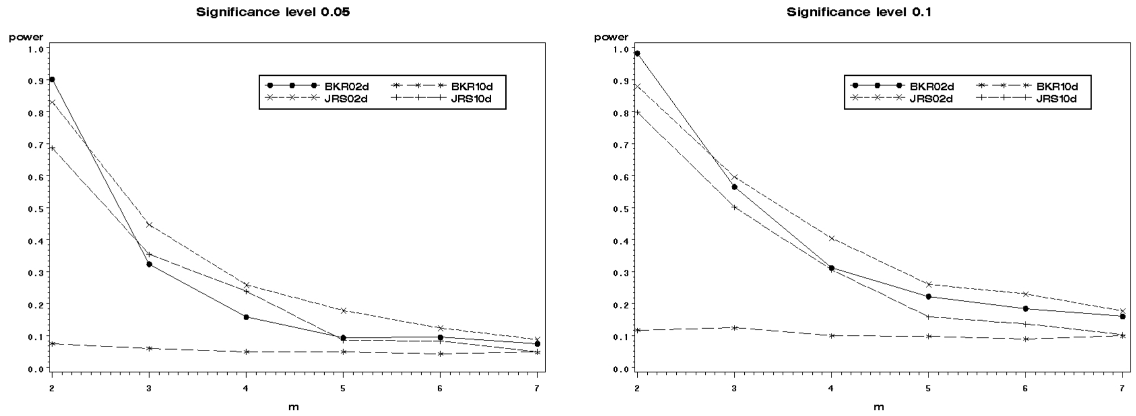

Figure 4: Power functions of the BKR test for the dimensions $d = 2$ ('BKR02d') and $d = 10$ ('BKR10d') and the corresponding power functions for the JRS test ('JRS02d' and 'JRS10d'); the significance level $\alpha = 0.05$ (left) and $\alpha = 0.1$ (right).

Table 1: The power of the tests BKR and JRS.

Dimensions $d_1 = d_2 = d/2$	Degrees of freedom (m)					
Dimension $d = 2$	2	3	4	5	6	7
BKR, $\alpha = 0.1$	98.2	56.5	31.2	22.2	18.4	15.9
JRS, $\alpha = 0.1$	87.8	59.5	40.3	26.0	23.0	17.6
BKR, $\alpha = 0.05$	90.0	32.3	15.7	9.3	9.4	7.4
JRS, $\alpha = 0.05$	83.0	44.6	25.8	17.8	12.3	8.7
BKR, $\alpha = 0.02$	58.8	11.2	4.7	4.9	4.7	3.6
JRS, $\alpha = 0.02$	67.8	31.6	15.6	11.7	4.3	5.0
Dimension $d = 10$	2	3	4	5	6	7
BKR, $\alpha = 0.1$	11.6	12.4	9.9	9.8	8.9	9.8
JRS, $\alpha = 0.1$	79.8	50.1	30.5	15.8	13.6	10.2
BKR, $\alpha = 0.05$	7.4	5.9	4.9	4.9	4.3	4.8
JRS, $\alpha = 0.05$	68.6	35.3	23.8	8.4	8.3	4.9
BKR, $\alpha = 0.02$	3.3	2.5	2.1	2.0	2.5	1.7
JRS, $\alpha = 0.02$	56.4	21.4	11.5	4.2	3.9	2.3

References

Baringhaus, L., and Henze, N. (1988). A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, *35*, 339-348.

Blum, J. R., Kiefer, J., and Rozenblatt, M. (1961). Distribution free tests for independence based on the sample distribution function. *Annals of Mathematical Statistics*, *35*, 138-149.

Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to statistical Learning Theory. In O. Bousquet, U. von Luxburg, and G. Rätsch (Eds.), *Advanced Lectures on Machine Learning* (p. 169-207). New York, Berlin: Springer.

Bowman, A. W., and Foster, P. J. (1993). Adaptive smoothing and density based tests of multivariate normality. *Journal of the American Statistical Association*, *88*, 529-537.

Genest, C., and Remillard, B. (2004). Tests of independence and randomness based on the empirical copula process. *Test*, *13*, 335-370.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The elements of Statistical Learning*. New York, Berlin: Springer.

Huang, L.-S. (1997). Testing goodness-of-fit based on a roughness measure. *Journal of the American Statistical Association*, *92*, 1399-1402.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. New York: John Wiley and Sons.

Polonik, W. (1999). Concentration and goodness-of-fit in higher dimensions: (asymptotically) distribrution free methods. *Annals of Statistics*, *27*, 1210-1229.

Szekely, G. J., and Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, *93*, 58-80.

Szekely, G. J., and Rizzo, M. L. (2006). *Testing for equal distributions in high dimension.*

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: John Wiley and Sons.

Vapnik, V. N., and Chervonenkis, A. (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory Probabability and its Applications*, *26*, 821-832.

Verdinelli, I., and Wasserman, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Annals of Statistics*, *26*, 1215-1241.

Zhu, L., Fang, K. T., and Bhatti, M. I. (1997). On estimated projection pursuit-type Cramér-von Mises statistics. *Journal of Multivariate Analysis*, *63*, 1-14.

Zhu, L.-X., and Neuhaus, G. (2000). Nonparametric Monte Carlo tests for multivariate distributions. *Biometrika*, *87*, 919-928.

Corresponding Authors' Address:

Marijus Radavičius
Probability Theory and Statistics Department
Institute of Mathematics and Informatics
Akademijos str. 4
LT-08663 Vilnius
Lithuania

E-mail: `mrad@ktl.mii.lt`