LMD mokslo darbai, III tomas, 370–375

Appropriateness of projection pursuit in classification

Gintautas JAKIMAUSKAS (IMI), Ričardas KRIKŠTOLAITIS (VMU)
e-mail: gnt@ktl.mii.lt, ricardas_krikstolaitis@fc.vdu.lt

1. Introduction

The most common method of estimating a posteriori probabilities in classification is "plug-in" principle when unknown parameters in calculation of a posteriori probabilities are replaced by the maximum likelihood estimates (MLE). However, for comparatively small sample size it is better to apply biased estimators in order to reduce the variance of the estimates. One of these methods is to reduce dimension of observations by projecting them to a subspace of lower dimension and to calculate MLE in the subspace.

Theoretical background of this problem is given, e.g., in [2] and [6]. We present computer simulation results that show that for comparatively small sample size classification using projection pursuit algorithm gives better accuracy of estimates of a posteriori probabilities and less classification error. We are thankful to prof. R. Rudzkis who gave the idea and many constructive and valuable remarks.

The introduction presents already known methods. Description of projection pursuit is given in more generalized form that is needed for our purposes for reader's convenience. Further studies are on the way, which will help to make practical decision (probably using bootstrap methods) in which situations use of projection pursuit is preferable (including computational costs of finding discriminant subspace).

Main definitions. Let us have q independent d-dimensional Gaussian random variables Y_i with different distribution densities $\varphi(\cdot; M_i, R_i) \overset{\mathrm{def}}{=} \varphi_i$, where means M_i and covariance matrices R_i, $i = 1, 2, \dots, q$, are unknown. Let ν be random variable (r.v.) independent of Y_i, $i = 1, \dots, q$, and taking on values $1, 2, \dots, q$ with unknown probabilities $p_i > 0, i = 1, 2, \dots, q$, respectively. In this paper we assume that number of classes q is known. We observe d-dimensional r.v. $X = Y_\nu$. Each observation belongs to one of q classes depending on r.v. ν. Distribution density of r.v. X is therefore a Gaussian mixture density

$$f(x) = \sum_{i=1}^{q} p_i\varphi_i(x) \overset{\mathrm{def}}{=} f(x, \theta), \quad x \in \mathbb{R}^d, \tag{1}$$

where $\theta = (p_i, M_i, R_i, i = 1, 2, \dots, q)$ is an unknown multidimensional parameter. Probabilities $p_i = \mathbf{P}\{\nu = i\}$ are *a priori* probabilities for r.v. X to belong to the ith class.

We will consider the general classification problem of estimating *a posteriori* probabilities $\pi(i,x) = \mathbf{P}\{\nu = i|X = x\}$ from the sample $\{X_1, X_2, \ldots, X_N\} \overset{\text{def}}{=} X^N$ of i.i.d. random variables with distribution density (1). Under assumptions above

$$\pi(i,x) = \pi_\theta(i,x) = \frac{p_i\varphi_i(x)}{f(x,\theta)}, \quad i = 1,2,\ldots,q, \quad x \in \mathbb{R}^d. \tag{2}$$

The problem is to estimate the unknown multidimensional parameter θ.

The EM algorithm. If number of classes q is known, then the maximum likelihood estimate θ^* is an efficient estimate of θ. The most common method for calculating the MLE for Gaussian mixtures is so-called EM (Expectation Maximization) algorithm. Let $\pi^N = \{\pi(i,X),\ i = 1,2,\ldots,q,\ X \in X^N\}$ be any given a posteriori probabilities for sample data points X^N. For given π^N, the parameter $\theta = (p_i, M_i, R_i, i = 1,2,\ldots,q)$ is calculated using the following equalities:

$$p_i = \frac{1}{N}\sum_{j=1}^{N}\pi(i,X_j), \quad i = 1,2,\ldots,q, \tag{3a}$$

$$M_i = \frac{1}{N}\sum_{j=1}^{N}\frac{\pi(i,X_j)}{p_i}X_j, \quad i = 1,2,\ldots,q, \tag{3b}$$

$$R_i = \frac{1}{N}\sum_{j=1}^{N}\frac{\pi(i,X_j)}{p_i}(X_j - M_i)(X_j - M_i)^{\mathrm{T}}, \quad i = 1,2,\ldots,q. \tag{3c}$$

For a given θ probabilities π^N are calculated using formula (2). The EM algorithm is an iterative procedure which starts either from a given parameter θ or given probabilities π^N applying in turn formulae (3) and (2). The EM algorithm usually ends after some predefined number of iterations. Parameter θ in the EM algorithm converges to MLE if starting parameter θ^0 is sufficiently close to θ^*.

For mixture distributions the EM algorithm was proposed independently by Schlesinger[11], Hasselblad[7], and Behboodian[3]. On the convergence properties of the EM algorithm see [13]. Also see, e.g., monographs [2], [4], [8], [12]. For further references see [2] and [9].

Discriminant space. Let $V = \mathrm{cov}(X,X)$ be the covariance matrix of r.v. X. Define the scalar product of arbitrary vectors $u, h \in \mathbb{R}^d$ as $(u,h) = u^{\mathrm{T}}V^{-1}h$ and denote by u_L the projection of arbitrary vector $u \in \mathbb{R}^d$ to a linear subspace $L \subset \mathbb{R}^d$. Discriminant space H is defined as a linear subspace $H \subset \mathbb{R}^d$ with the property $\mathbf{P}\{\nu = i|X = x\} = \mathbf{P}\{\nu = i|X_H = x_H\}, i = 1,2,\ldots,q$, $x \in \mathbb{R}^d$, and the minimal dimension. It is known that for Gaussian mixture densities (1) with equal covariance matrices we have $\dim H < q$.

Let $k = \dim H$ and vectors $u_1, u_2, \ldots, u_k$ be a basis in the discriminant space H. Denote $U = (V^{-1}u_1, V^{-1}u_2, \ldots, V^{-1}u_k)^{\mathrm{T}}$. Then $\pi(i,x) = \mathbf{P}\{\nu = i|UX = Ux\}, i = 1,2,\ldots,q$, $x \in \mathbb{R}^d$. This means that projected sample $\{UX_1, UX_2, \ldots, UX_N\}$ is a sufficient statistics for estimating the a posteriori probabilities. The distribution density of the r.v. UX is a Gaussian mixture density

$$f^H(z) = \sum_{i=1}^{q} p_i \varphi_i^H(z) \stackrel{\text{def}}{=} f_q^H(z, \theta_H), \quad z \in \mathbb{R}^k, \tag{4}$$

where $\varphi_i^H = \varphi(\cdot, M_i^H, R_i^H)$, $i = 1, 2, \ldots, q$, are k-dimensional Gaussian distribution densities with means $M_i^H = UM_i$ and covariance matrices $R_i^H = U^{\mathrm{T}} R_i U$, $\theta_H = (p_i, M_i^H, R_i^H,\ i = 1, 2, \ldots, q)$ is the multidimensional parameter.

Projection pursuit algorithm. One of methods to find discriminant space is projection pursuit algorithm. It is a step-by-step procedure to find the basic vectors of the discriminant space. Projection pursuit method was introduced by Friedman and Tukey [6]. Properties of the projection pursuit method also have been studied well enough, see, e.g., [1] and [5]. See also Aivazyan *et al.* [2]. Description below is based on [10].

Let $\boldsymbol{F}$ be the set of one-dimensional Gaussian mixture distribution functions, $\rho = \rho(G_1, G_2)$, $G_1, G_2 \in \boldsymbol{F}$, be some functional satisfying the following conditions: $\rho(G_1, G_1) = 0$ and $\rho(G_1, G_2) > 0$, if $G_1 \neq G_2$. For arbitrary non-zero $u \in \mathbb{R}^d$ define a projection index $Q(u) = \rho(F_u, \Phi)$, where F_u is the distribution function of the standardized r.v. $u^{\mathrm{T}}X$, Φ is the standard Gaussian distribution function.

Let orthonormal vectors $u_1, u_2, \ldots, u_k$ be found step-by-step as follows: set $U_0 = \{0\}$, for $i = 1, 2, \ldots, d$, calculate $u_i = \operatorname{argmax}\{Q(u),\ u \in U_{i-1}^{\perp},\ \|u\| = 1\}$, $U_i = \operatorname{span}\{u_1, u_2, \ldots, u_i\}$, and stop when $Q(u_i) = 0$. We set $k = \min\{i : Q(u_{i+1}) = 0\}$ (by definition $Q(u_{d+1}) = 0$). Of course, in real calculations we use projection index estimate $\widehat{Q}(u) = \widehat{Q}(u, X^N)$ based on the sample X^N and use stopping condition $\widehat{Q}(u_i) < \varepsilon_i$, where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_d$ is some sequence of small positive numbers satisfying certain conditions.

Assume that the covariance matrices of components of X are equal. If the functional ρ is shift- and scale-invariant and $\rho(G_1 * \Phi, G_2 * \Phi) < \rho(G_1, G_2)$ for any Gaussian G_2 and $G_1 \neq G_2$, $G_1 \in \boldsymbol{F}$, then the vectors $u_1, u_2, \ldots, u_k$ form a basis of the discriminant space.

2. Computer simulation results

In this paper we present two typical examples (all performed tests are not covered by this paper) of computer simulation results that demonstrate better accuracy of estimates of the a posteriori probabilities and less classification errors when projection pursuit algorithm is used. We assume that we have sufficiently good starting parameters for the EM algorithm – we start from parameters that were used for simulation of the sample X^N. For projected sample we use corresponding theoretical density (4). We also use theoretical basic vectors of the discriminant space. So presented results do not contain errors due to selection of starting parameters for the EM algorithm and errors due to finding basic vectors of the discriminant space.

In all performed tests we have studied dimensions d in the range 5..10. We have tested the projection pursuit algorithm based on generalized Ω^2 metrics for finding the basic vectors of the discriminant space. Dimension of the discriminant space was in the range 1..2. The sample size N varied from 100 to 400 in order to achieve differences in estimation. Number of classes q was in the range 2..5. Covariance matrices of all partial distribution densities were unit.

We studied accuracy of estimation of the a posteriori probabilities, number of Bayesian classification errors (i.e., classification using estimated parameters vs. classification using theoretical parameters) and true classification errors (i.e., Bayesian classification using estimated or theoretical parameters vs. known true class numbers of the sample). Accuracy of estimation of the a posteriori probabilities is measured as mean absolute distance $l(\widehat{\pi}^N, \pi^N)$ between the estimated a posteriori probabilities $\widehat{\pi}^N$ and the theoretical a posteriori probabilities π^N, i.e.,

$$l(\widehat{\pi}^N, \pi^N) = \frac{1}{Nq} \sum_{i=1}^{q} \sum_{j=1}^{N} |\widehat{\pi}(i, X_j) - \pi(i, X_j)| \,. \tag{5}$$

We compare distance $l(\widehat{\pi}^N, \pi^N)$ and $l(\widehat{\pi}^N_H, \pi^N)$ where $\widehat{\pi}^N$ are obtained from MLE in the initial space and $\widehat{\pi}^N_H$ are obtained from MLE in the discriminant subspace H. Number of Bayesian classification errors is measured as percentage of differences in Bayesian classification comparing classification using known theoretical parameter versus classification using estimated parameter. Recall, that Bayes rule of classification assigns an observation $X \in X^N$ to the ith class if $i = \arg\max_{k=1,2,\dots,q} p_k \varphi_k(X)$.

In Example 1 we have 5-dimensional Gaussian mixture model with five clusters with means $(-2r, 0, 0, 0, 0)$, $(-r, 0, 0, 0, 0)$, $(0, 0, 0, 0, 0)$, $(r, 0, 0, 0, 0, 0)$, $(2r, 0, 0, 0, 0)$. In Figs. 1–3 on x axis we have parameter r. In each case we simulated 100 realizations (sample size $N = 300$). Example 1 shows that for moderate distance between clusters we can achieve significantly less mean absolute error (5) and number of classification errors.

In Example 2 we have 5-dimensional Gaussian mixture model with three clusters with means $(-3, -a, 0, 0, 0)$, $(0, 2a, 0, 0, 0)$, $(3, a, 0, 0, 0)$. In Figs. 4–6 on x axis we have parameter a. In each case we simulated 100 realizations (sample size $N = 300$). In this example $\dim H = 2$ (if $a > 0$). Projection to one-dimensional subspace therefore is not a projection to a discriminant subspace, but for small values of parameter a we achieve less classification errors. For bigger values of parameter a projection to the two-dimensional discriminant subspace gives significantly less classification errors. Example 2 shows that in real calculations we must be aware of selecting too small dimension of the discriminant subspace.

Performed tests show that the projection to the discriminant subspace is definitely recommended if possible. The advantages do not depend much on the sample size. Note that for Gaussian mixture models presented in the examples finding the discriminant subspace is quite simple. In general, finding the discriminant subspace in the higher dimensions is a very time consuming procedure and can significantly reduce advantages of using projection pursuit.

Fig. 1. **Mean absolute error (average of 100 realizations)**

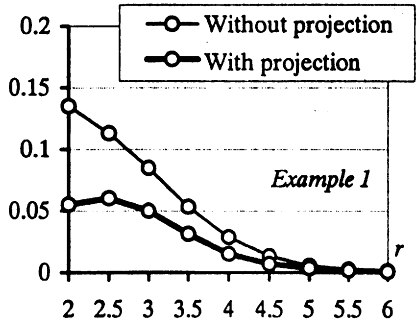


Fig. 2. **Number of classification errors (avg. of 100 realizations)**

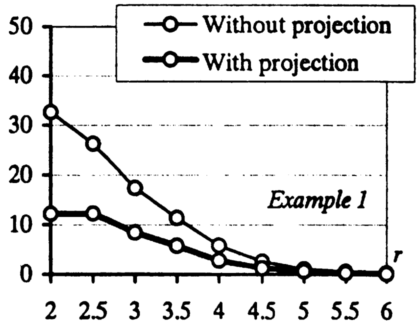


Fig. 3. **True classification errors (average of 100 realizations)**

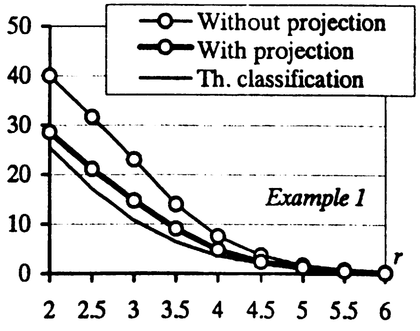


Fig. 4. **Mean absolute error (average of 100 realizations)**

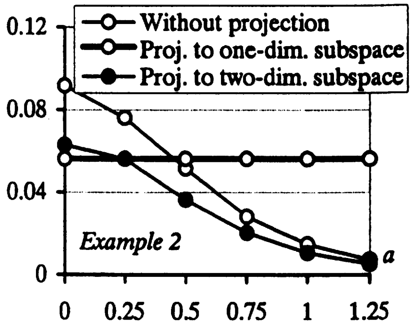


Fig. 5. **Number of classification errors (avg. of 100 realizations)**

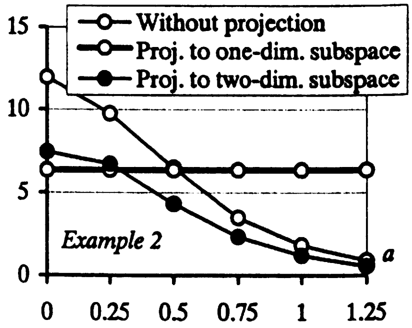
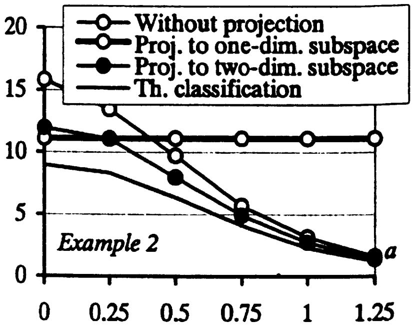


Fig. 6. **True classification errors (average of 100 realizations)**



References

[1] S.A. Aivazyan, Mixture approach to clustering via maximum likelihood, criteria of model complexity and projection pursuit, in: *Data Science, Classification and Related Methods, Abstracts of 5th IFCS Conference*, IFCS, Cobe, **1**, 36 (1996).

[2] S.A. Aivazyan, V.M. Buchstaber, I.S. Yenyukov and L.D. Meshalkin, *Applied Statistics. Classification and Reduction of Dimensionality*, Finansy i Statistika, M., (1989) (in Russian).

[3] J. Behboodian, On a mixture of normal distributions, *Biometrika*, **57**, 215–217 (1970).

[4] B.S. Everitt and D.J. Hand, *Finite Mixture Distributions*, Chapman and Hall, L. (1981).

[5] J.H. Friedman, Exploratory projection pursuit, *J. Amer. Statist. Assoc.*, **82**, 249–266 (1987).

[6] J.H. Friedman and J.W.Tukey, A projection pursuit algorithm for exploratory data analysis, *IEEE Trans. Comput.*, **C-21**, 881–889 (1974).

[7] V. Hasselblad, Estimation of parameters for a mixture of normal distributions, *Technometrics*, **8**, 431–444 (1966).

[8] G.J. McLacklan and K.E. Basford, *Mixture Models. Inference and Applications to Clustering*, Marcel Dekker, N.Y. (1988).

[9] R. Rudzkis and M. Radavičius, Statistical Estimation of a Mixture of Gaussian Distributions, *Acta Applicandae Mathematicae*, **38**, 37–54 (1995).

[10] R. Rudzkis and M. Radavičius, Projection pursuit in Gaussian mixture models preserving information about cluster structure, *Liet. Matem. Rink.*, **37**(4), 550–563 (1997) (in Russian).

[11] M.I. Schlesinger, On spontaneous discrimination of images, in: *Reading Automata*, Naukova Dumka, Kiev, 38–45 (1965) (in Russian).

[12] D.M. Titterington , A.F.M. Smith and U.E. Makov, *Statistical Analysis of Finite Mixture Distributions*, Wiley, N.Y. (1985).

[13] C.F.J. Wu, On the convergence properties of the EM algorithm, *Ann. Statist.*, **11**, 95–103 (1983).

Projektavimo taikymo tikslingumas klasifikavime

G. Jakimauskas, R. Krikštolaitis

Nagrinėtas tikslinio projektavimo algoritmo taikymo tikslingumas vertinant aposteriorines tikimybes iš imties, kai stebimas atsitiktinis dydis tenkina daugiamačio Gauso mišinio modelį. Pateikiami kompiuterinio modeliavimo rezultatai.