Composants principaux généralisés pour la définition de directions révélatrices en discrimination et régression

Composants principaux généralisés pour la définition des directions révélatrices en classification et régression

Igor Enyukov

StatPoint Ltd.
121069 Moscow, Novinsky bld. 16A, 14
igen@online.ru

RÉSUMÉ. *Des indices, basés sur le rapport de deux formes quadratiques définies positives, sont proposés pour la construction de projections révélatrices en discrimination et régression. Le problème d'optimisation de ces indices se réduit simplement à une analyse en composants principaux généralisés. Un exemple d'application est présenté.*

ABSTRACT. *Projection pursuit indexes based on a ratio two positive-defined quadratic forms are suggested for using in PP classification and regression. It allows us to reduce the problem of optimization of PP indexes to generalized principal components problem. An example of application of the approach to PP regression is given.*

MOTS-CLÉFS **:** *directions révélatrices, analyse exploratoire de données, indices de directions révélatrices pour l'analyse discriminante, indices de directions révélatrices pour la régression, composants principaux généralisés.*

KEYWORDS : *projection pursuit, exploratory data analysis, projection indexes for discriminant analysis, projection indexes for regression analysis, generalized principal components*

1. Introduction

Realization of projection pursuit (PP) approach is connected with the following problems: to design an appropriate projection index and to solve the problem of the index optimization. The optimization stage can be not so simple due to a number of local extremums. Reducing a PP problem to a generalized principal components, on the one hand, implicates new projection indexes and, on the other hand, allows us to avoid the local extremum problem. For this aim the indexes should be designed as a ratio of two positive defined quadratic forms. Then instead of solving an optimization problem, we can use generalized eigenvectors approach (Yenyukov, 1988, Caussinus *et al.*,1990, Yenyukov, 1992) which always leads to the global extremum. In this paper the approach is applied to exploratory PP in discriminant and regression analysis.

2. Projection index for discriminant analysis problem

In the discriminant analysis situation there are M classes of objects, which are presented by M training sets of p-dimensional vectors. The problem is to get some decision (classification) rule which allows us to classify p-dimensional vectors to one of the M classes with minimal error. If p-dimensional objects of each class follow p-dimensional normal distribution and all within-class covariance matrices are similar (so, classes are different only by their mean values vectors) the best way is to use canonical discriminant analysis (CDA). In particularly, for data projecting with keeping the between-class structure the canonical projection vectors $V_1,...,V_{M-1}$ can be used. They are defined as eigenvectors of the generalized eigenvectors problem $(\mathbf{S}-l\mathbf{W})V=0$, where **S** is total covariance matrix (or an estimate of it), and **W** is within classes covariance matrix(or an estimate). There are min(M, p-1) eigenvalues which are more than 1.

Below we give a description of some generalization of the approach when the CDA conditions are not true.

Assume the size of *i*-th (*i*=1,2,…,M) group is n_i and define the number of the nearest neighbors(NN) ***k***. Fix *i*-th training set. For each object X_j of this set look for its *k* NN. We will look for them through all the training sets until the number *k* is reached or an object from another training set (not *i*-th) is found. So the real number of the found NN (k_j) can be less than *k,* but all of them are from *i*-th training set. Estimate a covariance matrix $\tilde{\mathbf{W}}_{ij}$ as follows

$$\tilde{\mathbf{W}}_{ij}=\sum_{l=1}^{k_j}(X_{jl}-X_j)'(X_{jl}-X_j)\text{, where } X_{jl} \text{ is } l\text{-th NN of } X_j .$$

Then define kNN-based within-group covariance matrix of *i*-th set as the sum

$$\tilde{\mathbf{W}}_i=\sum_{j=1}^{n_i}\tilde{\mathbf{W}}_{ij}/\mathrm{k(i)}\text{, where k(i) is the total number of NN for objects from } i\text{-th set.}$$

Repeating the procedure for all the training sets we have as a result the pooled kNN-based within-covariance matrix

$$\tilde{\mathbf{W}} = (\tilde{\mathbf{W}}_i + \ldots + \tilde{\mathbf{W}}_M)\,/\,\mathrm{M}$$

The projection vectors $U_1, U_2 \ldots$ we are interested can be found now as solution of the following generalized eigenvectors problem (with maximal eigenvalues)

$$(\mathbf{S} - l\tilde{\mathbf{W}})U = 0$$

The projections keep the kNN-topology found in p-dimensional space in the projection (one-, two-, three-dimensional) space. In fact the vectors are solutions of the problem of maximizing the ratio of two quadratic forms

$$\lambda = U'\mathbf{S}U \,/\, U'\tilde{\mathbf{W}}U$$

The numerator of the ratio is proportional to the sum of the distances between-all pairs of the points projected from p-dimensional space on vector U . At the same time the denominator is proportional the sum of the distances between projections of NN.

After the projections are defined they can be used for visual inspection of between class relations and for designing a decision rule by means of a kind of interactive graphic procedure. Another possibility is to create a classification rule using the projections as input variables of a neural network.

3. PP regression

The method for approximating a regression function with the help of PP was proposed in (Friedman J.H. *et al.*, 1981). Assume we have a sample of size *n* from $(p+1)$ -dimensional distribution (y_i, X_i) $(i = 1,\ldots,n)$ and we want to fit a regression function of variable *y* on *p* components of vector *X* in the form:

$y = \sum_{i=1}^{m} g_i(U_i'X) + \varepsilon$, where $g_i()$ are unknown functions, U_i unknown vectors, *m* the number of vectors. Here we will regard the case *m* = 1.

To apply the above developed method to the regression problem let's form two artificial classes (groups) by using the dependent variable *y* as a grouping variable.

The first class is defined as the set of objects for which *y*-value is more than its mean value (*y-m(y)* > 0), and the second class is formed by such vectors for that *y*-value is less than its mean value (*y-m(y)* < 0). Of course, it is possible to organize any reasonable number of groups using *y* as a grouping variable.

Then the above suggested PP procedure is applied. The obtained projection vector can give a more or less suitable solution of the regression problem (see example below). But in any case it can be used as a reasonable starting point for an optimizing procedure.

4. Example

A data matrix consisting of 200 cases and 9 variables was generated. The first variable $ymod$ is related to the variables $x_1,...,x_8$ by the following regression equation: $ymod = \sin(a_1x_1 + ... + a_8x_8) + \varepsilon$.

The noise component ε is chosen so that the theoretical value of the nonparametric determination coefficient is approximately equal to 0.7. The usual (linear) determination coefficient is zero. Figure 1 shows the curves produced by the smoothing of scatter-plots for the linear least squares regression method (left plot, *y* versus the predicted values) and by the PP-regression procedure based on the suggested approach (right plot, *y* versus linear combination $(U_1'X)$).

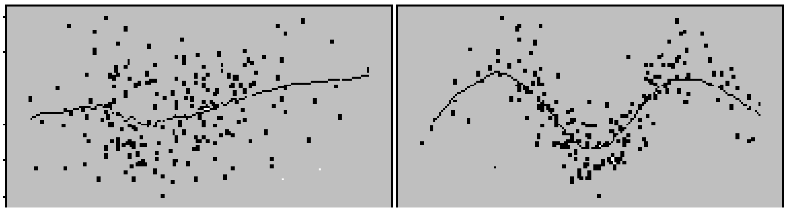

Fig. 1 Screen for the PP-procedure for simulated data

5. References

Caussinus H., Ruis A. "Interesting projections of multidimensional data by means of generalized principal components analysis", *Proceeding of COMPSTAT 90*, 1990, p. 121-126.

Friedman J.H., Stuetzle W. "Projection pursuit regression" *JASA,* 76, 1981, p.817-823.

Friedman, J.H. "Exploratory projection pursuit", *JASA*, 1987, 82, p.249-266.

Yenyukov, I.S. "Detecting Structures by Means of Projection Pursuit", *Proceeding of COMPSTAT 88*, 1988, p. 48-58.

Yenyukov I. "Gradient Filtering Projections for Recovering Structures In Multivariate Data", *Information and Classification. Proceedings of the 16th Annual Conference of the "Gesellshaft fur Klassification e.V",* University of Dortmund, April 1-3,1992. Springer-Ferlag, p.214-218