Statistical Estimation of Discriminant Space
Using Various Projection Indices

R. Krikštolaĭtis
Vytautas Magnus University
Vieikos 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 $\omega^2$, $\Omega^2$, Kolmogorov-Smirnov goodness-of-fit mea-
sures, entropy index and Friedman’s index. It is supposed that ob-
erved random variable satisfies a multidimensional Gaussian mixture
model.

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

1 Introduction

Rather frequently data sets consist of high-dimensional observations. Pro-
jection 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 successful implementation of
projection pursuit method was by Friedman and Tukey [3], whose also sug-
gested 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], Rudzakis and Raudivičius [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.Rudzakis 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 \( \nu \) 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

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

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

Discriminant space

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

$$(u, h) = u^TV^{-1}h, \ u, v \in \mathbb{R}^d,$$

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

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

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

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

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

This means that, given a sample $\{X_1, X_2, \ldots, X_N\} \overset{def}{=} \mathbf{X}^N$ 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

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

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 $R_i^H = UT_i^TR_i$, $i = 1, 2, \ldots, q$. $\theta_H = (p_i, M_i^H, R_i^H, \ i = 1, 2, \ldots, q)$ 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:

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

\[
\rho(G, G) = 0; \tag{5}
\]

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

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. \( u^T X \), \( \Phi \) is the standard Gaussian distribution function.

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

\[
U_0 = \{0\}, \tag{7}
\]

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

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

and set

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

Then, under some additional 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 \( \hat{Q}(u) = \hat{Q}(u, X^N) \) based on the sample \( X^N \).

### 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 \( g \) (\( \psi \)) stands for the d.d. of \( G \) (respectively, \( \Psi \in \mathcal{F} \)).

Denote \( Y_j = \Phi\left( \frac{X_{j} - \overline{X}}{S} \right) \), where \( \Phi \) is the standard Gaussian d.f.,
and
\[
S^2 = \frac{1}{N-1} \sum_{j=1}^{N} (X_j - \overline{X}).
\] (12)

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

1. The projection index based on \( \omega^2 \) goodness-of-fit measure (\( \omega^2 \) PI)
\[
\rho_1(G, \Psi) = N \cdot \int_{-\infty}^{\infty} (G(x) - \Psi(x))^2 d\Psi(x).
\] (13)

Statistical estimate is given by equality
\[
\hat{\rho}_1 = \frac{1}{12N} + \sum_{j=1}^{N} \left( Y_j - \frac{2j - 1}{2N} \right)^2.
\] (14)

2. The projection index based on \( \Omega^2 \) goodness-of-fit measure (\( \Omega^2 \) PI)
\[
\rho_2(G, \Psi) = N \cdot \int_{-\infty}^{\infty} \frac{(G(x) - \Psi(x))^2}{\Psi(x)(1 - \Psi(x))} d\Psi(x).
\] (15)

Its statistical estimate is
\[
\hat{\rho}_2 = -N - 2 \sum_{j=1}^{N} \left( \frac{2j - 1}{2N} \ln Y_j + \left( 1 - \frac{2j - 1}{2N} \right) \ln(1 - Y_j) \right).
\] (16)

3. The projection index based on Kolmogorov-Smirnov goodness-of-fit measure (KSPI):
\[
\rho_3(G, \Psi) = \sup_x |G(x) - \Psi(x)|.
\] (17)

Its statistical estimate is
\[
\hat{\rho}_3 = \max(D^+_N, D^-_N),
\] (18)
where
\[
D^+_N = \max_{1 \leq j \leq N} \left( \frac{j}{N} - Y_j \right)
\]
and
\[
D^-_N = \max_{1 \leq j \leq N} \left( Y_j - \frac{j - 1}{N} \right).
\]
4. The entropy index (EPI)

\[ \rho_4(G, \Psi) = \int_{-\infty}^{\infty} \ln \left( \frac{\psi(x)}{g(x)} \right) \psi(x) dx. \]  

To estimate \( \rho_4 \) we use \( k \)-nearest neighbors method:

\[ \hat{\rho}_4 = \frac{1}{N - 2k} \sum_{j=k+1}^{N-k} \ln \left( \frac{2k + 1}{N} \cdot \frac{1}{(Y_{j+k} - Y_{j-k})} \right), \]

where \( k = \lfloor \sqrt{N} \rfloor + 1. \)

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

\[ \rho_5(G, \Psi) = \int_{-\infty}^{\infty} \left( \frac{g(x)}{\psi(x)} - 1 \right)^2 \psi(x) dx. \]

Here for \( \rho_5 \) estimating we use another form of \( \rho_5 \):

\[ \rho_5(G, \Psi) = \int_{-\infty}^{\infty} \frac{g^2(x)}{\psi(x)} dx - 1. \]

Statistical estimate \( \hat{\rho}_5 \) of \( \rho_5 \) is based on the kernel method

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

where \( h = \frac{1}{\sqrt{N}}. \)

\[ \lambda(Y_j, Y_l) = \begin{cases} 
1, & \text{when } Y_j \geq h \text{ and } 1 - Y_l \geq h, \\
\frac{1}{1 - \frac{Y_j}{h}}, & \text{when } Y_j < h, \\
\frac{1}{1 - \frac{1-Y_l}{h}}, & \text{when } 1 - Y_l < h, 
\end{cases} \]

and

\[ W(t) = \begin{cases} 
(1 - |t|), & \text{when } |t| \leq 1, \\
0, & \text{when } |t| > 1, \end{cases} \text{ is triangle kernel.} \]
The first three indices, namely, $\omega^2$ PI, $\Omega^2$ 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 PTs.

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

\[
D(\hat{H} | H) = \frac{1}{k} \sum_{j=1}^{k} \left| \hat{\alpha}^{(j)} - \left( \hat{\alpha}^{(j)} \right)_H \right|^2,
\]

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

Thus, $D(\hat{H} | H)$ is an average of squared distance of the basic vectors of the estimated DS to the true DS. In fact, $k \cdot D(\hat{H} | H)$ is equal to the squared Hilbert-Schmidt norm of the projection operator onto the orthogonal complement of DS $H$ restricted on $\hat{H}$. Hence the measure of accuracy $D(\hat{H} | H)$ 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 equal covariance matrices. Since the PTs 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 PTs enumerated in the same order as in section 3. One can observe, that for all $a$ values FPI gives better accuracy, $\Omega^2$ PI, $\omega^2$ 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, $\omega^2$ 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 $H = 3$, $b = 0$)
$-b, 0, 0), (0, 0, 3b, 0, 0)$ 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, $\omega^2$, PI, $\Omega^2$ 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, $\omega^2$, $\Omega^2$ 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 $\Omega^2$, $\omega^2$ 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


