Pattern classification using principal components regression

Ciuiu, Daniel
Technical University of Civil Engineering, Bucharest, Romania, Romanian Institute for Economic Forecasting

January 2008

PATTERN CLASSIFICATION USING PRINCIPAL COMPONENT REGRESSION

Daniel CIUIU
Technical University of Civil Engineering, Bd. Lacul Tei, no. 124, Bucharest, Romania
E-mail: dciuiu@yahoo.com

Abstract: In this paper we will classify patterns using an algorithm analogous to the *k*-means algorithm and the principal components regression (*PCR*).

We will also present a financial application in which we apply *PCR* if the points represent the interests for accounts with different terms.

Mathematics Subject Classification (2000): 62J05, 62H25, 68T10
Keywords: Principal components regression, pattern classification, *k*-means

1. Introduction

Let be n points in $\mathbf{R}^p$: $X^{(1)},\ldots, X^{(n)}$. The orthogonal linear variety of the dimension k ($0<k<p$) is that linear variety with the minimum sum of the squares of Euclidean distances. We know (see [4]) that this linear variety is generated by the eigenvectors of the sample covariance matrix corresponding to the first maximum k eigenvalues, and contains the gravity center of the given n points. These eigenvectors are called principal components, and for that the orthogonal regression is called also principal components regression (*PCR*). The principal components analysis is used in [5] to simplify the computations in the discriminant analysis by using the Kolmogoroff distance.

For n points from $\mathbf{R}^p$ we can find the orthogonal regression linear variety of the dimension k (we use the first k principal components). But in this case all the n points are in the same class. A modality to classify n points from $\mathbf{R}^p$ in k classes is to use the k-means algorithm (see [2]). First each class has only one point, which represents the class. The other points are introduced next into the class represented by the nearest point (the center of gravity of the points from the given class), and we compute the new center of gravity of this class. The next step is to check for each point if the distance to the center of gravity of its class is minimum. Otherwise we move the point from the current class such that the distance becomes minimum. We compute the centers of gravity for the source class and destination class, and the algorithm stops when no point is moved from its class.

2. The k-means algorithm and principal components regression

In the k-means algorithm the classes are given by their gravity centers Y_i, $i=\overline{1,k}$. These points minimize the sum

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-Y_i\right)^2, \tag{1}$$

where X_{ij}, $j=\overline{1,n_i}$ are the n_i points from $\mathbf{R}^p$ that are classified into the class i by the k-means algorithm.

In the same manner we can classify paterns from $\mathbf{R}^p$ using principal components regression, more exactly the first j with $0<j<p$ principal components. In this case each class has at least $j+1$ points, initially exactly $j+1$ points. The other points are classified first in the

class with the less Euclidean distance.

After the first classification, we take each point and if we have a distance less than those to the current class, we move the point to the new class. The algorithm stops when all the points are not moved.

When we add a point to a class we compute again the orthogonal linear variety for this class. If we move a point from a class to another one, we have to compute again the orthogonal linear variety for both classes (those from we move and those in which we move the point).

For this algorithm we have to compute the sample covariance matrices for the classes, and their eigenvectors and eigenvalues are computed using the Jacobi rotations method.

3. A financial application

For the following application X_1 is the annual interest for an account without term, X_2 is the annual interest for an account with the term one month, X_3 is the annual interest for an account with the term *3* months, X_4 is the annual interest for an account with the term *6* months, X_5 is the annual interest for an account with the term *9* months and X_6 is the annual interest for an account with the term one year. Consider *29* banks as follows.

Bank	X_1	X_2	X_3	X_4	X_5	X_6
ABN-Ambro Romania	*0.25%*	*3.5%*	*3.75%*	*3.75%*	*0*	*3.75%*
AlphaBank	*0.1%*	*6.25%*	*6.5%*	*7%*	*7%*	*7.25%*
Banc Post	*0*	*7.25%*	*7.25%*	*7.15%*	*0*	*7.15%*
Banca Comercială Carpatica	*1%*	*7.5%*	*7.55%*	*7.6%*	*7.75%*	*7.8%*
BCR	*0.25%*	*6%*	*6.25%*	*6.5%*	*6.75%*	*7.5%*
Banca Italo-Romena	*0*	*5.5%*	*5.75%*	*6%*	*6.15%*	*6.25%*
Banca Românească	*0.75%*	*7.3%*	*7.75%*	*8.05%*	*8.1%*	*8.1%*
Banca Transilvania	*0.25%*	*7.5%*	*7.5%*	*7.5%*	*7.75%*	*7.75%*
Bank Leumi Romania	*0.25%*	*7.5%*	*7.5%*	*7.75%*	*7.75%*	*8%*
Blom Bank Egypt	*0.1%*	*6%*	*6.5%*	*6.5%*	*6.75%*	*7%*
BRD-Groupe Société Générale	*0.25%*	*5.5%*	*5.6%*	*5.65%*	*5.65%*	*5.75%*
C.R. Firenze Romania	*0.1%*	*6.5%*	*6.75%*	*7%*	*7.25%*	*7.5%*
CEC	*0.25%*	*7%*	*7%*	*7.25%*	*0*	*7.25%*
Citibank Romania	*1%*	*4.28%*	*4.28%*	*4.28%*	*3.87%*	*3.46%*
Emporiki Bank	*0.5%*	*6.75%*	*7%*	*7.25%*	*7%*	*7%*
Finansbank	*0.1%*	*7.5%*	*8%*	*8%*	*8%*	*8.5%*
HVB-Ţiriac Bank	*0.1%*	*6.4%*	*6.3%*	*6.2%*	*6.1%*	*6.1%*
ING Bank	*6.85%*	*5.5%*	*5.75%*	*6%*	*6.25%*	*6.5%*
Libra Bank	*0*	*8%*	*8.1%*	*7.6%*	*7.6%*	*8.5%*
Mind Bank	*0.25%*	*7%*	*7%*	*7.25%*	*7.5%*	*7.75%*
OTP Bank	*0.25%*	*6.25%*	*6.5%*	*7%*	*7%*	*7.25%*
Piraeus Bank	*0.5%*	*7%*	*7.1%*	*7.25%*	*7.1%*	*7.35%*
Pro Credit Bank	*7%*	*7.5%*	*7.65%*	*7.7%*	*0*	*7.85%*
Raiffeisen Bank	*0.25%*	*4%*	*4.25%*	*4.5%*	*4.6%*	*4.75%*
Romanian International Bank	*0.25%*	*6.5%*	*6.75%*	*7%*	*7.5%*	*7.75%*
Romexterra	*0.25%*	*7.5%*	*7.75%*	*7.75%*	*8.1%*	*8.1%*
San Paolo IMI Bank	*0.1%*	*6.5%*	*6.7%*	*6.8%*	*7%*	*7.2%*
Uni Credit Romania	*0.1%*	*5%*	*5%*	*5.25%*	*5.5%*	*5.5%*
Wolksbank	*0.1%*	*4.5%*	*4.75%*	*4.5%*	*3.5%*	*3.25%*

The orthogonal regression line is

d:{$0.04837+0.00197X_1+0.53627X_2-0.79983X_3+0.26942X_4+0.00536X_5-0.0085X_6=0$,

$-0.40228-0.00842X_1-0.48134X_2-0.03852X_3+0.83753X_4-0.00819X_5-0.26124X_6=0$,

$-0.71162+0.03716X_1+0.48786X_2+0.35302X_3+0.05034X_4+0.06596X_5-0.79316X_6=0$,

$1.64803+0.82789X_1-0.21195X_2-0.19145X_3-0.16583X_4+0.42717X_5-0.1518X_6=0$,

$-6.64018+0.54586X_1+0.34282X_2+0.33464X_3+0.32816X_4-0.48029X_5+0.36627X_6=0\}$ and the error is 191.00977.

If we consider *2* classes we obtain the orthogonal regression lines d_1:$\{0.03475+0.00086X_1+0.41303X_2-0.79344X_3+0.44436X_4-0.02473X_5-0.04238X_6=0$, $-0.24446-0.0014X_1-0.62098X_2+0.10689X_3+0.74741X_4-0.01053X_5-0.2103X_6=0$, $-1.69801+0.05743X_1+0.52864X_2+0.41568X_3+0.17852X_4-0.07815X_5-0.71167X_6=0$, $-2.53695+0.05594X_1+0.2102X_2+0.25308X_3+0.25662X_4-0.78251X_5+0.45877X_6=0$, $-0.52383+0.9967X_1-0.03907X_2-0.03287X_3-0.01913X_4+0.0563X_5+0.02123X_6=0\}$ and d_2:$\{-5.55656-0.03768X_1+0.63086X_2-0.24548X_3+0.67669X_4-0.02752X_5-0.28577X_6=0$, $-1.2405-0.00603X_1-0.10724X_2+0.77483X_3+0.12312X_4+0.00681X_5-0.61065X_6=0$, $-1.87204-0.042X_1-0.64434X_2-0.01971X_3+0.70703X_4-0.04215X_5-0.23416X_6=0$, $-12.70244-0.12386X_1+0.38094X_2+0.57381X_3+0.13034X_4-0.14573X_5+0.68706X_6=0$, $-5.17038+0.94024X_1+0.05545X_2+0.08592X_3+0.09078X_4+0.2912X_5+0.11155X_6=0\}$, the classes C_1={*ABN Ambro Romania, Alpha Bank, BCR, Banca Italo-Romena, Blom Bank Egypt, BRD-Groupe Société Générale, C.R. Firenze Romania, Citibank Romania, Emporiki Bank, HVB-Țiriac Bank, ING Bank, Mind Bank, OTP Bank, Piraeus Bank, Raiffeisen Bank, Romanian International Bank, San Paolo IMI Bank, Uni Credit Romania, Volksbank*} and C_2={*Banc Post, Banca Comercială Carpatica, Banca Românească, Banca Transilvania, Bank Leumi Romania, CEC, Finansbank, Libra Bank, Pro Credit Bank, Romexterra*} and the error 84.49813.

If we consider *5* classes we obtain the orthogonal regression lines d_1:$\{-2.30278-0.01425X_1-0.62292X_2+0.59308X_3+0.30341X_4-0.27928X_5+0.29994X_6=0$, $-0.20254+0.93486X_1+0.05054X_2+0.25055X_3-0.07438X_4+0.04281X_5-0.23096X_6=0$, $-3.67379-0.11524X_1+0.62336X_2+0.50028X_3-0.33188X_4-0.42384X_5+0.24098X_6=0$, $-3.23606+0.2531X_1+0.27791X_2-0.40285X_3+0.62327X_4-0.41266X_5-0.37103X_6=0$, $-1.20155-0.22007X_1+0.23705X_2+0.28491X_3+0.5425X_4-0.08105X_5-0.71467X_6=0\}$, d_2:$\{X_5=0$, $-0.12781+0.01422X_1+0.7146X_2-0.69939X_3+0.00108X_4+0.00138X_6=0$, $-9.04924-0.08943X_1+0.14685X_2+0.1497X_3+0.97363X_4-0.00869X_6=0$, $-8.60955-0.10667X_1+0.13211X_2+0.13467X_3-0.04172X_4+0.97534X_6=0$, $-6.7629-0.0409X_1+0.66957X_2+0.68257X_3-0.21146X_4-0.19846X_6=0\}$, d_3:$\{-3.10123-0.32265X_1+0.01865X_2+0.03342X_3+0.02148X_4+0.93412X_5-0.14623X_6=0$, $-1.21571-0.01912X_1+0.90218X_2-0.35507X_3-0.2439X_4-0.00779X_5-0.00946X_6=0$, $-1.05915-0.37093X_1-0.10422X_2-0.18675X_3-0.12004X_4+0.02357X_5+0.89538X_6=0$, $-0.82081-0.07122X_1-0.00005X_2-0.56058X_3+0.82377X_4-0.029X_5-0.03523X_6=0$, $-6.71939-0.31954X_1+0.39497X_2+0.70794X_3+0.45481X_4-0.13435X_5+0.12576X_6=0\}$, d_4:$\{-1.9058+0.078X_1+0.00401X_2-0.00456X_3-0.57944X_4+0.81123X_5-0.00656X_6=0$, $-1.87683-0.00754X_1-0.00069X_2+0.00025X_3-0.38536X_4-0.26737X_5+0.88315X_6=0$, $-3.14662-0.02419X_1-0.00226X_2+0.95173X_3-0.22277X_4-0.15259X_5-0.14388X_6=0$, $-11.18819-0.03841X_1+0.96176X_2+0.08416X_3+0.18495X_4+0.13249X_5+0.12122X_6=0$, $6.39659+0.80545X_1+0.18929X_2-0.15124X_3-0.34991X_4-0.33115X_5-0.24586X_6=0\}$ and d_5:$\{-0.07787-0.28138X_1+0.19615X_2-0.55873X_3+0.62899X_4-0.39359X_5+0.14366X_6=0$, $0.04796+0.47872X_1+0.52896X_2-0.52138X_3-0.38659X_4+0.12704X_5+0.23152X_6=0$, $-0.20828-0.54277X_1+0.28241X_2-0.23547X_3-0.06916X_4+0.57523X_5-0.48428X_6=0$, $-0.49119+0.62627X_1-0.14361X_2-0.1017X_3+0.51736X_4+0.27532X_5-0.48306X_6=0$, $0.4072+0.0693X_1+0.61565X_2+0.37429X_4-0.49283X_5-0.48213X_6=0\}$, the classes C_1={*ABN Ambro Romania, Alpha Bank, BCR, Blom Bank Egypt, C.R. Firenze Romania, OTP Bank, Romanian International Bank*}, C_2={*Banc Post, CEC, Pro Credit Bank*}, C_3={*Citibank Romania, ING Bank, Volksbank*}, C_4={*Banca Comercială Carpatica, Banca Românească, Banca Transilvania*} and C_5={*Banca Italo-Romena, Bank Leumi Romania, BRD-Groupe Société Générale, Emporiki Bank, Finansbank,* *HVB-Țiriac Bank, Libra Bank, Mind Bank, Piraeus Bank, Raiffeisen Bank, Romexterra, San Paolo IMI Bank, Uni Credit Romania*} and the error 3.09442.

The orthogonal regression hyper-plane is H: $0.04837+0.00197X_1+0.53627X_2-0.79983X_3+0.26942X_4+0.00536X_5-0.0085X_6$ and the error is 0.23192.

If we consider *2* classes we obtain H_1: $-0.32534-0.0064X_1-0.65531X_2+0.75341X_3-0.05268X_4-0.01057X_5-0.00293X_6=0$ and H_2: $-0.0803-0.40475X_1-0.20348X_2-0.15505X_3+0.71788X_4+0.15299X_5-0.48164X_6=0$, the classes C_1={*ABN Ambro Romania, Alpha Bank, Banc Post, Banca Comercială Carpatica, BCR, Banca Italo-Romena, CEC, Emporiki Bank, ING Bank, Libra Bank, OTP Bank, Piraeus Bank, Pro Credit Bank, Romanian International Bank, Volksbank*} and C_2={*Banca Românească, Banca Transilvania, Bank Leumi Romania, Blom Bank Egypt, BRD-Groupe Société Générale, C.R. Firenze Romania, Citibank Romania, Finansbank, HVB-Țiriac Bank, Mind Bank, Raiffeisen Bank, Romexterra, San Paolo IMI Bank, Uni Credit Romania*}, and the error 0.03317.

If we consider *4* classes we obtain H_1: $-0.38315+0.08054X_1-0.66972X_2+0.73788X_3-0.00026X_4-0.0167X_5-0.0154X_6=0$, H_2: $0.09818+0.4909X_1+0.23129X_2-0.11737X_3-0.5868X_4-0.14256X_5+0.57191X_6=0$, H_3: $-1.04913+0.07788X_1-0.08447X_2+0.88105X_3-0.23123X_4-0.0201X_5-0.39583X_6=0$ and H_4: $-0.32802+0.34105X_1-0.48925X_2+0.59427X_3-0.44424X_4+0.3X_5+0.06173X_6=0$, the classes C_1={*ABN Ambro Romania, Alpha Bank, Banc Post, Banca Comercială Carpatica, BCR, Banca Italo-Romena, Romanian International Bank, Volksbank*}, C_2={*Banca Românească, Banca Transilvania, Bank Leumi Romania, Blom Bank Egypt, BRD-Groupe Société Générale, C.R. Firenze Romania, San Paolo IMI Bank, Uni Credit Romania*}, C_3={*CEC, Citibank Romania, Emporiki Bank, Finansbank, HVB-Țiriac Bank, ING Bank, Romexterra*} and C_4={*Libra Bank, Mind Bank, OTP Bank, Piraeus Bank, Pro Credit Bank, Raiffeisen Bank*} and the error 0.00115.

4. Conclusions

The applied *k*-means algorithm finds the minimum of error because there exists a finite number of classifications, and when we move a point to another class we obtain a smaller error. The error is smaller even if we only move the point and we consider the same orthogonal regression linear varieties.

We can also remark that if we increase the number of classes the error decrease. We can explain this as follows. Suppose that at a given moment we have *k* optimal classes given by their orthogonal regression linear varieties of the dimension *d*. From some classes with at least $d+2$ points we can move $d+1$ points to a new class given by the linear variety of the dimension *d* containing these points. Even if we consider the previous *k* classes given by the same linear varieties, the error decrease.

If we increase the dimension *d* the error decrease due to the fact that the orthogonal regression linear variety with the dimension *d* is included in those with the dimension $d+1$, and the three perpendiculars theorem.

References

[1] Ciucu, G. and Craiu, V.: *Statistical Inference*, Didactic and Pedagogic Publishing House, Bucharest, 1974 (Romanian).

[2] Dumitrache, I., Constantin, N. and Drăgoicea, M.: *Neural Networks*, Matrix Rom, Bucharest, 1999 (Romanian).

[3] Petrehus, V. and Popescu, A.: *Probabilities and Statistics*, UTCB Publishing House, Bucharest, 1997 (Romanian).

[4] Saporta, G.: *Probabilités, analyse des donées et statistique*, Editions Technip, Paris, 1990.

[5] Mahjoub, S. and Saporta, G.: Une méthode de discrimination non paramétrique, *Revue de Statistique Appliquée*, **Vol. XLII, No. 2** (1994), 99-113.