Liet. mat. rink. LMD darbai, **50**, 2009, 293–297

Efficient algorithm for testing goodness-of-fit for classification of high dimensional data

Gintautas JAKIMAUSKAS
Institute of Mathematics and Informatics
Akademijos 4, LT-08663 Vilnius, Lithuania
e-mail: gnt@ktl.mii.lt

Abstract. Let us have a sample satisfying d-dimensional Gaussian mixture model (d is supposed to be large). The problem of classification of the sample is considered. Because of large dimension it is natural to project the sample to k-dimensional ($k = 1, 2, \ldots$) linear subspaces using projection pursuit method which gives the best selection of these subspaces. Having an estimate of the discriminant subspace we can perform classification using projected sample thus avoiding 'curse of dimensionality'. An essential step in this method is testing goodness-of-fit of the estimated d-dimensional model assuming that distribution on the complement space is standard Gaussian. We present a simple, data-driven and computationally efficient procedure for testing goodness-of-fit. 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

Keywords: Gausian mixture model, goodness-of-fit.

Introduction

Let $X = X^N$ be a sample of size N satisfying d-dimensional Gaussian mixture model (we assume that d is large) with distribution function (d.f.) F.

Because of the high dimension of the considered space it is natural to project the sample X to linear subspaces of dimension k ($k = 1, 2, \ldots$) using projection pursuit method. If the distribution of the standardized projected sample on the complementary space is standard Gaussian this linear subspace H is called discriminant subspace. E.g., if we have q Gaussian mixture components with equal covariance matrices then the dimension of the discriminant subspace is equal to $q - 1$.

Having the estimate of the discriminant subspace it is easier to perform the classification using the projected sample.

The step-by-step procedure applied to the standardized sample is the following (here $k = 1, 2, \ldots, d$, until hypothesis of standard Gaussian distribution on the complementary space holds for some k):

1. Finding the best linear subspace of dimension k using the projection pursuit method (see, e.g., [4]).
2. Estimation of the parameters of Gaussian mixture (see, e.g., [3]) from the sample projected to the linear subspace of dimension k.

3. Test goodness-of-fit of the estimated model in the d-dimensional space assuming that distribution on the complementary space is standard Gaussian. If the test fails we increase k and go to Step 1.

The problems related with the Steps 1 and 2 are considered in abovementioned papers and in their references. If we use common methods in the Step 3 the problem is the comparison of some non-parametric density estimate with some parametric density estimate in high dimensional space. Problems related with high dimensional data are often referred to as 'curse of dimensionality' (see, e.g., [1]). As an alternate approach we use Monte-Carlo method and special sequential data partition procedure. More precisely, we resample the given sample assuming that the distribution on the complementary space is standard Gaussian. For the test statistics we use the joined sample and calculate number of data points corresponding to the initial and resampled samples in each partition element. Test statistics is selected in such a way that if the hypothesis holds the distribution of the test statistics weakly depends on the dimension d and of the distribution in the linear subspace. Test criterion is obtained by simulating sufficiently large number (e.g., 100 or 1000) of independent resampled samples for which the hypothesis holds and comparing test criterion value with predefined level.

The efficiency of the algorithm is based on the weak dependence of the test criterion on the dimension d and the distribution in the linear subspace. Computational efficiency is based on the very efficient dyadic data partition procedure and very simple computation of the test statistics.

We will present some computer simulation results. This approach can be used in other situations, e.g., for testing independence of high-dimensional random vectors (see [2]).

1. Test criterion

Assume that we have standardized d-dimensional Gaussian mixture model with corresponding d.f. F. Denote by F_H the d.f. of the corresponding d-dimensional Gaussian mixture model for which distribution on the complementary subspace of dimension d–k is standard Gaussian (recall that k is dimension of the linear subspace obtained by projection pursuit method). Consider the 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 denote a random vector with the distribution function F_p. Let $\pi(Y)$ denote the posterior probability of the population Ω given Y, i.e.

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

Here f and f_H denote distribution densities of F and F_H, respectively.

Let X_H be a sample of size M of i.i.d. vectors from Ω_H independent of X. The joint sample is denoted by Y, and Z_j, $j = 1, 2, \ldots, N+M$, is the corresponding sequence of indicators of the population Ω. Let $P = \{P_k,\ k = 0, 1, \ldots, K\}$, $P_0 = \mathbb{R}^d$ be a sequence of partitions of $\mathbb{R}^d$, possibly dependent on Y and let A_k, $k = 0, 1, \ldots, K$,

be the corresponding sequence of σ-algebras generated by these partitions. A computationally efficient choice of P is the sequential dyadic coordinate-wise partition minimizing at each step mean square error in partition sets. The natural choice of the test statistics would by χ^2-type statistics

$$T_k = \hat{\mathbf{E}}(Z_k - p)^2, \quad p = N/(N+M),$$

where $\hat{\mathbf{E}}$ stands for the expectation with the respect to the empirical distribution $\hat{F}$ of Y and $Z_k = \hat{\mathbf{E}}(Z|A_k)$, $k \in \{1, 2, \ldots, K\}$.

2. Computer simulation results

For the computer simulation we selected $M = N$, and the test statistics in explicit form is given by the following formula:

$$T_k = \frac{S_k - (k-1)}{\sqrt{(2k-1)}}, \quad k = 1, 2, \ldots, K,$$

where

$$S_k = \frac{1}{2N} \sum_{j=1}^{k} \left(n^{j,k} - m^{j,k}\right)^2, \quad k = 1, 2, \ldots, K,$$

$n^{j,k}$ and, respectively, $m^{j,k}$, are number of elements of sample X (respectively, sample X_H in jth partition element in partition P_k.

We assumed that discriminant space is known exactly (no errors in finding the best linear subspace). We performed simulations with 100 independent realizations. We obtained maximum and minimum values of the test statistics of corresponding joint realizations. Also we obtained minimum and maximum values of the test statistics excluding 5 per cent highest and 5 per cent lowest values. Dimensions up to 100, typically 10, were considered. Dimension of the discriminant subspace was chosen in range 1–4 (i.e., this dimension depends on the number of mixture components and its parameters), and corresponding range of dimensions of linear subspaces were considered.

The results showed very weak dependence on the selected mixture model and the dimension. Maximum of test statistics excluding 5 per cent highest values appeared to be the suitable criterion to accept or reject the considered hypothesis.

In Fig. 1 and Fig. 2 we present minimum and maximum values of the test statistics (and excluding 5 per cent extreme values) for an example of 3 component Gaussian mixture in 10-dimensional space with component means $(-4, -1, 0, \ldots, 0)$, $(0, 2, 0, \ldots, 0)$, $(4, -1, 0, \ldots, 0)$ and unit covariance matrices. Clearly, the dimension of the discriminant subspace is equal 2. In Fig. 1 (respectively, in Fig. 2) we projected to 10-dimensional data to 1-dimensional (respectively, 2-dimensional) subspace.

In Fig. 3 and Fig. 4 we present minimum and maximum values of the test statistics (and excluding 5 per cent extreme values) for an example of 2 component Gaussian mixture in 10-dimensional space with zero component means and diagonal covariance

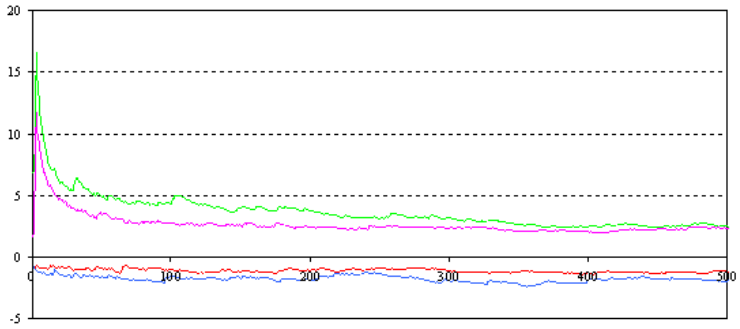

Fig. 1. Behaviour of maximum and minimum of the test statistics ($k = 1$).

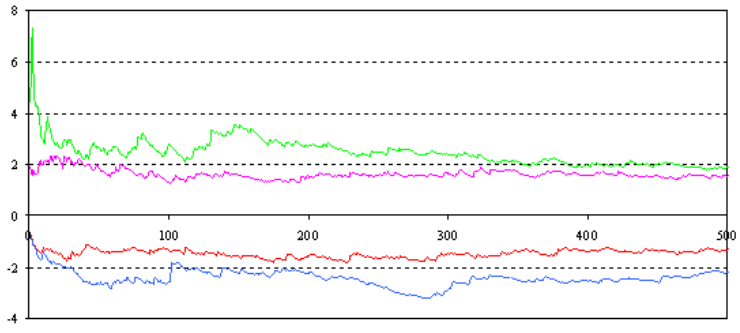

Fig. 2. Behaviour of maximum and minimum of the test statistics ($k = 2$).

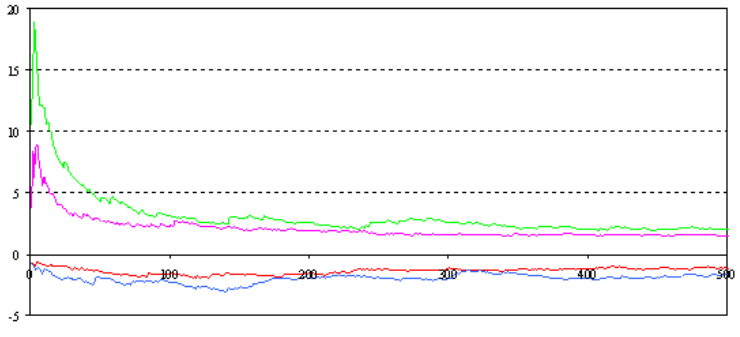

Fig. 3. Behaviour of maximum and minimum of the test statistics ($k = 1$).

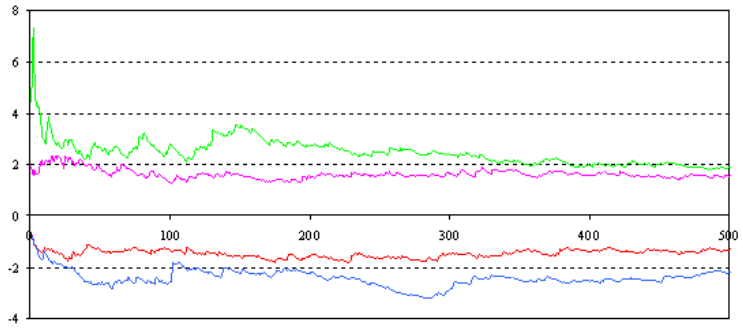

Fig. 4. Behaviour of maximum and minimum of the test statistics ($k = 2$).

matrices with diagonal elements $(10, 1, 10, \ldots, 1)$ and $(1, 10, 1, \ldots, 1)$, respectively. Dimension of the discriminant subspace is equal 2. In Fig. 3 (respectively, in Fig. 4) we projected to 10-dimensional data to 1-dimensional (respectively, 2-dimensional) subspace.

References

1. T. Hastie, R. Tibshirani, J.H. Friedman. *The Elements of Statistical Learning*. Springer, 2001.
2. G. Jakimauskas, M. Radavičius, J. Sušinskas. A simple method for testing independence of high-dimensional random vectors. *Austrian Journal of Statistics*, 37(1):101–108, 2008.
3. R. Rudzkis, M. Radavičius. Statistical estimation of a mixture of Gaussian distributions. *Acta Applicandae Mathematicae*, 38:37–54, 1995.
4. R. Rudzkis, M. Radavičius. Characterization and statistical estimation of a discriminant space for Gaussian mixtures. *Acta Applicandae Mathematicae*, 58:279–290, 1999.

REZIUMĖ

G. Jakimauskas. Efektyvus modelio adekvatumo testavimo algoritmas didelio matavimo duomenų klasifikavimui

Tegul turime matavimo d imtį, tenkinančią Gauso mišinių modelį (laikoma, kad d yra didelis). Nagrinėjama imties klasifikavimo problema. Dėl didelio matavimo yra natūralu projektuoti imtį į matavimo k ($k = 1, 2, \ldots$) tiesinį poerdvį, naudojant tikslinio projektavimo metodą, kuris duoda geriausią šių poerdvių parinkimą. Turėdami diskriminantinės erdvės įvertį galime atlikti klasifikavimą naudodami projektuotą imtį, tuo išvengdami taip vadinamojo „didelių matavimų prakeiksmo“ (curse of dimensionality). Esminis žingsnis šiame metode yra įvertinto matavimo d modelio adekvatumo testavimas, laikant, kad papildomoje erdvėje pasiskirstymas yra standartinis Gauso. Mes pateikiame paprastą, veikiančią pagal duomenis ir skaičiavimų prasme efektyvią procedūrą modelio adekvatumo testavimui. Ši procedūra remiasi gerai žinoma modelio adekvatumo testavimo interpretacija kaip klasifikacijos problema, specialia nuoseklia duomenų skaidymo procedūra, randomizacija ir pakartotiniu imties generavimu, nuosekliosios analizės elementais. Procedūros efektyvumo įvertinimui naudojami Monte-Carlo metodu generuojami duomenys.

Raktiniai žodžiai: Gauso mišinių modelis, modelio adekvatumo testavimas.