DETECTING STRUCTURES BY MEANS OF PROJECTION PURSUIT

I.S.Yenyukov

Central Economics Mathematical Institute of the USSR Academy of Sciences, Moscow 117418, Krasikova 32.

Summary

In this paper, the consideration of the projection pursuit for testing the presence of clusters is based on the model of the ellipsoidally symmetric unimodal densities mixture. It is shown that under this model the use of projections indices based on Renyi entropy or on third or forth moments results in obtaining an estimate of discriminant subspace. For estimating the Renyi indices values some forms of the order statistics are used. For detecting outliers the ratio of the standard variance estimate to a robust one is proposed as projection index. Indices for discriminant analysis problem are introduced.

Keywords: projection pursuit, discriminant subspace estimating, cluster analysis, outliers detecting, discriminant analysis.

1. Introduction

The projection pursuit (PP) technique (Friedman, Tukey (1974), Buchshtaber, Maslov (1975, 1977), Huber (1985), Jones, Sibson (1987), Friedman (1987)) can be used to reveal any singularities of multivariate data and to find out a geometric structure of it, for instance the presence of clusters, outliers, concentration of points in the vicinity of nonlinear manifolds. The PP approach offers the visual analysis of "interesting" linear projections of the initial data for this goal.

More formally, let us have a sample $\mathbf{X}^n = (X_1,...,X_n)$ of p-variate observations and let $\mathbf{U}$ be a linear orthogonal projection operator from R^p into R^q *(q<<p)* and $Q(\mathbf{U},\mathbf{X}^n)$ be the sampling value of some projection index (PI). The PI is designed in such a way that its value is the larger the more expressed the explored structure in *R* is. We select the q-variate projection $Z = \mathbf{U}_1, X$ with the help of the solution of the problem

$$\mathbf{U}_1 = \arg\max_{\Xi(\mathbf{U})} Q(\mathbf{U},\mathbf{X}^{(n)}) \qquad (1)$$

where $\Xi(\mathbf{U})$ is the set of orthogonal linear operators from R^p into R^q.

Then we can do the visual analysis of q-variate points $Z_i = \mathbf{U_1} X_i$.

The problem (1) is usually rather difficult to solution if q>1. That's why a step-wise approach is often used for obtaining an approximation of solution of (1). In this case we determine the sequence of the projection vectors $U_1,....,U_q$. At first the vector U_1 is determined by the solution of the problem

$$U_1 = \arg\max_U Q(U, \mathbf{X}^{(n)}) \qquad (1')$$

Each following vector U_j is defined from (1') but on additional condition that the influence of all the before obtained vectors is excluded in a suitable way. So the q-dimensional projection $Z = \mathbf{U}_2' X$ can be defined, where the columns of matrix $\mathbf{U_2}$ are just these vectors. Certainly, generally speaking, the projectors $\mathbf{U_2}$ and $\mathbf{U}_1$ result in the different projections. But the problem (1') is more simple and it is possible to show that under some assumptions the use of the step-wise approach does not lead to loss of the information about the explored structure. Such a probability model for cluster structure we discuss in Sections 2. In Section 3 we consider the PI family based on Renyi entropy (Renyi, (1970)). The main result obtained in Section 4 is that the subspace spanned on the projection vectors selected by the step-wise maximization of such PI is the discriminant subspace introduced by Rao (1965). In Section 5 we consider the problem of estimating the PI values. Forms based on the order statistics are proposed. Sections 6, 7 contain some results connected with the use of PI based on high degree moments, including asymptotical results when $p \to \infty, n \to \infty, p/n = \eta > 0$. In Section 8 some consequences of violation of the probability model assumptions are discussed. It is shown by example that PP procedure can select the projection with visual pattern like "dense kern with outliers" even if there are clearly discernable clusters and there are no actual outliers. Detecting the outliers is discussed in Section 9. In Section 10 we consider PP for discriminant analysis.

2. Probability model for describing the cluster structure.

We consider that the underlying distribution of the sample $\mathbf{X}^{(n)}$ is a k-component mixture with the density

$$p(X) = c(d, p, \mathbf{W}) \sum_{i=1}^{k} a_i d((X - M_i)' \mathbf{W}^{-1} (X - M_i)), \qquad (2)$$

where

$d(y)$ is a positive monotonously decreasing function of y (y>0), that has finite moment of (p/2+1) – degree;

c(*d,p,***W**) is the normalizing constant;

$a_i > 0$ is the weight of the *i*-th component;

M_i is the means vector of the *i*-th component;

W is the within-component scatter (covariance) matrix.

So the density of *i*-th component is the unimodal ellipsoidally symmetric function. For instance, if $d(y) = \exp(-y/2)$ it is the normal distribution density. The mixture (2) may be considered as some model of the cluster structure. The density (2) has *k* modes (if the mixture components are enough scattered). And the points surrounding some mode may be regarded as the members of the same cluster.

Now let *z* be an one-dimensional projection *z=U'X* . The density of *z* is *k*-components unimodal symmetric densities mixture

$$f(z) = \sum_{i=1}^{k} a_i e_i(z), e(z) = e((z - m_i)/w)/w,$$

where

$$e(z) = c(d)\int \rho^{p-2} d(z^2 + p^2)dp; m_i = U'M_i; w^2 = U'\mathbf{W}U.$$

The variance of the *z* is $s^2 = U'\mathbf{S}U.$

Now define the discriminant subspace (Rao (1965)). It is good known that the covariance matrix **S** of the *X* may be presented as the sum the within – components scatter matrix **W** and between- components scatter matrix **B**, i. e. **S=B+W**, where $\mathbf{B} = \sum a_i(M_i - M)(M_i - M)', M = \sum a_i, M_i$ is means vector of the *X*. In according to it the variance $s^2(U)$ of one-dimensional projection is $s^2(U) = b^2(U) + w^2(U),$ where $b^2(U) = U'\mathbf{B}U.$ Regard ratio

$$t^2(U) = b^2(U)/w^2(U) \qquad (3)$$

The value $t^2(U)$ is a measure of dissimilarity of the mixture components for the projection with vector *U*. Maximizing the $t^2(U)$ leads to the solution of the generalized eigenvectors problem

$$(\mathbf{B} - t\mathbf{W})V = 0.$$

There are *q'* eigenvectors $V_1,...,V_q$, with corresponding positive eigenvalues $t_1,...,t_q,(q < \min(p,k-1), t_i = t^2(V_i)).$

The discriminant subspace is defined as subspace

$$\Omega = span(V_1,...,V_{q'}).$$

The another definition of the Ω is $\Omega = span(\mathbf{W}^{-1}M_j,...,M_k)$. The number *q'* depends on the geometric configuration of the mean vectors $M_1,....,M_k$.

The Ω contains the complete information about the differences among the mixture (2) components. The PP gives the way to get the Ω without any information about the matrix **W** unlike discriminant analysis problem where we can estimate **W** and **B**. Further without loss generality we will consider the case where *M=0* (centered data).

3. PI suitable for testing the presence of clusters.

We consider the one -parametric family of the PI (Yenyukov, (1986b), Jones, Sibson (1987)) based on Renyi entropy

$$Q_\beta(U,X) = s^\beta E_f f^\beta(z) = s^\beta \int f^{1+\beta}(z)dz, (\beta > 0). \quad (4)$$

Here the "theoretical" quantity of PI is designated by $Q_\beta(U,X)$ in contrast to the sampling quantity $Q_\beta(U,\mathbf{X}^{(n)})$.

We give without the proof the inequalities connecting the $Q_\beta(U,X)$ with the ratio $t^2(U)$ (3)

$$g(e,\beta)(\sum a_i^{1+\beta})(1+t^2(U))^{\beta/2} < Q_\beta(U,X) < g(e,\beta)(1+t^2(U))^{\beta/2} \quad (5)$$

where $g(e,\beta) = E_e e^\beta(z)$ does not depend on the *U*. When $t^2(U)=0$ then $Q_\beta(U,X) = g(e,\beta)$, it is the minimal value of this PI. If the Mahalanobis distances between all the mixture component pairs increase then the low bound is reached asymptotically. And so one may expect that if there are projections with the large values of the t^2 then the maximizing procedure of the Q_β selects these. Certainly, the last reasoning is heuristic rather than exact. However, it is possible to get some exact results. We formulate it as Lemma 1.

Lemma 1. Let the model (2) be true. Assume that the vectors $U_1,...U_q$, are found by the step-wise (sequentional) maximization of the $Q_\beta(U,X)$, and so that the vector U_i is **S**-orthogonal to the subspace $span U_1,...,U_{i-1}$. Then every vector U_i belongs to the discriminant subspace Ω and what is more $\Omega = span(U_1,...,U_{q'})$.

Proof. The PI $Q_\beta(U,X)$ is some function *G* of the $s,w,m_1,...,m_k$, i.e. $Q_\beta(U,X) = G(s,w,m_1,...,m_k)$. Differentiating the G with respect to *U* and equating the result to zero gives

$$(G'_s/s)\mathbf{S}U + (G'_w/w)\mathbf{W}U + \sum G'_{m_i} M_i = 0$$

Multiplying this equality by the **W** (from the left) gives after little algebra the equations for the vector U_1

$$(h(U_1)\mathbf{I}_p + \mathbf{W}^{-1}\mathbf{B})U_1 - V = 0, \qquad (6)$$

where $V = \sum G'_{m_i} \mathbf{W}^{-1} M_i ; h(U) = G'_w + G'_s$.

The vector V is the linear combination of the vectors $\mathbf{W}^{-1} M_i$ from the Ω and therefore, itself belongs to the Ω. Now show that the maximizing vector U_1 belongs to the Ω.

Assume that it is not true, i.e. $U_1 = c_1 U^+ + c_2 U^-$,where $U^+ \in \Omega, U^- \in A$ (the subspace A is **S**-orthogonal to the subspace Ω) and $c_2 \neq 0$.Then the substitution of the U_1 into the(6) yields

$$c_2 h(U_1) U^- = V - c_1 (h(U_1)\mathbf{I}_p + \mathbf{W}^{-1}\mathbf{B})U^+.$$

The right sight of this equality is a vector from Ω. Further, the value *h(U)*=0 only if $t^2(U) = 0$, i.e. if $U \in A$ (it can be checked by direct calculating the derivatives G_s and G_w).

Therefore, the last equality can be true either if $U_1 \equiv A$ (and then $V = 0, c_1 = 0$ and U_1 is not maximizing vector) or if $U^- = 0$. Thus the maximizing vector U_1 belongs to Ω. By analogy it is proved that the vectors $U_2,...,U_{q'}$ belong to the Ω (under the condition **S**-orthogonality). Since rank$(\Omega) = q'$ and these vectors are **S**-orthogonal the $\Omega = span(U_1,...,U_{q'})$ and ones are the basis of Ω. But, generally saying, this basis is other than canonical basis.

The following example shows that the PI (4) are not, generally saying, monotonic functions of the variances ratio t^2 and it explains some behavior details of the step-wise optimization of PI (4).

Example 1. Take two one-dimensional mixtures with the densities

$$f_1(z) = a(2n_1(0,1;z) + n_1(\delta_1,1;z));$$
$$f_2(z) = a(n_1(-\delta_2,1;z) + n_1(0,1;z) + n_1(\delta_2,1;z)),$$

where $n_1(m,\delta^2;z)$ is normal density with parameters *m* and $\delta^2, a = \frac{1}{3}$.

We use PI (4) with $\beta = 1$,i.e. $Q_1 = s\int f^2(z)dz$. The values of the t_i^2, s_i^2 and $E_i = \int f_i^2(x)dx$ *i*=(1,2) are given below

1	E_i	t_i^2	s_i^2
1	$g_0(5+4\exp(-\delta_1^2/4))/9$	$2\delta_1^2/9$	$1+2\delta_1^2/9$
2	$g_0(3+4\exp(-\delta_{21}^2/4)+2\exp(-\delta_2^2))/9$	$2\delta_2^2/3$	$1+2\delta_2^2/3$

When the values δ_1^2 and δ_2^2 are large enough the values
$q_1 = s_1 E_1 \cong (5 \setminus 9) g_0 (1 + 2\delta_1^2 / 9)^{1/2}, q_2 = sF_2 \cong (3/9) g_0 (1 + 2\delta_2^2 / 3)^{1/2}$.
Here $g_0 = \int n_1^2 (0,1; x) dx.$

It is possible to find after some algebra that the equality $\sqrt{27/5}\delta_1 < \delta_2 < \sqrt{3}\delta_1$ entails inequalities $t_1^2 < t_2^2$ but $q_1 > q_2$.

One from the consequences of the example is as follows. If some densities mixture (2) admits projections of both types - $f_1()$ and $f_2()$, then a procedure of selecting the projections from maximizing PI (4) will be at first select the bimodal projections. The effect have been noted as experimental fact in Sibson, Jones (1987) (see to Freedman, Tukey (1974)).

4. Estimating of the PI values.

In practice it is necessary to be able to estimate the PI values for any onedimensional projection. Hear we consider the estimate based on the high –order gaps. This approach have been used for estimated usual entropy functional in Vasicek (1976), Cressie (1978). We use it to estimating the (4).

Let $z_1, ..., z_n$ be projections of the initial observations by a vector U, i.e. $z_i = U'X$ and let $z_{(1)}, ..., z_{(n)}$ be the corresponding order statistics. Then we can design the following estimate of $E_f f^{\beta}(z)$ (we denote it by $E(\beta, U)$)

$$E(\beta, U) = (2r/n)^{\beta} \sum_{i=1}^{n} (z_{(i+r)_-} - z_{(i-r)_+})^{-\beta}$$

where r is integer value $(r < n/2, n = 0(n^d), d > 0)$;
$(i + r)_- = \min(n, 1 + r), (i - r)_+ = \max(n, i - r).$
It is possible to prove that this estimate is consistent and asymptotically normal. Finally, the estimate of the PI (4) is

$$Q_{\beta}(U, \mathbf{X}^{(n)}) = \hat{s} E(\beta, U).$$

The estimate $E(\beta, U)$ depends of the value of *r*. But we have found experimentally that the influence of *r* is not so great. In practice we use the values $r = cn^{1/2} (0.5 < c < 1.5)$.

As the estimates of $\hat{s}^2$ we use the routine estimate and the weighted robust estimate with weights $v_i (i = 1, ..., n)$ defined in section 10.

6. Discriminant subspace estimate based on high degree moments.

The third and forth moments were proposed as PI by Huber (1985), Jones, Sibson (1987). We also used them (Yenyukov, Neuschtadt (1978)). Here we discuss the simple way to obtain with the help of these moments some families of vectors from Ω. We consider here only the case of the third moment. But these results can be applied to the moments of order higher than the third one, and to the characteristic function (about the forth moment see Yenyukov, Neuschtadt (1978)). At first we note that if a vector $U \in \mathbf{M} = span(M_1,...M_k)$ then the vectors $\mathbf{W}^{-1}U$ and $\mathbf{S}^{-1}U$ belong to the Ω. So it is enough to obtain the estimate of vector from … and then multiply it by an estimate of the matrix **S**.

The further results are based on the following expressions for the third central moment of the projection of mixture (2) by some vector V

$$\mu_3(V) = \sum_{i=1}^{k} a_i (M'_i V)^3 \qquad (7)$$

and it's estimate

$$\hat{\mu}_3(V) = (1/n)\sum_{j=1}^{n} (V' X_j)^3 \,. \qquad (7')$$

Evidently

$$E\hat{\mu}_3(V) = c\mu_3(V)\,, \qquad (8)$$

where constant c=$c(n)$ and Ez is the average of random value z.

Lemma 2. Let V be a p-dimensional vector, **H** is any positively defined matrix with sizes $p * p$.

Consider the vectors

$$\hat{U}(V) = (1/n)\sum_{j=1}^{n} (V' X_j)^2 X_j \qquad (9)$$

$$\hat{U}(\mathbf{H}) = (1/n)\sum_{j=1}^{n} (\dot{X}_j \mathbf{H} X) X_j \qquad (10)$$

Then

$$E\hat{U}(V) = 3c\sum_{i=1}^{k} a_i (M'_i V)^2 M_i \qquad (11)$$

$$E\hat{U}(\mathbf{H}) = \sum_{u=1}^{k} a_i (M'_i \mathbf{H} M_i) M_i = cU(\mathbf{H}) \qquad (12)$$

Thus the means of the vectors of (9) and (10) belong to the **M**.

Proof. Let us take the derivatives from both parts of expression (7') with the respect to *V* and then use averaging *E* for them. Since differentiation is a linear operation we can change the order of operations of averaging and differentiation for the left part of expression (7') that entails equality (11). Equality (12) may be proved if we additionally consider that the vector $VN_p(O,\mathbf{H})$ and use proper average.

7. The case when the means of components are placed on the same direct line (q'=1).

This case is interesting in several applications. On the other hand, we can obtain some results about the proposed estimates when the dimension *p* is great. In this case we can write $M_i = b_i D \ (\sum a_i b_i = 0)$ and the vectors *U*(**H**) and *D* are collinear. Suppose that the mixture (2) components densities are normal and matrix $\mathbf{W} = \mathbf{I}_p$. Let the ϕ be the angle between $\hat{U}(\mathbf{H})$ and D and $\gamma = \cos(\phi)$.

*Lemma 3.*Let $n \to \infty, p \to \infty, \eta = p/n > 0$. Then

$$E\gamma = \mu_3(L)/[\mu_3^2(L) + (p/n)(15 + \mu_4(L) - (L'SL)^2 + 2(L'SL))]^{1/2} + O(1/r),$$
$$E(\gamma - E\gamma)^2 = O(1/r),$$

where $L = D/\|D\|$;

r is *n* or *p*; $\mu_4(L)$ is the forth moment of the $z = L'X$.

Remark: The case when $\mathbf{W} = \mathbf{I}_p$ may be reduced by the linear transformation of the variables to the one considered above. The scalar product of *D* and *U* becomes $D'\mathbf{W}^{-1}U$.

If *k*=2 the main term of $E\gamma$ transforms in

$$E\gamma = A/(A + (p/n)B))^{1/2}$$

where $A = a(1-a)(1-2a)\tau^3, B = 19 + 6a(1-a)\tau^2 + a(1-a)(1-2a)^2\tau^4$

τ is Machalanobis distance between the mixture components. Values γ for several sets (a,τ) give in the table below (the ratio $\eta = 0.1$)

a	τ	$E\gamma$
0.2	2.5	0.67
0.2	2	0.45
0.05	4	0.83

8. Some consequences of violating the probability model assumptions.

Let us discuss some consequences of violating the condition of equality of the within-covariance matrices **W** of mixture (2) components. Consider the simple example of the density of two-variate vector *X*

$$p(X) = an_2(M_1, \mathbf{W_1}; X) + (1-a)n_2(M_2, \mathbf{W_2}; X), (0 < a < 1)$$

where $n_q(M, \mathbf{W}, X)$ is the normal density of *q*-variate vector with means vector *M* and covariance matrix **W**;

$$M'_1 = (0, -m), \qquad M'_2 = (0, m);$$

$$\mathbf{W_1} = \begin{bmatrix} 1/a - (1-a)/c & 0 \\ 0 & 1/\psi \end{bmatrix}, \qquad \mathbf{W_2} = \begin{bmatrix} a/c & 0 \\ 0 & 1/\psi \end{bmatrix}$$

$$m^2 = (1 - 1/\psi)/2a(1-a), (\psi > 1)$$

It is easy to check that the *X* have the means vector *M'*=(0,0) and covariance matrix $\mathbf{S} = \mathbf{I_2}$. For the first coordinate axis the induced density is

$$f_1(z) = an_2(0, 1 - a(1-a)c; z) + (1-a)n_2(0, a(1-a)/c; z);$$

and for the second

$$f_2(z) = an_2(-m/, 1/\psi; z) + (1-a)n_2(m/, 1/\psi; z).$$

Let us calculate the values of PI (4) under $\beta = 1$. Note that $f_1(z)$ does not depend on ψ and $f_2(z)$ does not depend on the value of *c*. Therefore:

- for any value of *c* there exists such $\psi_0 = \psi_0(c)$ that if $\psi > \psi_0$ the projection of the second coordinate axis gives maximum to $Q_1(U_2, X)$, i.e. optimal projection vector $U'_2 = (0,1)$;
- for any value of ψ there exists such $c_0 = c_0(\psi)$ that if $c > c_0$ the projection of the first coordinate axis gives maximum to $Q_1(U_1 X)$, i.e. optimal projection vector $U'_1 = (1,0)$.

Therefore, the value of PI (4) will increase also due to the increase (or decrease) of the variances ratio $w_1^2(U)/w_2^2(U)$.

The projection with the vector U_2 gives visual pattern of two clearly discernable clusters if *m* is large enough. At the same time if $c > c_0(m)$ the projection of U_1 will be selected by condition of the PI maximum. But this projection gives visual pattern rather like a dense kern with outliers.

8. Choosing the projection for detecting outliers.

As the PI suitable for choosing the projections vectors for detecting outliers we propose the ratio

$$Q(U, \mathbf{X}^{(n)}) = s^2(U)/s_r^2(U), \qquad (13)$$

where $s^2(U)$ is the usual estimate of the variance of the projection with the vector *U*, and $s_r^2(U)$ is robust estimate of it.

It is known that the usual estimate of the s^2 is rather sensitive to outliers and if they are present its value increases as a rule. So, the projection where the value of the PI (13) is maximal may be reasonably considered as the one where the influence of outliers is the most expressed. An approximate maximization of the PI (13) can be achieved by the solution of the generalized eigenvectors problem

$$(\mathbf{S} - h\mathbf{S}_r)U = 0, \tag{14}$$

where $\mathbf{S}$ is the routine estimate of covariance matrix and $\mathbf{S}_r$ is some robust estimate of it.

For visual analysis it is necessary to use the projection with maximal eigenvalues.

The index (14) may be improved if we take into account the difference between the location parameter estimates (usual and robust). For example, we can use index

$$Q(U, \mathbf{X}^{(n)}) = (s^2(U) + \|M - M_r\|^2) / s_r^2(U)$$

where M is the usual estimate of the vector and M_r is some robust estimate of it.

Approximated solution may be again defined as solution of the generalized eigenvalues (vectors) problem

$$(\mathbf{S} + (M - M_r)(M - M_r)' - h\mathbf{S}_r)U = 0$$

The estimates $\mathbf{S}_r, \mathbf{M}_r$ that we use have been proposed by Meshalkin (1970) and they are solution of the following set of equations system

$$M_r = \sum_{i=1}^{n} w(t_i) X_i / \sum_{i=1}^{n} w(t_i)$$

$$\mathbf{S}_r = (1+\lambda) \sum_{i=1}^{n} w(t_1)(X_i - M_r)(X_i - M_r)' / \sum_{i=1}^{n} w(t_i), \tag{15}$$

where $t_i^2 = (X_i - M_r)' S_r^{-1} (X_i - M_r)$,

$$w(t_i) = \exp(-\lambda t_i^2 / 2).$$

It is possible to use an iterative procedure for seeking the solution of equations (15).

The factor $1+\lambda$ affords the unbiasedness of the estimate $\mathbf{S}_r$ when the underlying distribution of the *X* is in fact normal (Meshalkin (1970)). Therefore, in such case the eigenvalues are close to 1. The values λ that we use are $c/(p+1), 1<c<4$.

This approach gives the weights $v_i = w_i / \sum w_i$ for every observation X_i. These weights can be used for several purposes: automatical removal of the observations with the low weights (suspected as outliers), marking such ones on visual display, obtaining robust scale estimates for calculation of PI values (Section 4).

9. Projection pursuit for discriminant analysis DA.

Assume that the p-variate sample X is divided into the k subsets $\mathbf{X}_1,...,\mathbf{X}_k$ with the sizes $n_i (i = 1,...,k), \sum n_i = n$.

The classical DA is based on the assumption that the mixture model (2) is true and the vectors from the group $\mathbf{X}_i$ are vectors corresponding to the i-th component of the mixture.

On these assumptions if k=2, the PI for selecting of the most interesting one –dimensional projection is Fisher ratio

$$Q(U, \mathbf{X}^{(n)} = (U'\overline{X}_1 - U'\overline{X}_2)^2 / W^2(U), \qquad (16)$$

where $\overline{X}_i$ is means vector of the $\mathbf{X}_i$.

Robust variants of the PI may be obtained if in (16) we replace the values $X_i, W^2(U)$ by some robust analogs (Huber (1985)).

If k>2 we can use the canonical vectors V_i (see Section 2) as the vectors of interesting projections.

When the assumptions of classical model of DA are violated even in robust it is necessary to use PI that based on more complete information about distributions of vectors in $\mathbf{X}_i$.

If k=2 such a PI is for example

$$Q(U, \mathbf{X}^{(n)}) = \gamma(w_1, w_2) \int \hat{f}_1(z) \hat{f}_2(z) dz$$

where w_i^2- is variances estimate of one- dimensional projection of i-th component and $\hat{f}_i(z)$ is induced density estimates.

Now, the most interesting projection may be obtained as the solution of minimization problem

$$U = \underset{Y \in R^p}{\arg\min}\, Q(Y, \mathbf{X}^{(n)})$$

References.

Buchschaber V.M., Maslov V.K. (1975) Factor analysis on manifolds and problem of selecting the variates in pattern recognition. Proceedings of Ac. Sc. of USSR, ser. “Engineering Cibernetika”, N 6 , p. 194-200 (in Russian).

Buchschaber V.M., Maslov V.K. (1977) Factor analysis and extremal problems on Grassman manifolds. In “Mathematical methods of solving the economical problems”. Moscow: Science, №7, p. 85-102 (in Russian).

Cressie N. (1978) The power results for tests based on high – order gaps. Biometrika, 65, p.214-218.

Friedman J.H., Tukey J.W.(1974) A projection pursuit algorithm for exploratory data analysis. IEEE Trans. Comput., C-23, p. 881-889.

Friedman J.H.(1987) Exploratory projection pursuit. J. Amer. Stat. Ass., 82 p.249-266.

Huber P.J. (1985) Projection pursuit. Ann. Statist., 12, p.433-474.

Jones M.C., Sibson R . (1987) What is Projection Pursuit. J. Royal Stat. Soc., 150, Ser. A, p.1-38.

Meshalkin L.D. (1970) Approximation of multidimensional densities by normal distributions.-7-th International Biometric Conference, Hannover.

Maronna R.A. (1976) Robust M-estimators of multivariate location and scatter. Ann. Statist. 4, p.51-56.

Rao C.R. (1965) Linear statistical inferences and it’s applications. New-York: Wiley.

Renyi A. (1970) Probability Theory. Amsterdam: North Holland.

Vacicek O. (1976) A test of normality based on Sample Entropy. J. Royal Stat. Soc. Ser. B, 38, p.54-59.

Yenyukov I.S., Neuschtadt (1978) Estimating the discriminant subspace when the sample is unclassificated. In “Applied Multivariate Statistical Analysis”, Moscow: Science, p.326-333.

Yenyukov I.S. (1986a) Methods, algorithms and programs of the multivariate statistical analysis (in Russian). Moscow: Financy & Statistica.

Yenyukov I.S. (1986b) Detecting outliers and clusters in multivariate data based on projection pursuit. In “Proceedings of the World Congress of Bernoully Society, Tashkent, USSR, Sept. 1987. Utrecht: VHU Science Press, 2, p.137-141.