Nonlinear Principal Component Analysis for Withdrawal of Employment Time Guaranty Fund

Li Weigang 1, Aiporê Rodrigues de Moraes 2, Shi Lihua 3
and Raul Yukihiro Matsushita 2

1 Department of Computer Science, C.P. 4466, The University of Brasilia,
2 Department of Statistics, The University of Brasilia,
Brasilia – DF, CEP: 70919-970, Brazil
{weigang, aipore, raul}@unb.br
3 Department of Aero/Astro, Stanford University, USA

Abstract. To improve the management of Employment Time Guaranty Fund (FGTS), a research in Brazil is conducted to analyze past data and anticipate future trends of this fund. In this paper, Nonlinear Principal Component Analysis (NLPCA) is used to reduce data dimension in describing various causes of withdrawal from FGTS. Properties of these withdrawals are analyzed. Nonlinear time series corresponding to each cause of withdrawal in 75 months from 1994 to 2000 are collected from administrator of FGTS. Using NLPCA, 17 small quantity time series (group 1) are combined into one series and then combine with other 7 middle quantity series (group 2) to form another series. Finally, four combined time series (group 3) are formed which can well represent features of totally 27 kinds of withdrawals with respect to different causes. As a criterion of dimension reducing, the correlation coefficient of group 1 is 0.8486 and that of group 2 is 0.9765.

1 Introduction

The Employment Time Guaranty Fund (FGTS) is a kind of severance pay fund administered by Brazilian Federal Government. Under the FGTS employers deposit 1/12 of a worker’s salary into a restricted bank account, the balance of which is released to worker if and when the worker is fired without good reasons. In February of 2001, there are almost 63 million accounts with the total of 40 billion US$ in FGTS[1, 2].

Under the FGTS system, the worker can use the fund with variety of reasons (27 at this moment). It makes little difference whether a company fired an employee or not because severance pay had already been deposited every month. To counteract this undesired trend, employers are now required to make an additional payment equivalent to 40% of all previous payments to the employee’s account every time they want to fire someone. Other reasons to withdraw their fund balance include financing a home under a government-sponsored housing program, SIDA/AIDS patients and etc.
To improve the administration of FGTS, an Actuarial System of FGTS is proposed with the general objective to establish a technical base for the actuarial evaluation of FGTS. The development of system starts from the analysis of the data and other possible information. From the actuarial technical description and detailed diagnosis of the data, it will be possible to evaluate the future behavior of the fund in terms of the stability of the balance between incomes (employer’s deposits, interests of financial application etc.) and expenses (employee’s withdrawal, administration fee and 3% annual interests for every account etc.).

As an important factor which affects the balance of FGTS, employee’s withdrawal should be well represented in the Actuarial System of FGTS. However, according to various causes and long time records, there are a huge amount of data need to be taken account in the system. To reduce the required data dimension, we analyzed the part of the employee’s withdrawal according to the reasons of dismissal, extinction or retirement and etc. Based on 27 time series of monthly withdrawal with respect to various reasons from 1994 to 2001, four time series were finally extracted to represent all the 27 series.

For dimension reduction, the most well known method is the principal component analysis (PCA) or Karhunen-Loeve expansion, which effectively approximates the data by a linear subspace using the mean squared error criterion [5, 6]. Another linear method, independent component analysis (ICA) is also proposed to extract linear feature combinations that define independent sources [7]. Because to represent data sets just by linear combination implies a potential oversimplification of the original data, with the advent of neural network (NN) model, nonlinear mapping was introduced to PCA problem and led to the occurrence of Nonlinear Principal Component Analysis (NLPCA) [8]. These dimension reduction methods have been widely used in multivariate time series analysis, such as signals processing, climate forecasting, financial analysis and etc [8-13].

In this paper, we use PCA and NLPCA to perform dimension reduction of time series of withdrawal in FGTS. The remainder of the text is organized as follows. Section 2 describes 27 reasons of withdrawal from FGTS. Section 3 gives a brief description of methods used in this paper. Analysis results with PCA and NLPCA for the combination of multivariate time series of FGTS are shown in Section 4. Finally, the conclusions are described in section 5.

2 Withdrawal of FGTS

Based on the Brazilian labor law, in principal, Consolidation of Labor (CLT) [1], there are 27 reasons to use FGTS for an employee. These reasons can be classified into following quaternaries:
- Dismissal payment (as the code of 01, 01s and 02);
- Extinction of company (03);
- Finish the contract (04, 04s);
- Retirement (05, 05a);
- Death (23, 23a);
Lecture Notes in Computer Science 3

- Illness of SIDA/AIDS, cancer and etc. (80 and 81);
- Judicial determination (88, 88p and 88r);
- Home finance (91, 92, 93, 95);
- Other reasons (06, 07, 10, 26, 27, 86, 87 and 87n);

We analyzed 27 time series which were collected from July of 1994 to September of 2000 [2]. Table 1 gives some statistical results for these withdrawals. Some time series of them are also shown in Figures 1 to 3. As an initial analysis, we will describe four main properties of the withdrawal of FGTS in this section.

Table 1. Properties of some withdrawal of FGTS [2]

<table>
<thead>
<tr>
<th>Code</th>
<th>Reason</th>
<th>Average use per operation (R$)</th>
<th>Average percent (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>In operation*</td>
<td>In value**</td>
</tr>
<tr>
<td>01</td>
<td>Dismissal without cause</td>
<td>864</td>
<td>65.22</td>
</tr>
<tr>
<td>02</td>
<td>Dismissal by unexpected reason</td>
<td>946</td>
<td>0.01</td>
</tr>
<tr>
<td>04</td>
<td>Finish the contract in a defined period</td>
<td>88</td>
<td>9.72</td>
</tr>
<tr>
<td>05</td>
<td>Retirement</td>
<td>3605</td>
<td>3.82</td>
</tr>
<tr>
<td>23A</td>
<td>Use for himself home (Type III)</td>
<td>539</td>
<td>0.01</td>
</tr>
<tr>
<td>80</td>
<td>SIDA/AIDS patients</td>
<td>1036</td>
<td>0.10</td>
</tr>
<tr>
<td>91</td>
<td>Home finance</td>
<td>8935</td>
<td>0.96</td>
</tr>
</tbody>
</table>

* average yearly account operations of 27 withdrawal are 100% from 07/1994 to 09/2000;
** average yearly value of 27 withdrawal are 100% from 07/1994 to 09/2000;

2.1 Variety of scales for the time series

The scales of 27 time series varies greatly. We can see in Table 1, the largest value, withdrawal for dismissal without cause (for code 01) takes 60.21% of all withdrawal, but the smallest value, withdrawal for dismissal by unexpected reason only takes 0.01%. The five main withdrawals (60.21% for code 01, 15.83% for code 05 and 13.15 for code 91, 92 and 93) take nearly 90% of all withdrawal values while other 22 reasons only take 10.81%. The unit of variates 01 is in million Reais (R$, Brazilian currency), but for some other varieties (such as 2, 27 and etc.) the unit is in Reais. This fact increases the difficulty to scale the input variety of the system. Therefore, we divided 27 series into 3 groups: a) Group 1 (small quantity) includes 17 series (01S, 02, 03, 04S, 05A, 06, 07, 10, 23A, 26, 27, 80, 81, 87, 88P, 88R, 95), in which, the average yearly operation number or value are less than 1% of total. Sum all of 17 items in Group 1 just takes 2.16% of all withdrawal value. b) Group 2 (medium quantity) includes 7 series (04, 23, 86, 87N, 88, 92, 93), in which, the average yearly operation number or value are more than 1% but less than 5% of
Sum of all items in the group 2 and the output of group 1 takes the 16.06% of all withdrawal value. c) Group 3 (large quantity) includes 3 series (01, 05, and 91), in which, the average yearly operation number or value are more than 7%.

Fig. 1. Deposit, withdrawal, difference between them and balance of FGTS, monthly data from 07/1994 to 09/2000 [2]

2.2 Non-linearity

Withdrawal variates show high order of non-linearity. This may also increase the difficulty to use conventional signal analysis methods to process the data. Figure 1 shows the deposit (DT), total withdrawal (TW), difference between DT and TW, and Balance of FGTS (there are still some other income to the balance which will not be discussed in this paper). Figure 2 shows the first four largest withdrawal series: 01, 05, 91 and 95, in which the values changes from R$ 5 million to R$ 200 million. Figure 3 shows the four smallest withdrawal series 01S, 02, 07, and 23A, in which the values change from R$ 0 to R$ 1 million. It is obviously very hard to derive a regular models for these variates so as to reduce the amount of data.
Fig. 2. The largest four withdrawal reasons: 01 (60.21% of average yearly of all withdrawal), 05 (15.83%), 91 (7.9%) and 92 (3.96), monthly data from 07/1994 to 09/2000 [2]

Fig. 3. The smallest four withdrawal reasons: 01S (0.02% of average yearly all withdrawal), 02 (0.01%), 07 (0.01%) and 27A (0.01%), monthly data from 07/1994 to 09/2000 [2]
2.3 Open system

FGTS is an open system. Its deposit, financial application and credit for public construction, withdrawal, administration fee and etc depend on the country’s social, economic and politics situation. For example, the dismissal caused withdrawal is closely related with the rate of employment. The SIDA/AIDS patient’s withdrawal has the close relationship with health condition of population, policies of government. Since 1996, Brazilian Ministry of Health has developed AIDS I [3] and AIDS II [4] projects to supply free treatment for patients of SIDA/AIDS, as the result of the politics, the death rate of AIDS case in Brazil reduces 12.5%. From 1996 to 1998, in 100 thousands population there are 14 AIDS patients. But in 1999, this number reduced to 11.2. This result also directly influences the withdrawal of FGTS (code 80) by these kinds of persons. The government’s cost are shown in Table 2. The average monthly withdrawal of this item is US$1.2265 million in 1995, US$1.1126 million in 1996, and reduced to US$1.0349 million in 1997. We can also observe that, the AIDS cases increase 13% from 1995 to 1996, 1.89% from 1996 to 1997, and reduced 54 cases from 1997 to 1998. We can get two main points from these results: 1) The action of government effectively reduced the AIDS case in Brazil; 2) Ministry of Health spends money to treat AIDS patients while FGTS’s withdrawal by patients of SIDA/AIDS is in a stable level. FGTS comes from the Ministry of Finance, therefore the government is saving money at some extent.

Table 2 The Cost of Ministry of Health for AIDS Patients and Withdrawal from FGTS

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AIDS cases in Brazil*</td>
<td>18424</td>
<td>20168</td>
<td>22745</td>
<td>23173</td>
<td>23119</td>
<td>18288</td>
<td>8596</td>
</tr>
<tr>
<td>Government treatment cost*</td>
<td>2.6758</td>
<td>4.5450</td>
<td>8.7667</td>
<td>9.4417</td>
<td>1.1079</td>
<td>5.130</td>
<td>7.2361</td>
</tr>
<tr>
<td>Withdrawal from FGTS by AIDS patients**</td>
<td>0.6716</td>
<td>1.2265</td>
<td>1.1126</td>
<td>1.0349</td>
<td>1.1620</td>
<td>0.8711</td>
<td>0.9991</td>
</tr>
</tbody>
</table>

*3, 4* [US$, million, average monthly.]  
**3, 4** [US$, million, average monthly; in 1994, from July to December; in 2000, from January to September. We do not analyze the withdrawal from FGTS by AIDS patients in 1999 and 2000 using USD, because in this period, the exchange rate from $R to $US is complicated.

2.4 Difference between quantity and value of the operation

Quantity of account operation by employee and value in account are not always in proportion. For example, withdrawal of finishing the contract (04) takes 9.72% of all the average yearly operations, but the value in account of these people just takes
0.93% of the whole (average yearly, Table 1). And average yearly operations for the retirement (05) take 3.82%, but the value of this account takes 15.83% of average yearly of all withdrawal, which is the second largest one. Intensive analysis of this combination is interesting for account administration in further data mining research.

Complicity of the properties of withdrawal means that we can neither merely discard any cause of the withdrawal nor make a simple combination to replace all of them. The main object of this paper is to reduce dimensions of withdrawal reasonably and preserve information as much as possible. The basic idea is to combine the series with small values into one, make it comparable with large groups while capable of reproducing the original small value series. Figure 4 shows the combination tree. There will be four series to describe the withdrawal of FGTS: 01, 05, 91 and combination of others. Every final series will take at lest 7% of the total withdrawal.

![Combination Tree](image)

**Fig. 4. Combination tree**

### 3 Methods of dimension reduction

Change of employees’ withdrawal with time and causes can be expressed in the form \(X(t) = [x_1(t), x_2(t), \ldots, x_m(t)]\), where \(t = 1 \sim n\), is the observation time, and \(x_i(t)\) \((i = 1 \sim m)\) is the monthly withdrawal according to \(m\) reasons. Dimension-reduction in pre-processing of this multivariate time series is to use a mapping \(g: \mathbb{R}^m \rightarrow \mathbb{R}^p\) \((p < m)\) to map \(X_{(max)}\) to a \(p\)-dimensional space and preserves as much information as possible at each point. To map back to the original space, a second mapping \(h: \mathbb{R}^p \rightarrow \mathbb{R}^m\) will be used. The optimal \(g\) and \(h\) are searched so as to ensure the reconstruction error \(\| X - h(g(X)) \|\) to be minimal. To do this, two methods are used in this paper: principal component analysis (PCA) and non-linear principal component analysis (NLPCA).
3.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a well-understood and useful method for modeling data sets. When applied to a $m$-dimensional data set $X$, it performs forward and backward mapping with linear transforms,

$$V = W^T X$$  \hspace{1cm} (1)

$$X^* = WV$$  \hspace{1cm} (2)

where $W = [w_1, w_2, ..., w_p]$, is the linear transform, $V$ is a $p$-dimensional feature vector representation of $X$, $X^*$ is the reconstructed $X$.

If vectors of $W$ are chosen to be the $p$ eigenvectors corresponding to $p$ largest eigenvalues of $X^T X$, then the approximation error

$$\|X - WW^T X\|$$  \hspace{1cm} (3)

will be minimized [6].

3.2 Nonlinear principal component analysis (NLPCA)

Nonlinear principal component analysis (NLPCA) also attempts to find mappings between a multidimensional data set and a lower-dimensional feature-space while minimizing reconstruction error, but allows the mappings to be nonlinear. In stead of using equation (1) and (2), NLPCA uses a nonlinear mapping function to get feature vectors

$$V = g(X)$$  \hspace{1cm} (4)

and another nonlinear mapping function to reconstruct the original data

$$X^* = h(V)$$  \hspace{1cm} (5)

Neural network is very suitable to realize such nonlinear mapping. In order to do this, an feed-forward neural network with three hidden layers is trained with its input and desired output both are the original data set $X$ [8]. The second hidden layer is designed to have lower dimension than the input and output layers. Therefore the output at this layer can be seen as a low-dimensional representation of the input $X$.

The advantage of NLPCA over PCA is that it is capable to represent and learn more general transformations. This is necessary in cases where one wishes to eliminate correlations between dimensions in a set of data or the lower-intrinsic dimensionality of a data set arises from a nonlinear relationship between different dimensions of the data set. However, NLPCA also has important disadvantages compared to PCA, such as the feature vector will no longer have a physical meaning like eigenvectors, more computation time and local optimal solution.
4 Dimension reduction for time series of monthly withdrawal

NLPCA were applied to reduce dimensions of 27 time series of withdrawal in FGTS. First, PCA is used to try to eliminate some unimportant variates. Then, NLPCA is used to combine the multivariates. Results show that NLPCA is superior to PCA in our applications.

4.1 Parameters and criteria

As mentioned in Figure 4, the sequence of the combination is that: from Group 1 of 17 time series to form 1 vector and then from Group 2 of 7 time series and the combination of Group 1 to form the final vector.

When using NLPCA, the architecture of the five-layer neural network is formed to extract 1D approximation to the data set [8, 11, 12]. In Group 1, the first (input) and fifth (output) layers each contain 17 neurons (Figure 5). Layers 2 and 4 are called the encoding and decoding layers respectively in which there are 5 neurons. The third layer is referred as the bottleneck layer which just contains a single neuron. The output of this neuron is what we need, 1D approximation of 17 vectors. We simply note this structure by 17:5:1:5:17. In Group 2, the architecture of the network is formed as 8:3:1:3:8.

![Fig. 5. The architecture of the five-layer neural network for NLPCA](image)

All of the time series are normalized by removing the mean and divided by the standard deviation of each variate. In case of groups of variates, they are normalized by the largest mean and standard deviation (or the second largest mean sometimes).

We use MSE_train (training mean square error), MSE_test (testing mean square error), \( J_{\text{train}} \) (minimum cost function \( J = \|X - X^*\|^2 \) from training) and \( J_{\text{test}} \) (minimum cost function \( J = \|X - X^*\|^2 \) from testing) as criteria to measure the quality of network training [13]. Usually, training is stopped if the MSE increases in the test data by more than the threshold. To avoid the local minima in the cost function, we also use an ensemble of optimization runs from random initial weight parameters. The best member of the ensemble selected as the solution. In this paper, we train network 20 times to select the best one result.

The criterion to verify the quality of the result of dimension reduction is based on the correlation coefficient between the output of the method (NLPAC etc) and the sum of all of varieties of that group.
4.2 Using PCA to reduce dimensions

As the first step, we use PCA to pre-process 17 varieties of Group 1. We take the vectors with the contribution more than 10% in Group 1, PCA eliminated 17 variates to 2 (see Figure 6(a)). If we increase that level to 20% or decrease it to 1%, PCA will give one vector or 15 vectors, respectively, to represent 17 variates.

Then we use PCA to pre-process 7 variates and the sum of Group 1 (17 time series). Figure 6(b) gives the 2 feature vectors that PCA extracted when the contributing level is set to be 10%. To get fewer feature vectors like one, the contributing level should be set to 20%. If we want to preserve more information and select a contribution level of 1%, the output of PCA shows 8 varieties. In this case, PCA can eliminate nothing.

With above results, we need not to mention the quality of processing using PCA for FGTS. We can not to eliminate 20% of information to get just one variety to represent both Group 1 and 2. If we want to preserve 99% information of the group, the quantity of varieties is more than 8 in both groups.

Fig. 6. PCA results for group 1 and 2 with 10% the rate of elimination. The blue line shows the sum of all of varieties of the group. a) 2 vectors to represent 17 varieties of group 1; b) 3 vectors to represent 8 varieties of group 2
4.3 Using NLPCA to reduce dimensions

There are 75 points of data for each of 27 time series. 80% of them are used to train the network and rest 20% are used for test. In order to avoid the local minimal problem, we train the NLPCA 20 times and use the one with the minimum MSE as the best feature vectors.

4.3.1 NLPCA for Group 1

By using a 17:5:1:5:17 neural network for Group 1, NLPCA can generate one time series to represent 17 varieties. Figure 7 shows the best one selected from 20 times of training. After 25346 interactions, the MSE_train is 0.5441 and MSE_test is 0.4675. The MSE_test is less than MSE_train. The correlation coefficient of regression between the output of NLPCA and the sum of all of 17 series is 0.8486. From the figure, we can observe that in some special point as in 01/1997, the result of 88r from NLPCA can not reach the real point. Table 3 shows the distribution of the data of 88r (Juristic determination for change the manner of contract) from October of 1996 to March of 1997. The value of January of 1997 is 270 times of December of 1996 and 15.8 time of February of 1997. This large variation may help to explain the lower correlation coefficient, but the reason of this sharp change is still in study.

![Fig. 7. NLPCA result from Group 1 (upper: 17 time series; middle: output of middle layer compared with sum of 17 time series; bottom: NN output versus sum of withdrawal)](image)
Table 3. Distribution of withdrawal 88r from 10/1996 to 03/1997

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Withdrawal 88r, R$</td>
<td>310903</td>
<td>286914</td>
<td>140539</td>
<td>37987606</td>
<td>2400623</td>
<td>196894</td>
</tr>
</tbody>
</table>

4.3.2 NLPCA for Group 2

After the pre-processing of the data in Group 1, a neural network with a structure of 8:3:1:3:8 is used to reduce dimension of the 7 time series in Group 2 together with the feature vector of Group 1. Figure 8 shows the best result from 20 times of training. After 18902 interactions, the MSE_train is 12.0508 and MSE_test is 18.4052 (the reason for big value is that we do not use the largest mean and stand division to normalize all varieties). The MSE_test is larger than MSE_train. The correlation coefficient between the output of NLPCA and the sum of all of 8 varieties is 0.9765.

Fig. 8. NLPCA result from group 2 using the result for NLPCA group 1 (upper: 8 time series; middle: output of middle layer compared with sum of 8 time series; bottom: NN output versus sum of withdrawal)
5 Conclusions

The main objectives for the development of the Actuarial System of FGTS are non linearity testing, feature extraction, correlation and causality analysis among the varieties of FGTS, prediction and etc. As an initial research of the project, this paper just focused on the part of feature extraction and dimension reducing of data set. After some experiments with 27 time series of withdrawal, we have the following conclusions.

![Fig. 9. Final four time series for withdrawal of FGTS](image)

1. Using NLPCA, 24 time series are combined to one series which is expected to represent 16.06% values of average yearly of all withdrawal. Together with time series 01, 05 and 91, these four time series are shown in Figure 9.

2. The elimination methods, such as PCA method, are not suitable for dimension reduction in the case of analyzing withdrawal of FGTS. With PCA, we can not reduce the variates of Group 1 and Group 2 to one that can well represent the main feature of the data set.

3. NLPCA method is suitable for dimension reduction of withdrawal time series. With NLPCA, we can get one time series to represent 17 ones in Group 1 with the correlation coefficient 0.8486, and get one to represent 8 in Group 2 with the correlation coefficient 0.9765. These results may not be the optimum, but at least the best among 20 training.
4. The results from Group 1 and Group 2 show that there are high order correlations between the output from NLPCA and sum of time series from this group. The reason of this fact should be deeply studied in further research.

5. The outside influence to FGTS is significant even we just analyzed the case of withdrawal by SIDA/AIDS patients. The politics of government for free treatment of these kinds of persons get well effect to society, specially to reduce the cost of FGTS. We will continue to study the social, economic and politics effects to FGTS.

Further research will continue to analyze the property of withdrawal, deposit, financial application and 3% yearly interest for every account of FGTS and the relationship among them. We will formally test the non-linearity of data set and to analyze correlation and causality using data mining method. In order to efficiently predict the future behavior of FGTS, dynamic model will be constructed based on the result of this paper.

References