TESTING OF INDEPENDENCY FOR HIGH-DIMENSIONAL DATA

M. Radavičius, G.Jakimauskas, J.Sušinskas
Institute of Mathematics and Informatics
Vilnius, LITHUANIA
e-mail: mrad@ktl.mii.lt

Abstract

A simple, data-driven and computationally efficient procedure for testing independence of variables in high-dimension data 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 used to assess the performance of the procedure.

1 Introduction

Let $\mathbf{X} := \{X(1), \dots, X(N)\}$ be a sample of the size N of i.i.d. random vectors with a common 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:

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

Testing the independency of two components $X_1 \in \mathbf{R}^{d_1}$ and $X_2 \in \mathbf{R}^{d_2}$, $d_1 + d_2 = d$, of $X = (X_1', X_2')' \sim F$ 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 of the components X_1 and X_2, respectively.

Our goal is to propose a 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 well-known interpretation of testing goodness-of-fit as the classification problem, a special sequential data partition procedure, randomization and resampling, 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 [1], nonparametric distribution density estimators and smoothing [3, 5], and classical univariate nonparametric statistics calculated for data projected onto the directions found via the projection pursuit [13, 8].

More advanced technique is based on Vapnik-Chervonenkis theory, the uniform functional central limit theorem and inequalities for large deviation probabilities [10, 2]. Recently, especially in applications, the Bayes approach and Markov chain Monte Carlo methods are widely used (see, e.g. [11] 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, for instance, [4]). However, these results are not directly applicable in our setting since the components X_1 and X_2 themselves have 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 monograph by Hyvrinen et al. [6]. An efficient method for testing of (conditional) independence is essential here. Related references to our approach are [9, 7, 12].

2 Test criterion

Test statistic. Let $\mathcal{F} := \mathcal{F}_H \bigcup \mathcal{F}_A$. Suppose that the mapping $\Psi\colon\ \mathcal{F} \to \mathcal{F}_H$ is such that $\mathcal{F}_H = \{G \in \mathcal{F}\colon \Psi(G) = G\}$. 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

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

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 $Z = Z_{(p)}$ be the posterior probability of the population Ω given Y, i.e.

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

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* $\ell(F, F_0) := \mathbf{E}(Z-p)^2$. It is clear that $\ell(F, F_H) = 0 \ \Leftrightarrow\ F = F_H$, since the posterior probability Z is equal to the prior probability p if and only if $F = F_H$.

Let $\mathbf{X}^{(H)} := \{X^{(H)}(1), \ldots, X^{(H)}(M)\}$ be a sample of size M of i.i.d. random vectors from Ω_H. It is also supposed that $\mathbf{X}^{(H)}$ is independent of $\mathbf{X}$. Set

$$\mathbf{Y} := \mathbf{X} \bigcup \mathbf{X}^{(H)}.$$

Let $\mathcal{P} := \{P_k,\ k = 0, 1, \ldots, K\}$, $P_0 := \{\mathbf{R}^d\}$, $P_{k-1} \subset P_k$, $k = 0, 1, \ldots, K$, be a sequence of partitions of $\mathbf{R}^d$, possibly dependent on $\mathbf{Y}$, and let $\{\mathcal{F}_k,\ k = 0, 1, \ldots, K\}$ 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.

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

$$T_k := \mathbf{E}_{MN}\big(Z_k - p\big)^2, \quad \text{where} \quad Z_k := \mathbf{E}_{MN}[Z|\mathcal{F}_k],\ p := \frac{N}{N+M},$$

for some $k \in \{1, \ldots, K\}$ which can be treated as a *"smoothing" parameter*. It characterizes how small is the partition. Here $\mathbf{E}_{MN}$ stands for the expectation with respect to the empirical distribution $\hat{F}$ of $\mathbf{Y}$. However, since the optimal value of k is unknown, we prefer the following definition of the *test statistic*

$$T := \max_{1 \le k \le K} (T_k - a_k)/b_k, \tag{3}$$

where 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 $k = 0$; (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).

Null distribution of the test statistic. Let τ be a random permutation of $\{1, \ldots, N+M\}$ with equal probabilities and $\mathbf{Y}^\tau$ denote the corresponding permutation of $\mathbf{Y}$. Under the hypothesis H, $\mathbf{Y}^\tau = \mathbf{Y}$ in distribution. Therefore conditional distribution of $\mathbf{Y}^\tau$ given $\mathbf{Y}$ is determined by τ distribution.

Fix the sample $\mathbf{Y}$. For any statistic ξ, let ξ^τ indicate that this statistic is calculated for the randomized sample $\mathbf{Y}^\tau$. In particular, $\mathbf{X}^\tau = \{Y^\tau(1), \ldots, Y^\tau(N)\}$.

Let $P_k = \{S_{k,1}, \ldots, S_{k,J_k}\}$,

$$\begin{aligned} n^\tau(k) &= (|S_{k,j} \bigcap \mathbf{Y}^\tau|,\ j = 1, \ldots, J_k), \\ \nu^\tau(k) &= (|S_{k,j} \bigcap \mathbf{X}^\tau|,\ j = 1, \ldots, J_k), \quad k = 1, \ldots, K. \end{aligned}$$

Then $Z_k^\tau = \nu_j^\tau(k)/n_j^\tau(k)$ on $S_{k,j}$. Since the discrete random vector $\nu(k)$, given $n_j^\tau(k)$, has the *multivariate hypergeometric* distribution with the parameters N and $n(k)$, the condition distribution of T^τ given $\mathbf{Y}$ and the partition $\mathcal{P}$ depends on $\mathbf{Y}$ only through the *"sizes" of the partition sets*, $n^\tau(k)$ $k = 1, \ldots, K$. This determines a_k and b_k in (3).

3 Testing independence: an example

To generate a sample from $F_H = F_1 \cdot F_2$ we apply *bootstrap* and resample from the empirical distribution $\hat{F}_H := \Psi(\hat{F}) = \hat{F}_1 \cdot \hat{F}_1$ 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 computer simulations have been performed for $d \leq 10$, $200 \leq N = M \leq 1000$, and $m = 1, 3, 7, \infty$. To find critical values c_α for the test we use Monte Carlo method.

Preliminary results of Monte Carlo simulations show that the procedure proposed is promising. The dependence of c_α on the dimension d and the partition procedure is weak and can be reduced by imposing appropriate additional requirements on the latter.

References

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

[2] Bousquet O., Boucheron S., Lugosi G. (2004) Introduction to Statistical Learning Theory In: *Advanced Lectures on Machine Learning* Lecture Notes in Artificial Intelligence 3176, pp. 169-207.

[3] Bowman A.W., Foster P.J. (1993) Adaptive smoothing and density based tests of multivariate normality. *J. Amer. Statist. Assoc.* Vol. **88**, pp. 529-537.

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

[5] Huang L.-S. (1997) Testing goodness-of-fit based on a roughness measure. *J. Amer. Statist. Assoc.* Vol. **92**, pp. 1399-1402.

[6] Hyvrinen A., Karhunen J., Oja E. (2001) *Independent Component Analysis*. John Wiley and Sons, Inc.

[7] Polonik W. (1999) Concentration and goodness-of-fit in higher dimensions: (asymptotically) distribrution free methods. *Ann. Statist.* Vol. **27**, pp. 1210-1229.

[8] Szekely G.J., Rizzo M.L. (2005) A new test for multivariate normality. *J. Multiv. Anal.* Vol. **93**, pp. 58-80.

[9] Szekely G.J., Rizzo M.L. (2006) Testing for Equal Distributions in High Dimension. interstat.statjournals.net/YEAR/2004/articles/0411005.pdf

[10] Vapnik V.N. (1998). *Statistical Learning Theory*. Wiley, New York.

[11] Verdinelli I., Wasserman L. (1998) Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.* Vol. **26**, pp. 1215-1241.

[12] Zhu L.-X. , Neuhaus G. (2000) Nonparametric Monte Carlo Tests for Multivariate Distributions. *Biometrika*, Vol. **87**, pp. 919-928.

[13] Zhu L.-X., Fang K.T., Bhatti M.I. (1997) On estimated projection pursuit-type Cramer-von Mises statistics. *J. Multiv. Anal.* Vol. **63**, pp. 1-14.