Nonlinear Analysis: Modelling and Control, 2001, v. 6, No. 1, 67-77

Statistical Estimation of Discriminant Space Using Various Projection Indices

R. Krikštolaitis

Vytautas Magnus University
Vileikos 8, 3031 Kaunas, Lithuania
ricardas@if.vdu.lt

Received: 09.04.2001
Accepted: 21.04.2001

Abstract

Projection pursuit is a method for finding interesting projections of high-dimensional multivariate data. Typically interesting projections are found by numerical maximizing some measure of non-normality of projected data (so-called projection index) over projection direction. The problem is to select the index for projection pursuit. In this article we compare performance of five projection indices: projection indices based on ω^2, Ω^2, Kolmogorov-Smirnov goodness-of-fit measures, entropy index and Friedman's index. It is supposed that observed random variable satisfies a multidimensional Gaussian mixture model.

Keywords: Gaussian mixture model, discriminant space, projection pursuit, projection index.

1 Introduction

Rather frequently data sets consist of high-dimensional observations. Projection pursuit (PP) is a method for finding interesting projections of high-dimensional multivariate data. The first research into projection pursuit is accredited to Kruskal [6]. However, the first succesful implementation of projection pursuit method was by Friedman and Tukey [3], whose also suggested the name *Projection pursuit*. The first examination of the theoretical aspects of PP was made by Huber [4].

Usually PP procedures are used for high-dimensional distribution density and regression function estimation. However, they also yield a natural way to estimate a discriminant space (see, e.g., Aivazyan et al [1], Rudzkis and Radavičius [7, 8]) and thus are a promising alternative to the principal component method.

The two basic elements of projection pursuit are: a PP index and a PP algorithm. A projection index (PI) is a measure of how interesting a projection is. Usually PI is defined as pseudodistance between the distribution of the projected observation and some "uninteresting" distribution. Frequently the maximum of the index over projections corresponds to the most interesting projection. A natural "uninteresting" distribution is a normal distribution. When applying PP to data, a sample version (estimate) of PI based on the data is used.

In this paper we investigate the role of PI in PP procedures. Five traditional PI are compared by their impact on discriminant space (DS) estimating accuracy by means of computer simulation. The underlying model is a mixture of Gaussian random vectors with equal covariance matrices. The accuracy of DS estimate obtained via PP is evaluated by average of squared distances of its basic vectors to the true DS.

In the next section necessary notation and definitions are introduced. In Section 3 five PI's and their sample estimators are given. The simulation results and preliminary conclusions are presented in 4 and 5 sections, respectively.

I am grateful to prof. R.Rudzkis for the problem formulation and stimulating discussions. I am also thankful to reviewer for constructive remarks.

2 Notation and definitions

Let Y_i, $i = 1, 2, \ldots, q$, be d-dimensional Gaussian random variables with means M_i and covariance matrix R_i, $i = 1, 2, \ldots, q$. Let ν be random variable (r.v.) independent of Y_i, $i = 1, 2, \ldots, q$, and taking on values $1, 2, \ldots, q$ with unknown probabilities $p_i > 0$, $i = 1, 2, \ldots, q$, respectively. We observe d-dimensional r.v. $X = Y_\nu$. The distribution density (d.d.) of r.v. X is therefore a mixture of Gaussian d.d.'s

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

where $\theta = (p_i, M_i, R_i,\ i = 1, 2, \ldots, q)$ is an unknown multidimensional parameter. Probabilities $p_i = \mathbf{P}\{\nu = i\}$ are called *a priori* probabilities.

Discriminant space

Let $V = \mathrm{cov(X, X)}$ be the covariance matrix of r.v. X and suppose for simplicity $\mathbf{E}X = 0$. Define a scalar product on $\mathbf{R}^d$ by the equality

$$(u, h) = u^T V^{-1} h, \quad u, v \in \mathbf{R}^d,$$

and denote by u_H the projection of arbitrary vector $u \in \mathbf{R}^d$ to a linear subspace $H \subset \mathbf{R}^d$.

DEFINITION. A linear subspace $H \subset \mathbf{R}^d$ satisfying the condition

$$\mathbf{P}\{\nu = i | X = x\} = \mathbf{P}\{\nu = i | X_H = x_H\},\ \forall x \in \mathbf{R}^d,\ i = 1, 2, \ldots, q, \tag{2}$$

and having the minimal dimension is called a *discriminant space*.

It is known that for Gaussian mixture densities (1) with equal covariance matrices we have $\dim H < q$ (see Aivazyan et al [1]).

Let $k = \dim H$ and vectors $u_1, u_2, \ldots, u_k$ be a basis in the discriminant space H. Denote $U = (V^{-1}u_1, V^{-1}u_2, \ldots, V^{-1}u_k)^T$. Then

$$\mathbf{P}\{\nu = i | X = x\} = \mathbf{P}\{\nu = i | UX = Ux\},\ i = 1, 2, \ldots, q,\ x \in \mathbf{R}^d.$$

This means that, given a sample $\{X_1, X_2, \ldots, X_N\} \stackrel{def}{=} \mathbf{X}^N$ of X, the projected sample $\{UX_1, UX_2, \ldots, UX_N\}$ is a sufficient statistics for estimating a posteriori probabilities. The distribution density of r.v. UX is a mixture of Gaussian d.d.'s

$$f^H(z) = \sum_{i=1}^{q} p_i \varphi_i^H(z) \stackrel{def}{=} f_q^H(z, \theta_H), \quad z \in \mathbf{R}^k, \tag{3}$$

here $\varphi_i^H = \varphi(\cdot, M_i^H, R_i^H)$, is k-dimensional Gaussian d.d. with mean $M_i^H = UM_i$ and covariance matrix $R_i^H = U^T R_i U,\ i = 1, 2, \ldots, q,\ \theta_H = (p_i, M_i^H, R_i^H,\ i = 1, 2, \ldots, q)$ is a multidimensional parameter.

Projection pursuit algorithm

One of methods to find discriminant space (DS) is projection pursuit algorithm. This is a step-by-step procedure to find the basic vectors of DS.

Let $\mathcal{F}$ be the set of all one-dimensional Gaussian mixture distribution functions, $\rho = \rho(G, \Psi)$, $G, \Psi \in \mathcal{F}$, be some functional satisfying the following conditions:

$$\rho \quad (G, \Psi) > 0, \quad \text{if} \quad G \neq \Psi; \tag{4}$$

$$\rho \quad (G, G) = 0; \tag{5}$$

$$\rho \quad \left(G\left(\frac{\cdot + \lambda}{c}\right), \Psi\left(\frac{\cdot + \lambda}{c}\right)\right) = \rho(G, \Psi), \quad c > 0, \quad \lambda \in \mathbf{R}^1. \tag{6}$$

For arbitrary non-zero $u \in \mathbf{R}^d$ define a projection index $Q(u) = \rho(F_u, \Phi)$, where F_u is the distribution function of the standardized r.v. $u^T X$, Φ is the standard Gaussian distribution function.

Let orthonormal vectors $u_1, u_2, \ldots, u_k$ be found step-by-step as follows:

$$U_0 = \{0\}, \tag{7}$$

$$u_i = \arg\max\{Q(u), u \in U_{i-1}^{\perp}, \|u\| = 1\}, \tag{8}$$

$$U_i = \text{span}\{u_1, u_2, \ldots, u_i\}, \quad i = 1, 2, \ldots, d, \tag{9}$$

and set

$$k = \min\{l : Q(u_{l+1}) = 0\} \tag{10}$$

Then, under some aditional conditions, we have $H = U_k$ [9], i.e., the vectors $u_1, u_2, \ldots, u_k$, determined by (7)-(10) constitute a basis in the DS H. In real calculations we use projection index estimate $\widehat{Q}(u) = \widehat{Q}(u, \mathbf{X}^N)$ based on the sample $\mathbf{X}^N$.

3 Projection indices

The choice of the projection index is the most critical aspect of projection pursuit technique. In this section we define the five PI's whose appropriateness for DS estimation are to be investigated in the last section.

Let $X_1, X_2, \ldots, X_N$ be independent identically distributed random vectors with common distribution function (d.f.) $G \in \mathcal{F}$. In the sequel g (ψ) stands for the d.d. of G (respectively, $\Psi \in \mathcal{F}$).

Denote $Y_j = \Phi(\frac{X_j - \overline{X}}{S})$, where Φ is the standard Gaussian d.f.,

$$\overline{X} = \frac{1}{N}\sum_{j=1}^{n} X_j, \tag{11}$$

and

$$S^2 = \frac{1}{N-1}\sum_{j=1}^{N}(X_j - \overline{X}). \tag{12}$$

The following five projection pursuit indices are to be compared. Let $G, \Psi \in \mathcal{F}$.

1. The projection index based on ω^2 goodness-of-fit measure (ω^2 PI)

$$\rho_1(G,\Psi) = N \cdot \int_{-\infty}^{\infty} (G(x) - \Psi(x))^2 \mathrm{d}\Psi(x). \tag{13}$$

Statistical estimate is given by equality

$$\widehat{\rho}_1 = \frac{1}{12N} + \sum_{j=1}^{N}\left(Y_j - \frac{2j-1}{2N}\right)^2. \tag{14}$$

2. The projection index based on Ω^2 goodness-of-fit measure (Ω^2 PI)

$$\rho_2(G,\Psi) = N \cdot \int_{-\infty}^{\infty} \frac{(G(x) - \Psi(x))^2}{\Psi(x)(1-\Psi(x))} \mathrm{d}\Psi(x). \tag{15}$$

Its statistical estimate is

$$\widehat{\rho}_2 = -N - 2 \cdot \sum_{j=1}^{N}\left(\frac{2j-1}{2N}\ln Y_j + \left(1 - \frac{2j-1}{2N}\right)\ln(1-Y_j)\right). \tag{16}$$

3. The projection index based on Kolmogorov-Smirnov goodness-of-fit measure (KSPI):

$$\rho_3(G,\Psi) = \sup_x |G(x) - \Psi(x)|. \tag{17}$$

Its statistical estimate is

$$\widehat{\rho}_3 = \max(D_N^+, D_N^-), \tag{18}$$

where

$$D_N^+ = \max_{1 \le j \le N}\left(\frac{j}{N} - Y_j\right)$$

and

$$D_N^- = \max_{1 \le j \le N}\left(Y_j - \frac{j-1}{N}\right).$$

4. The entropy index (EPI)

$$\rho_4(G,\Psi)=\int_{-\infty}^{\infty}\ln\left(\frac{\psi(x)}{g(x)}\right)\psi(x)\mathrm{d}x. \tag{19}$$

To estimate ρ_4 we use k-nearest neighbors method:

$$\widehat{\rho}_4=\frac{1}{N-2k}\sum_{j=k+1}^{N-k}\ln\left(\frac{2k+1}{N}\cdot\frac{1}{(Y_{j+k}-Y_{j-k})}\right), \tag{20}$$

where $k=[\sqrt{N}]+1$.

5. The most popular is Friedman's index (FPI)

$$\rho_5(G,\Psi)=\int_{-\infty}^{\infty}\left(\frac{g(x)}{\psi(x)}-1\right)^2\psi(x)\mathrm{d}x. \tag{21}$$

Here for ρ_5 estimating we use another form of ρ_5:

$$\rho_5(G,\Psi)=\int_{-\infty}^{\infty}\frac{g^2(x)}{\psi(x)}\mathrm{d}x-1. \tag{22}$$

Statistical estimate $\widehat{\rho}_5$ of ρ_5 is based on the kernel method

$$\widehat{\rho}_5=\frac{2}{(N-1)Nh}\sum_{j=1}^{N}\sum_{l=j+1}^{N}\lambda(Y_j,Y_l)\cdot W\left(\frac{Y_l-Y_j}{h}\right)-1, \tag{23}$$

where $h=\frac{1}{\sqrt{N}}$,

$$\lambda(Y_j,Y_l)=\begin{cases}1, & \text{when} \quad Y_j\geq h \text{ and } 1-Y_l\geq h,\\ \frac{1}{1-\frac{1}{2}(1-Y_j/h)^2}, & \text{when} \quad Y_j<h,\\ \frac{1}{1-\frac{1}{2}(1-(1-Y_l)/h)^2}, & \text{when} \quad 1-Y_l<h,\end{cases}$$

and

$$W(t)=\begin{cases}(1-|t|), & \text{when} \quad |t|\leq 1,\\ 0, & \text{when} \quad |t|>1,\end{cases} \quad \text{is triangle kernel.}$$

The first three indices, namely, ω^2 PI,Ω^2 PI, and KSPI, are traditional statistics to test normality. EPI and FPI are most popular indices used in projection pursuit algorithms. The foregoing explains our choice of PI's.

The accuracy of estimated DS is measured by the following pseudodistance (discrepancy):

$$D(\widehat{H}\,|H) = \frac{1}{k}\sum_{j=1}^{k}\left|\widehat{u}^{(j)} - \left(\widehat{u}^{(j)}\right)_H\right|^2, \tag{24}$$

where $\widehat{u}^{(j)}$ and $u^{(j)}$ are the estimated and the true basic vectors of the DS.

Thus, $D(\widehat{H}\,|H)$ is an average of squared distance of the basic vectors of the estimated DS to the true DS. In fact, $k \cdot D(\widehat{H}\,|H)$ is equal to the squared Hilbert-Schmidt norm of the projection operator onto the orthogonal complement of DS H restricted on $\widehat{H}$. Hence the measure of accuracy $D(\widehat{H}\,|H)$ is invariant with respect to affine transformations.

4 Simulation results

We investigated 5-dimensional Gaussian mixture models with 3 and 4 components having different means and eqaul covariance matrices. Since the PI's and the accuracy measure D invariant with respect to affine transformations, without loss of generality the covariances are taken to be unit matrices. The dimension of the DS's varies from one to three.

For the first test, we selected 5-dimensional Gaussian mixture model with three clusters with the means $(-r,-a,0,0,0)$, $(0,2a,0,0,0)$, $(r,-a,0,0,0)$, where $r=3$ and a is a parameter. The sample size of simulated data is taken to be $N=100$.

Let us note, that the dimension of DS is one in case $a=0$ and dimension of DS is two for the other a values. However in spite of that, we suppose that always $k=2$. For this case the results are presented in Fig.1. The curves in Fig.1 corresponds to the PI's enumerated in the same order as in section 3. One can observe, that for all a values FPI gives better accuracy, Ω^2 PI, ω^2 PI and KSPI accuracy is similar, while EPI is the "worst" projection index for all a values. However, calculation of FPI is very time consuming procedure as compared with others projection indices, e.g. finding DS basic vectors using KSPI, ω^2 PI and EPI takes approximately 4,5 time less than using FPI.

For the second test, we selected 5-dimensional Gaussian mixture model with four clusters with the means $(-r,-a,-b,0,0)$, $(0,2a,-b,0,0)$, $(r,-a,$

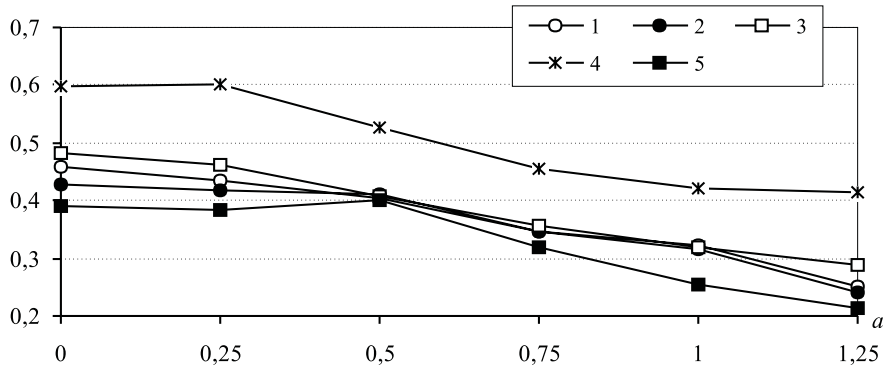

Figure 1: The accuracy of DS estimate vs. the parameter a ($\dim H = 2$)

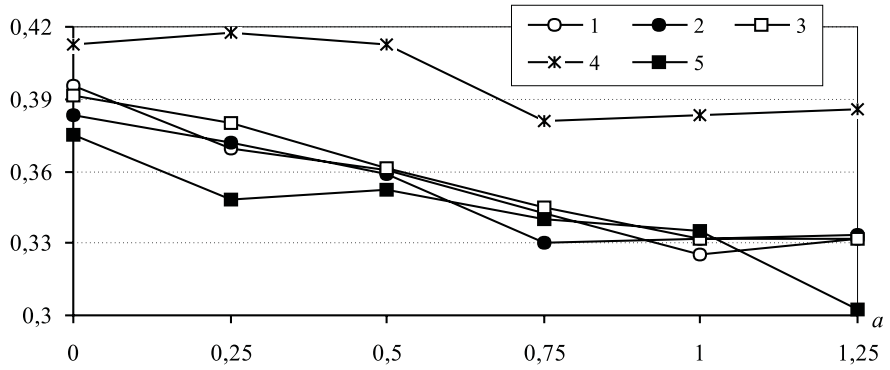

Figure 2: The accuracy of DS estimate vs. the parameter a ($\dim H = 3$, $b = 0$)

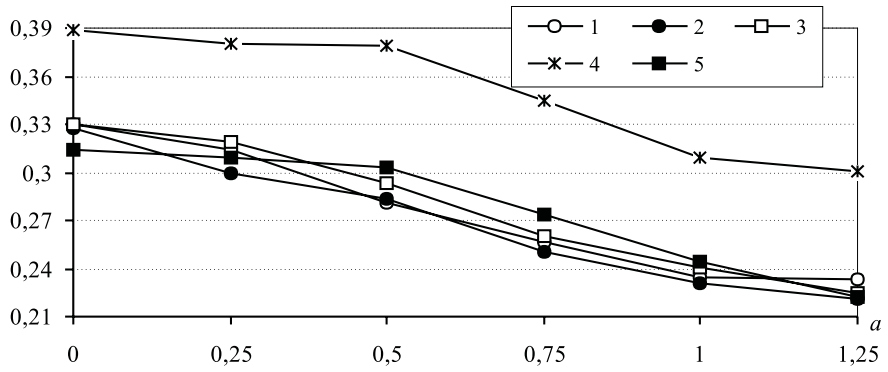

Figure 3: The accuracy of DS estimate vs. the parameter a ($\dim H = 3$, $b = 0,75$)

$-b, 0, 0)$, $(0, 0, 3b, 0, 0)$ respectively, where $r = 3$, a and b are parameters. Calculation were carried out for values: $b = 0; 0,25; 0,5; 0,75; 1,0; 1,25$.

Notice, that in this test the dimension of DS is one in case $a = 0, b = 0$, it is two when $a = 0$ or $b = 0$, and three for other a and b values. However, we suppose that DS dimension is three.

We present results for $b = 0$, $b = 0,75$ and $b = 1,25$. In Fig. 2 we observe FPI advantage for small a values ($a < 0,75$) as compare to other projection indices. However, for greater a values ($a \geq 0,75$) FPI becomes worse, i.e. other projection index gives better accuracy. For large b values (Fig. 3-4) accuracy of estimated DS are similar for all projection indices (except EPI). Therefore, indices which takes less time for calculations, KSPI, ω^2 PI, Ω^2 PI, have an advantage.

5 Conclusions

The tests performed show that the Friedman's projection index gives better accuracy in cases where distance between clusters is close. When the distance increases projection indeces based on Kolmogorov-Smirnov, ω^2, Ω^2 goodness-of-fit measures and Friedman's projection index yields the similar accuracy. In general, taking into account the calculation results and time we can conclude that projection indices based on Ω^2, ω^2 and Kolmogorov-Smirnov goodness-of-fit measures are better than the other. However, this is only the preliminary results and further investigations are necessary for final conclusions.

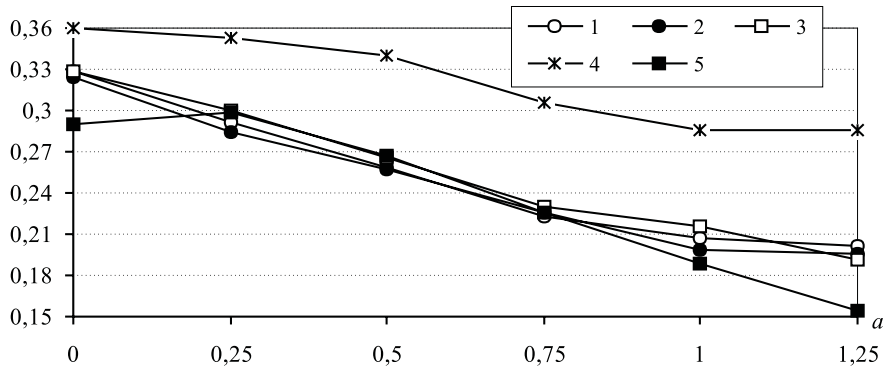

Figure 4: The accuracy of DS estimate vs. the parameter a ($\dim H = 3$, $b = 1,25$)

References

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

2. Friedman J. H. "Exploratory projection pursuit", *J. Amer. Statist. Assoc.*, **82**, p. 249, 1987.

3. Friedman J. H., Tukey J. W. "A projection pursuit algorithm for exploratory data analysis", *IEEE Trans. Comput.*, **C-21**, p. 881, 1974.

4. Huber P. J. "Projection pursuit (with discussion)", *Ann. Statist.*, **13**, p. 435, 1985.

5. Jakimauskas G., Krikštolaitis R. "Influence of Projection Pursuit on Classification Errors: Computer Simulation Results", *Informatica*, **11**(2), p. 115, (2000).

6. Kruskal J. B. "Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'", In *Statistical Computation*, R. C. Milton and J. A. Nelder, ed. New York: Academic Press, 1969.

7. Rudzkis R., Radavičius M. "Statistical Estimation of a Mixture of Gaussian Distributions", *Acta Applicandae Mathematicae*, **38**, p. 37, 1995.

8. Rudzkis R., Radavičius M. "Projection pursuit in Gaussian mixture models preserving information about cluster structure", *Liet. Matem. Rink.*, **37**(4), p. 550, 1997. (in Russian).

9. Rudzkis R., Radavičius M. "Characterization and statistical estimation of a discriminant space for Gaussian mixtures", *Acta Applicandae Mathematicae*, **58**, p. 279, 1999.

10. Sun J. "Projection Pursuit", *Encyclopedia of Statistical Sciences (updated volumes)*, **2**, p. 554, 1998.