Lietuvos matematikos rinkinys. LMD darbai
51 tomas, 2010, 402–407

ISSN 0132-2818
www.mii.lt/LMR/

Application of the empirical Bayes approach to nonparametric testing for high-dimensional data

Gintautas Jakimauskas, Jurgis Sušinskas

Matematikos ir informatikos institutas
Akademijos g. 4, LT-08663 Vilnius
E-mail: gnt@ktl.mii.lt; jur@ktl.mii.lt

Abstract. In [5] 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

Keywords: empirical Bayes, chi-square test, high-dimensional data, nonparametric maximum likelihood estimator, nonparametric testing, posterior mean, simulations.

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, 4], multivariate nonparametric Monte Carlo tests [12], 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 [9, 2]. Recently, especially in applications, the Bayes approach and Markov chain Monte Carlo methods are widely used (see, e.g., [10] and references therein).

In [5] 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.

1 Simple testing procedure

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

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

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 [5] the following discrepancy measure T_0 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 $\mathbf{X}_0$,

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

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

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

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

$$\eta_0 := (Y_P - Y_0)/\sqrt{Y_P + Y_0} \in \mathbf{R}^k, \tag{3}$$

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_0 can be approximated by $(k-1)$-dimensional standard normal distribution. Therefore it is natural to take χ^2 statistic $|\eta_0|^2$ as the discrepancy measure between $\hat{P}$ and $\hat{P}_0$ and to use it as a test statistic for (1). Actually, with the the statistic $|\eta_0|^2$, the null hypothesis

$$H_0^\eta\colon \mathbf{E}\eta_0 = 0_k \quad \text{versus} \quad H_1^\eta\colon \mathbf{E}\eta_0 \neq 0_k$$

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

The approximate covariance matrix of the statistics T_0, however, depends on the alternative H_P. Therefore variance-stabilizing transformation is used giving a new discrepancy vector

$$\eta := \sqrt{Y_P + Y_0}\left(\arcsin\left(\sqrt{\frac{Y_P}{Y_P + Y_0}}\right) - \arcsin\left(\sqrt{\frac{Y_0}{Y_P + Y_0}}\right)\right). \tag{4}$$

Moreover, χ^2 test has a small power when the dimension n of η is large and either each component of the mean $\theta := \mathbf{E}\eta$ only slightly differs from 0_n 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 powerful criterion to test H_0^η and hence H_0.

2 Auxiliary testing problem and empirical Bayes

Let us consider an auxiliary testing problem

$$H_0^\eta\colon \mathbf{E}\eta = 0_n \quad \text{versus} \quad H_1^\eta\colon \mathbf{E}\eta \neq 0_n, \tag{5}$$

where $\eta \sim \mathit{Normal}_n(\theta, I_n)$ and $\theta \in \mathbf{R}^n$ is an unknown mean vector. 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

$$\eta = \theta + z, \quad \theta \text{ and } z \text{ are independent}, \tag{6}$$

$$z \sim \mathit{Normal}_n(0_n, I_n), \tag{7}$$

$$\theta_i \sim G, \quad \{\theta_i,\ i = 1, \dots, n\} \text{ are iid}. \tag{8}$$

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

$$\mu_\nu(y \mid G) := \frac{\varphi_\nu(y \mid G)}{\varphi_0(y \mid G)}, \tag{9}$$

$$\varphi_\ell(y \mid G) := \int_{\mathbf{R}} u^\ell \varphi(y - u)\, dG(u), \quad \ell \geqslant 0. \tag{10}$$

Here φ denotes the standard normal distribution density.

The homogeneity hypothesis (5) states that in fact there is no mixture, G is the degenerated at 0 distribution. Since $\mathbf{E}|\eta|^2 = n\mathbf{E}\theta_1^2 + n$, a criterion for testing the null hypothesis H_0^η can be based on an estimator of the functional

$$\mu_2 = \mu_2(G) := \int_{\mathbf{R}} u^2\, dG(u) = \mathbf{E}\theta_1^2. \tag{11}$$

Alternatives to the direct estimator $(\hat{\mu}_2)_{\chi^2} := n^{-1}|\eta|^2 - 1$ are the *nonparametric maximum likelihood estimator* (NMLE)

$$(\hat{\mu}_2)_{ML} := \mu_2(\hat{G}_{ML}), \tag{12}$$

and the *nonparametric empirical Bayes* (NEB) estimator

$$(\hat{\mu}_2)_{EB} := \frac{1}{n}\sum_{j=1}^{n} \mu_2(\eta_j \mid \hat{G}_{ML}). \tag{13}$$

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., [8]). We consider also the NEB statistic

$$(\hat{\mu}_1^2)_{EB} := \frac{1}{n}\sum_{j=1}^{n} \mu_1^2(\eta_j \mid \hat{G}_{ML}). \tag{14}$$

which is a biased toward 0 estimator of μ_2.

Jiang and Zhang [6] have proved that the NEB estimator

$$\hat{\theta} := \big(\mu_1(\eta_j \mid \hat{G}_{ML}),\ j = 1, \dots, n\big)$$

of θ asymptotically achieves the minimial in the class of separable statistics mean square error R_n^* provided

$$(\log n)^{9/2} \min\big(\sqrt{\log n}, \|\theta\|_\infty\big) = o\big(nR_n^*\big) \quad (n \to \infty).$$

They also have shown via simulations that in some cases $\hat{\theta}$ significantly outperforms other known counterparts including James–Stein estimator. Since $\hat{\theta}$ is location invariant, this suggests that the criterion for testing (5) based on the statistics $(\hat{\mu}_1^2)_{EB}$ might be more powerful especially for close alternatives.

The asymptotic properties of $(\hat{\mu}_1^2)_{EB}$ can be derived from that of $|\hat{\theta} - \theta|^2$. In this paper we are interested in finite sample properties of $(\hat{\mu}_1^2)_{EB}$ and present simulation results for some natural alternatives.

3 Simulation experiment and concluding remarks

The following three alternatives of θ_i distribution are considered:

(a1) $\theta_i = au_i$, $u_i \sim Normal(0, 1)$;

(a2) $\theta_i = a(2z_i - 1)$, $z_i \sim Binomial(1, 1/2)$;

(a3) $\theta_i = a(-1)^i \cdot \mathbf{1}\{i \leqslant m\}$, $1 < m < n$.

For various combinations of the parameters a, n and m, simulations with 1000 replications have been performed. The parameter $a > 0$ represents the difficulty of the testing problem. The simulations show some improvements in power of NEB test in comparison with χ^2 test. Figs. 1–3 illustrate the typical results. Here power plots for the test statistic $(\hat{\mu}_1^2)_{EB}$ and for χ^2 test versus a are given for $n = 50$ and $m = 8$.

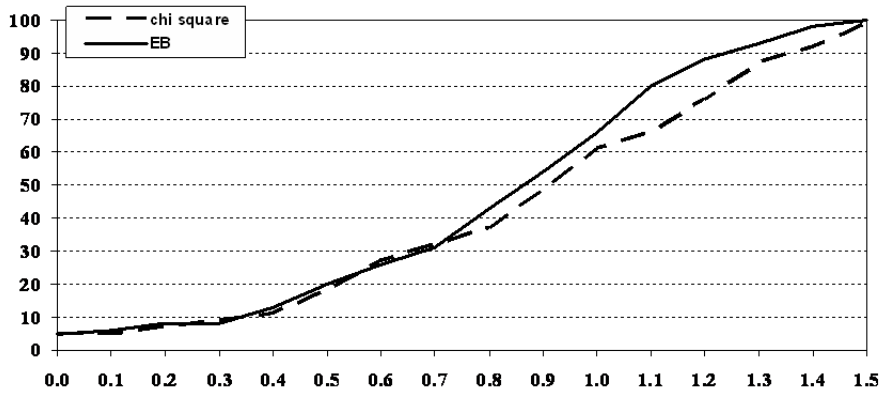

Fig. 1. The Power of the tests for alternative (a1).

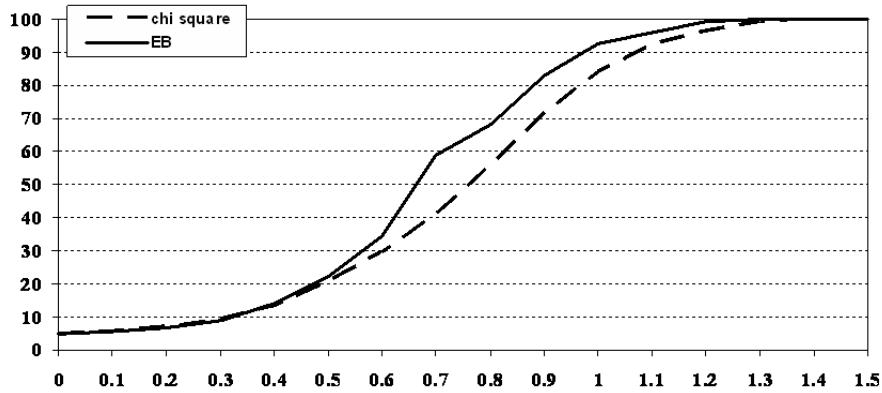

Fig. 2. The Power of the tests for alternative (a2).

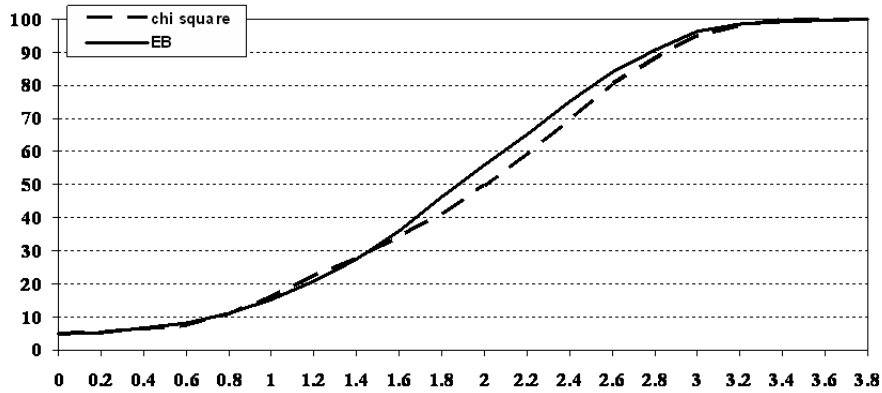

Fig. 3. The Power of the tests for alternative (a3).

The NMLE $\hat{G}_{ML}$ is calculated by making use of the EM algorithm for a finite Gaussian mixture with prespecified and fixed centers of the mixture components (see, e.g., [6]). The number of the components $m = 15$. This means that actually the *restricted* NMLE is substituted for $\hat{G}_{ML}$.

Concluding remarks

The initial nonparametric testing problem (1) for high-dimensional data is reduced to the auxiliary testing problem (5) using the method proposed in [5]. In the empirical Bayes setting the null hypothesis H_0^η can be restated as $G = \delta_0$, where G is the prior distribution of the unknown parameters θ_i, $i = 1, \ldots, n$, and δ_0 is the degenerate at 0 distribution. Thus, any discrepancy measure between the δ_0 and NMLE $\hat{G}_{ML}$ of G can be used for testing (5), in particular, χ^2 test or nonparametric likelihood ratio criterion. In the paper the finite sample properties of the test based on the NEB statistic $(\hat{\mu}_1^2)_{EB}$ (see (14)) are investigated by means of simulations.

Preliminary simulation results show some improvements of NEB test as compared with χ^2. Since NMLE $\hat{G}_{ML}$ calculation is an iterative and time consuming procedure the results can depend on the calculation method and the number of iterations.

References

[1] L. Baringhaus and N. Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, **35**:339–348, 1988.

[2] O. Bousquet, S. Boucheron and G. Lugosi. Introduction to statistical learning theory. In O. Bousquet, U. von Luxburg and G. Rätsch(Eds.), *Advanced Lectures on Machine Learning, 2004*, Lecture Notes in Artificial Intelligence, vol. 3176, pp. 169–207. Springer, 2004.

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

[4] L.-S. Huang. Testing goodness-of-fit based on a roughness measure. *J. Amer. Statist. Assoc.*, **92**:1399–1402, 1997.

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

[6] W. Jiang and C.-H. Zhang. General maximum likelihood empirical bayes estimation of normal means. *Ann. Statist.*, **37**:1647–1684, 2009.

[7] G.J. Szekely and M.L. Rizzo. A new test for multivariate normality. *J. Multiv. Anal.*, **93**:58–80, 2005.

[8] S. van de Geer. Asymptotic theory for maximum likelihood in nonparametric mixture models. *Computational Statistics and Data Analysis*, **41**:453–464, 2003.

[9] V.N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[10] I. Verdinelli and L. Wasserman. Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.*, **26**(4):1215–1241, 1995.

[11] L.-X. Zhu, K.T. Fang and M.I. Bhatti. On estimated projection pursuit-type Crameér-von mises statistics. *J. Multiv. Anal.*, **63**:1–14, 1997.

[12] L.-X. Zhu and G. Neuhaus. Nonparametric Monte Carlo tests for multivariate distributions. *Biometrika*, **87**:919–928, 2000.

REZIUMĖ

Empirinio Bajeso metodo taikymas didelio matavimo duomenų neparametriniams testams

G. Jakimauskas, J. Sušinskas

Straipsnyje [5] buvo pasiūlyta paprasta, adaptyvi ir skaitiškai efektyvi procedūra didelio matavimo duomenų (neparametrinėms) hipotezėms tikrinti. Procedūra remiasi randomizacija, saviranka, specialia duomenų rinkinio suskaidymo procedūra ir χ^2 tipo testais. Tačiau χ^2 testas turi mažą galią, kai nukrypimai nuo nulinės hipotezės yra maži arba išsklaidyti. Šiame darbe vietoje jo siūlomas kitas testas, kuris remiasi neparametriniu didžiausio tikėtinumo empiriniu Bajeso įvertiniu pagalbiniame neparametriniame skirstinių mišinių modelyje.

Raktiniai žodžiai: empirinis Bajeso metodas, chi-kvadrat testas, didelio matavimo duomenys, neparametrinis didžiausio tikėtinumo įvertinys, neparametriniai testai, aposteriorinis vidurkis, imitacinis modeliavimas.