EMPIRICAL BAYES TESTING GOODNESS-OF-FIT FOR HIGH-DIMENSIONAL DATA

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

Abstract

In [6] a simple, data-driven and computationally efficient procedure of (nonparametric) testing for high-dimensional data have been introduced. The procedure is based on randomization and resampling, a special sequential data partition procedure, and χ^2-type test statistics. However, the χ^2 test has small power when deviations from the null hypothesis are small or sparse. In this note test statistics based on the nonparametric maximum likelihood and the empirical Bayes estimators in an auxiliary nonparametric mixture model are proposed instead.

1 Introduction

Let $\mathbf{X} := (X(1), \dots, X(N))$ be a sample of the size N of iid observations of a random vector X having a distribution P on $\mathbf{R}^d$. We are interested in testing (nonparametric) properties of P in case the dimension d of observations is *large*.

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], multivariate nonparametric Monte Carlo tests [10], and classical univariate nonparametric statistics calculated for data projected onto the directions found via the projection pursuit [11, 7].

More advanced technique is based on Vapnik-Chervonenkis theory, the uniform functional central limit theorem and inequalities for large deviation probabilities [8, 2]. Recently, especially in applications, the Bayes approach and Markov chain Monte Carlo methods are widely used (see, e.g. [9] and references therein).

In [6] a simple, data-driven and computationally efficient procedure of nonparametric testing for *high-dimensional data* have been introduced. The procedure is based on randomization and resampling (bootstrap), a special sequential data partition procedure, and χ^2-type statistics.

The goal of this note is to propose more efficient than χ^2 test statistics based on the nonparametric maximum likelihood (NML) and the empirical Bayes (EB) estimators in an auxiliary nonparametric mixture model.

2 Simple testing procedure

Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be two disjoint classes of d-dimensional distributions, $\mathcal{P} := \mathcal{P}_0 \bigcup \mathcal{P}_1$. Consider a nonparametric hypothesis testing problem:

$$H_0: \ P \in \mathcal{P}_0 \quad \text{versus} \quad H_1: \ P \in \mathcal{P}_1. \tag{1}$$

Suppose that there exists a continuous (in some topology) mapping $\Psi\colon \mathcal{P} \to \mathcal{P}_0$ such that $\mathcal{P}_0 = \{P \in \mathcal{P}\colon \Psi(P) = P\}$. One can take, for example, $\Psi(P) = \operatorname{argmin}_{Q \in \mathcal{P}_0} \varrho(Q, P)$ where ϱ is a distance in $\mathcal{P}$.

Let $\hat{P}$ denote the empirical distribution based on the sample $\mathbf{X}$ and define $\hat{P}_0 := \Psi(\hat{P})$. Under the null hypothesis the empirical distributions $\hat{P}$ and $\hat{P}_0$ for large N should be close since they both are the approximations to the same distribution P_0. Thus, any measure of discrepancy between $\hat{P}$ and $\hat{P}_0$ can be taken as a test statistic for (1). In [6] the following discrepancy measure T has been calculated.

Generate two independent random samples $\mathbf{X}_P$ and $\mathbf{X}_0$ of size N from the distributions $\hat{P}$ and $\hat{P}_0$, respectively. Let $\mathbf{X}^*$ denote the joint sample of $\mathbf{X}_P$ and $\mathbf{X}_0$,

$$\mathbf{X}^* := \mathbf{X}_P \mid\mid \mathbf{X}_0 = (X_P(1), \dots, X_P(N), X_0(1), \dots, X_0(N)).$$

Further, let $\mathcal{S} := \{\mathcal{S}_k, \ k = 1, \dots, K\}$, be a sequence of partitions of $\mathbf{X}^*$ with $|\mathcal{S}_k| = k$ elements produced by some binary partition algorithm. Initially $\mathcal{S}_1 := \{\mathbf{X}^*\}$, and for $k = 2, \dots, K$ the next partition $\mathcal{S}_k$ is obtained from the previous $\mathcal{S}_{k-1}$ by splitting some set from $\mathcal{S}_{k-1}$ into two disjoint subsets.

For a fixed partition $\mathcal{S}_k = \{S_1^k, \dots, S_k^k\}$ and $Q \in \{P, 0\}$, define

$$Y_Q = Y_Q(k) := (Y_Q(1), \dots, Y_Q(k))^\top := (|S_j^k \bigcap \mathbf{X}_Q|, \ j = 1, \dots, k)^\top. \tag{2}$$

Thus, Y_Q is a k-dimensional vector with jth component equal to the number of elements of $\mathbf{X}_Q$ in the set S_j^k $(j = 1, \dots, k)$. Denote

$$\eta := (Y_P - Y_0)/\sqrt{Y_P + Y_0}; \tag{3}$$

here the operations are performed coordinatewise. When the number of observations $Y_P(j) + Y_0(j)$ in the each set S_j^k, $j = 1, \dots, k$, is large and the null hypothesis H_0 holds, the distribution of the vector η is approximately standard normal. Therefore it is natural to take χ^2 statistic $|\eta|^2$ as the discrepancy measure T between $\hat{P}$ and $\hat{P}_0$ and to use it as a test statistic for (1). Actually, with the statistic $T = |\eta|^2$, the null hypothesis

$$H_0^\eta : \mathbf{E}\eta = 0_k \quad \text{versus} \quad H_1^\eta : \mathbf{E}\eta \neq 0_k \tag{4}$$

is tested instead. (Here 0_k stands for the null vector in $\mathbf{R}^k$.)

However, χ^2 test has small power when the dimension k of η is large and either each component of the mean $\theta := \mathbf{E}\eta$ only slightly differs from 0_k or only a few θ components are nonzero.

In the next section we apply the nonparametric maximum likelihood estimator and the nonparametric empirical Bayes method to construct a more efficient criterion to test H_0^η and hence H_0.

3 Nonparametric maximum likelihood estimator and empirical Bayes

Consider auxiliary problem (4) where $\eta \sim Normal_k(\theta, I_k)$ and $\theta \in \mathbf{R}^k$ is a vector of unknown parameters. In the (empirical) Bayes approach, the unknown parameter θ is treated as random. Thus, we consider a nonparametric Gaussian mixture model with a mixture distribution G

$$\begin{aligned} \eta &= \theta + z, \quad \theta \text{ and } z \text{ are independent}, && (5)\\ z &\sim Normal_n(0_n, I_n), && (6)\\ \theta_i &\sim G, \quad \{\theta_i,\, i = 1, \ldots, n\} \text{ are iid.} && (7) \end{aligned}$$

For $\nu > 0$, by $\mu_\nu(y \mid G)$ we denote the posterior ν-moment of θ_1 given $\eta_1 = y$

$$\begin{aligned} \mu_\nu(y \mid G) &:= \frac{\varphi_\nu\left(y \mid G\right)}{\varphi_0\left(y \mid G\right)}, && (8)\\ \varphi_\ell\left(y|G\right) &:= \int_{\mathbf{R}} u^\ell \varphi\left(y-u\right) \mathrm{d}G(u), \quad \ell \geq 0. && (9) \end{aligned}$$

Here φ denotes the standard normal distribution density.

The homogeneity hypothesis (4) states that in fact there is no mixture, G is the degenerated at 0 distribution. Since $\mathbf{E}|\eta|^2 = k\mathbf{E}\theta_1^2 + k$, a criterion for testing the null hypothesis H_0^η can be based on an estimator of the functional

$$\mu_2 = \mu_2(G) := \int_{\mathbf{R}} u^2 \, \mathrm{d}G(u) = \mathbf{E}\theta_1^2. \tag{10}$$

Alternatives to the direct estimator $(\widehat{\mu_2})_{\chi^2} := k^{-1}|\eta|^2 - 1$ are the *nonparametric maximum likelihood (NML) estimator*

$$(\widehat{\mu_2})_{ML} := \mu_2\left(\hat{G}_{ML}\right), \tag{11}$$

and the *nonparametric empirical Bayes (NEB) estimator*

$$(\widehat{\mu_2})_{EB} := \frac{1}{k}\sum_{j=1}^{k} \mu_2\left(\eta_j \mid \hat{G}_{ML}\right). \tag{12}$$

Here $\hat{G} = \hat{G}_{ML}$ is the NMLE of the mixture distribution G. For Gaussian mixtures, it does exist and is strongly consistent (see, e.g., [4]). We consider also the NEB statistic

$$\left(\widehat{\mu_1^2}\right)_{EB} := \frac{1}{k}\sum_{j=1}^{k} \mu_1^2\left(\eta_j \mid \hat{G}_{ML}\right). \tag{13}$$

which a biased toward 0 estimator of μ_2.

The performance of proposed test statistics in auxiliary problem (4) and in nonparametric testing for high-dimensional data is compared by means of computer simulation. Preliminary results demonstrate their advantages as compared to χ^2 test especially when deviations from the null hypothesis are either small or sparse.

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] van de Geer S. (2003) Asymptotic theory for maximum likelihood in nonparametric mixture models. *Computational Statistics and Data Analysis*. Vol. **41**, pp. 453-464.

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

[6] Jakimauskas G., Radavičius M., and Sušinskas J. (2008) A simple method for testing independence of high-dimensional random vectors, *Austrian J. Statist.*, Vol. **44**, pp. 101–108.

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

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

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

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

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