how are we doing in education?: a SPCA approach

*L*uis A. Navarro H.
Universidad Nacional de Ingeniería -Perú, *luisnavarro@uni.edu.pe*

Abstract

This paper proposes the use of a new tool called Sparse Principal Component Analysis (SPCA) to support decision-making in educational management. Educational quality is influenced by some external factors that affect this training dynamic, such as drug use, child labor, crimes in their different forms, among others; where it is necessary to have synthetic indicators that are easy to interpret and built from many education indicators for contexts where few geographic areas are available. As an illustration, this work implements the SPCA in the creation of a Synthetic Indicator that measures the different intensities of crime occurrence in the constitutional province of Callao - Peru. The results show the adequacy of the proposal in considering the automatic choice of crimes with the highest occurrences and the management in the treatment of few geographic areas for many forms of crime. The different values of the Synthetic Indicator for each geographical area allow the different intensities to be compared and that educational policies must take into account the impact on the educational community

keywords SPCA, Multivariate Analysis, Statistical Inference

1 Introduction

It is common practice for the statistical units of the Ministries of Education to gather and analyze relevant statistical information to provide a comprehensive overview of the current situation of education in the Callao region, with educational indicators in their different forms for follow-up and monitoring purposes to ensure educational quality, such as infrastructure indicators, financing indicators, information technology indicators, indicators of crimes that affect the school environment, student/teacher rate indicators, access indicators, among others. In particular, as an illustration, the SPCA is used in the creation of a synthetic indicator from multiple crime indicators in a context where there are few geographic areas. The crime is an unlawful and imputable action that is punished by means of a criminal sanction. There are different types of crimes, those that are generally accepted in any society and others that are accepted only in certain societies; in both cases, some of these crimes present small changes in their definition as society evolves in their perception. To combat criminal activities it is necessary to understand the dynamics of intensities of occurrences of crime. In other words, to develop theoretical-practical approaches that describe these intensities of crime occurrences and that allow us to identify those geographic areas of populations with more or less criminal activities, contrasting certain hypotheses of interest McGuffog Et All [4].

In section 2, the problem to be solved is described, the construction of a synthetic crime indicator under an SPCA format; in Section 3, details the proposed solution by presenting the SPCA model; in section 4, the results found and adequate for the problem to be solved are shown; Finally, in the last section, some conclusions are presented on the successful implementation of the proposal of this work.

2 Description of the problem

According to the "Regional Report on Human Development, Security with a Human Face Diagnosis and Proposals for Latin America" (UNDP, 2013-2014), Peru is the country with the highest perception of insecurity and victimization, with respect to the others countries of the region. Certain types of crime have increased alarmingly in different geographical areas of Lima and Callao

in particular. The Public Ministry is in charge of managing the investigation and then prosecuting it by law, using the information provided by the National Police during the preliminary stage of the investigation. On the other hand, since the National Police are the ones who carry out the identification and intervention of the various actors who break the law, it is vitally important to design and implement some tools that allow them to continuously and easily interpret the effects of the fight against crime.

One of the reasons to sustain an efficient policy in the fight against crime is to understand the dynamics of crime in its different forms. From the above, making a crime intensity measurement tool available is of great importance.

One way to measure crime intensities is through the preparation of Synthetic Indicators obtained through the application of conventional Principal Component Analysis (PCA), used in contexts where the number of records is greater than the number of crimes, Atanu Et All [1], Shehu Et All [6]. However, the previous proposal is not pertinent with the generation of Synthetic Indicators for the geographical area of the Constitutional Province of Callao (Lima-Peru) for the period 2013-2018, which is made up of few districts and a large number of different types of crimes. Figure 1 shows the frequencies of occurrences of 49 different types of crime in 6 districts [1]

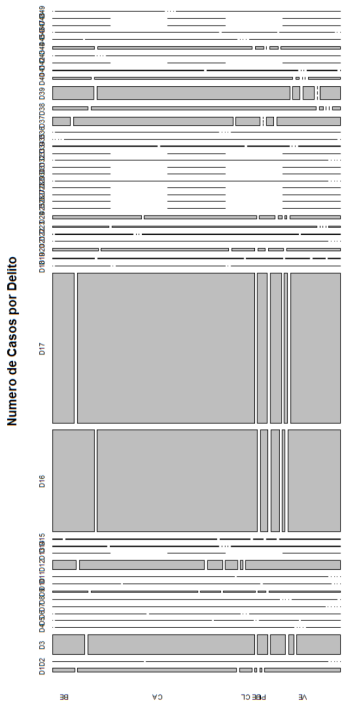

Figure 1: Crime Frequency - Mosaic Plot

En lo que sigue se trabaja con tasas de ocurrencias de delitos definidos como

$$x_{ij} = \frac{y_{ij}}{p_i} \times 100000$$

With y_{ij} being the frequency of occurrence of the crime j in the geographical area i, and p_i being the population of the geographical area i.

In any analysis of large amounts of information, it is common practice to identify some behavior patterns not seen with the naked eye. In this context, as mentioned above, a widely used technique is Principal Component Analysis (PCA) whose purpose is to reduce dimension and thus see the possibility of identifying some behavior patterns not previously seen in the original dimensions. The conventional PCA builds new synthetic features or main component as a linear combination of the original features, but does not allow working in situations where the number of records is greater than the number of features, nor does it allow the possibility of incorporating only the original features. of greater importance in the linear combinations of the new

[1]Distritos: Bellavista (BE), Callao (CA), Carmen de la Legua Reynoso (CL), La Perla (PE), La Punta (PU) y Ventanilla (VE)
Delitos: Homicidio (D1), Aborto (D2), Lesiones (D3), Exposición Al Peligro (D4), Atentados/Patria Potestad (D5), Omisión Asistencia Familiar (D6), Matrimonio Ilegal (D7), Delito c_Estado Civil (D8), Violación Libertad Personal (D9), Violación de la Intimidad (D10), Violación de Domicilio (D11), Violación Libertad Sexual (D12), Proxenetismo (D13), Ofensa Pudor Público (D14), Violación Secreto Comunicaciones_Libertad Reunión_Expresión_Trabajo_Secreto Profesional (D15), Hurto (D16), Robo (D17), Abigeato (D18), Apropiación Ilícita (D19), estafas y Defraudaciones (D20), Fraude en la administración (D21), Delitos informáticos (D22), Daños simples y agravados (D23), Receptación, usurpación y extorsión (D24), Acaparamiento, especulación y adulteramiento (D25), Negociación de bienes destinados a donaciones (D26), Funcionamiento ilegal de casinos de juegos (D27), Lucro indebido en importaciones (D28), Otros_Orden económico (D29), Delito financiero (D30), Delito monetario (D31), Contrabando (D32), Elaboración clandestina de productos (D33), Falsificación documentos en general (D34), Falsificación de sellos, timbres y marcas (D35), Otros_fé pública (D36), Peligro Común D(D37), T.L.D (D38), Micro comercialización de drogas (D39), Tenencia ilegal de armas (D40), Otros_seguridad pública (D41), Apología terrorismo (D42), Otros_contra tranquilidad (D43), Cometidos por particulares_estado (D44), Cometidos funcionarios públicos_estado (D45), Administración justicia_estado (D46), Pandillaje pernicioso (D47), Posesión de armas de guerra (D48), Otros delitos (D49)

main components−− in certain contexts different crimes occur with different intensities and it is pertinent to have a methodology that makes the selection of crimes with higher intensity of occurrences automatically.

The Sparse Principal Component Analysis (SPCA) rectifies the above and the proposals that exist depend on how the problem is formulated and the algorithm used in the proposed solution, see for example the work of Jolliffe *Et All* [3], Zou *Et All* [7] and d'Aspremont*Et All* [2]

3 Proposed solution

3.1 the SPCA model

There are some works on the design of some synthetic indicators based on some main previously defined crimes, see for example Atanu *Et All* [1] and Shehu *Et All* [6], who analyze 8 main crime rates using conventional Principal Component Analysis (PCA). However, the previous selection of which crimes would be for the analysis and subsequent generation of the synthetic indicators is not entirely clear (there is no standard protocol for how to select some of these crimes). This leads us to search for a variant of the conventional ACP method that also works in situations where the number of records (districts\geographic unit) is less than the number of variables (types of crimes\rates). It is the proposal of Zou *Et All* [7] that is used in the present work for the preparation of a Synthetic Crime Indicator

Suppose we have a sample of n vectors of dimension p and denoted as X_1, X_2, $\cdots$, X_n with vector of means $\overline{X}$, covariance matrix S and correlation matrix R. In what follows $x_j = [X_{j1}, X_{j2}, \cdots, X_{jp}]$ represents any of the n random vectors cited above and centered for convenience after a location transformation. In addition, the $n \times p$ dimension data matrix will be represented as X, where n is the number of observations and p is the number of variables.

3.1.1 Case $n > p$

It is common to use Principal Component Analysis (PCA) to reduce dimension and explore some pattern of behavior that at first glance cannot be identified in matrix X. Then, PCA allows us to represent a point x_j in R^p as $[z_1^j, z_2^j, \cdots, z_q^j]$ in R^q $(q < p)$, by a linear transformation $z_i^j = x_j a_i^{'}$ with $a_i = [a_{i1}, a_{i2}, \cdots, a_{ip}]$, and with the following properties:

- The z_i capture the maximum variance
- The z_i are orthogonal

That is, the interest is to explore the information contained in the data matrix X in a dimension space of R^p for a dimension R^q without losing information on the neighborhood between x_j seen from R^p, or as $[z_1^j, z_2^j, \cdots, z_q^j]$ seen from R^q.
The problem formulated in mathematical terms corresponds to finding vectors a_k (k≤ p) such that it maximizes

$$Var(z_i^j) \quad = \quad Var(x_j a_i^{'}) = a_i Cov(x_j, x_j) a_i^{'} = a_i S a_i^{'}$$

Subject to

$$a_i a_i^{'} = 1, \quad i \geq 2. \qquad a_i a_h^{'} = 0, \quad h < i$$

The solution to the previous problem is given by the a_i eigenvectors of the matrix S, see for example Johnson [5].
From the previous result we have that $x_j = z_i^j a_i$ for $j = 1, \cdots, n$ and $i = 1, \cdots, p$, which in matrix terms is $X = ZV^{'}$ with Z being a matrix of orthogonal columns of dimension $n \times p$ of the principal components $[z_i^j]$ and V an orthogonal matrix of dimension $p \times p$ of weights (eigenvectors of the matrix X). The above can also be obtained from the singular value decomposition of the matrix X, as $X = UDV^{'}$ with UD being the matrix of principal components Z, U of dimension $n \times p$ as a matrix of orthogonal columns and D of dimension $p \times p$ being a diagonal matrix. In this matrix format, it is true that $VV^{'} = I$ and that $VSV^{'}$ is a diagonal matrix
Finally, after implementing a traditional PCA, rotation techniques such as VARIMAX are generally used to interpret each of the main components obtained. The weights are generally different from zero, which often causes a difficult interpretation, especially when p takes a very large value.
In this context, the possibility of incorporating the option of reducing to zero some solution weights that are linked to some variables that do not reflect greater importance according to what the matrix X shows. In this direction, the SCoTLASS of Jolliffe *Et All* [3] successively proposes to maximize the variance

$$a_i S a_i^{'}$$

Subject to

$$a_i a_i^{'} = 1, \quad i \geq 2. \qquad a_i a_h^{'} = 0, \quad h < i$$

and

$$\Sigma_{j=1}^{p}|a_{ij}| \leq t$$

For some parameter t. Although for some small values ??of t the proposal generates weights equal to zero for some of the p variables, there is no guideline that indicates in general what values ??of t to use. Also here, the weights obtained are not sufficiently sparse when a high percentage of explained variance is required.

3.1.2 Case $n \leq p$

For any n, and in particular when $n \leq p$, the solution proposal of the Principal Components is obtained from an analysis of a Linear Regression model. To understand the above, the link between an PCA and a Linear Regression model will be shown where the weights of the main components are obtained as a solution to a problem of obtaining parameters (PC weights) of a Linear Regression model with certain additional properties and in which it is not necessarily true that $VV^{'} = I$ and that $VSV^{'}$ is a diagonal matrix. Finally, it must be made clear that the proposal to present is known as Ridge Regression and manipulates all kinds of cases, even when $n > p$.
A Principal Component (PC) is a linear combination of p variables, and the proposal is that the weights of these PCs can be recovered as parameter estimates in a Linear Regression model, where the PCs recovered from a Singular Value Decomposition, is the response variable and the p variables are the explanatory variables. According to Zou *Et All*) [7] the following results direct the obtaining of the PCs of a matrix X with the condition that some of the weights of these PCs are zero.

Result 1 The first result links the finding of the weights of each PC from the estimated parameters (weights) obtained in solving an estimation problem of a Linear Regression model with a Ridge penalty,

Theorem For each i, denote by $Z_i = U_i D_{ii}$ the ith principal component. Consider a positive λ, and the Ridge estimates $\widehat{ß}_{Ridge}$ given by

$$\widehat{ß}_{Ridge} = arg\ min_{\beta}\|Z_i - X\beta\|^2 + \lambda\|\beta\|^2$$

then $\widehat{V} = \frac{\widehat{ß}_{Ridge}}{\|\widehat{ß}_{Ridge}\|} = V_i$

The demonstration and details are in Zou *Et All* [7]. The previous result handles all types of matrix X, and the insertion of the Ridge penalty constraint $\lambda\|\beta\|^2$ guarantees the recovery of each PC from the weights V_i. It is important to note that this result does not consider the possibility of obtaining V_i sparse

Result 2 A second result, also seen in Zou *Et All* [7], includes an additional penalty that controls the possibility that some weights are equal to zero (sparse approximation) of the i-th PC. The latter is obtained if an additional penalty of type L_1 is added to the objective function stated in the previous theorem,

$$\widehat{\beta} = arg\ min_{\beta}\|Z_i - X\beta\|^2 + \lambda\|\beta\|^2 + \lambda_1\|\beta\|_1$$

Where $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$, with $\widehat{V}_i = \frac{\widehat{\beta}}{\|\widehat{\beta}\|}$ being an approximation of V_i, and $X\widehat{V}_i$ being the approximate i-PC. Since normalized adjusted coefficients are being used, the scale factor λ does not affect $\widehat{V}_i$. For a fixed λ the minimization problem can be solved for all λ_1 using for example the LARS-EN algorithm Zou *Et All* [7]

Result 3 A third result, also seen in Zou *Et All* [7], proposes the finding of the CPs incorporating the above criteria directly, and not in 2 stages as previously seen. For this, let us remember that in a Singular Value Decomposition $z_i^j = x_j a_i^{'}$ with a_i autovector, from where $x_j = z_i^j a_i = x_j a_i^{'} a_i$. That is, the idea is to propose the smallest discrepancy between the true value x_j and a projection $\alpha\beta^{'}x_j$ with α and β row vectors of dimension p. That is to say,

Theorem In obtaining the first k CPs, If $A_{p*k} = [\alpha_1, \cdots, \alpha_k]$ and $B_{p*k} = [\beta_1, \cdots, \beta_k]$, and for any positive λ, let

$$(\widehat{A}, \widehat{B}) = arg\ ?min_{A,B} \sum_{j=1}^{n}\|x_j - AB^{'}x_j\|^2 + \lambda\sum_{i=1}^{k}\|\beta_i\|^2$$

subject to

$$AA^{'} = I_{k\times k}$$

then, $\widehat{B} \propto V_i$ for $i = 1, \cdots, k$

Finally, the lasso approximation is used to generate sparse weights, by inserting the lasso penalty into the objective function of the previous theorem,

$$(\widehat{A},\widehat{B}) = arg\ ?min_{A,B} \sum_{j=1}^{n} \|x_j - AB^{'}x_j\|^2 + \lambda \sum_{i=1}^{k} \|\beta_i\|^2 + \sum_{i=1}^{k} \lambda_{1i}\|\beta_i\|_1$$

subject to

$$AA^{'} = I_{k\times k}$$

Different λ_{1i} are used to penalize the weights of the different principal components - the above optimization problem is known as the Sparse Principal Component Analysis (SPCA) whose obtaining of solution B, given $\widehat{A}$, follows the ideas seen above in solving a Ridge regression problem.

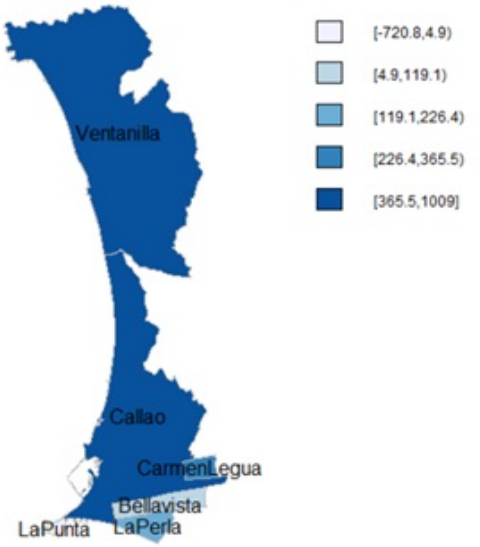

Figure 2: Crime Synthetic Indicator - 1ht PC

3.2 Calculation of Correlation Coefficients between Principal Components and original Variables

Once the PCs are obtained, it is necessary to label them to have an easy interpretation of the results found. In that direction, we define

$$\begin{aligned} z_i &= \widehat{\beta}_i^{'}x \\ a_j^{'} &= (0,\cdots,1,\cdots,0) \\ x_j &= a_j^{'}x \end{aligned}$$

from where

$$\begin{aligned} Cov(x_j,z_i) &= Cov(a_j^{'}x,\widehat{\beta}_i^{'}x) = a_j^{'}\Sigma\widehat{\beta}_i \approx a_j^{'}X^{'}X\widehat{\beta}_i \\ Var(x_j) &= Var(a_j^{'}x) \approx a_j^{'}X^{'}Xa_j \\ Var(z_i) &= Var(\widehat{\beta}_i^{'}x) \approx \widehat{\beta}_i^{'}X^{'}X\widehat{\beta}_i \end{aligned}$$

What you finally get

$$r_(z_i,x_j) = \frac{a_j^{'}X^{'}X\widehat{\beta}_i}{\sqrt[2]{a_j^{'}X^{'}Xa_i}\sqrt[2]{\widehat{\beta}_i^{'}X^{'}X\widehat{\beta}_i}} \quad i=1,\cdots,k \quad j=1,\cdots,p$$

4 Results

The application of the SPCA to the X data matrix does not fix the number of original characteristics to be used in the preparation of the new principal components, but rather allows the information contained in the same data matrix X to guide the selection of the original variables to be considered in the new principal components.

Based on the results found from the application of the SPCA algorithm, it was decided to consider only the first PC with a Total Variance of approximately 74%

```
  Call:
spca(x = my_data, K = 4,
  para = c(0.06, 0.16, 0.1, 0.5), type = "predictor",
  sparse = "penalty", lambda = 1e-06, trace = TRUE)
    4 sparse PCs
Pct. of exp. var. : 74.2 22.2  2.0  0.1
Num. of non-zero loadings :  5 5 5 5
Sparse loadings
       PC1
D3  -0.243
D16 -0.558
D17  0.776
D37 -0.160
D39  0.046
```

From the correlations between the first CP and crimes whose weights are different from zero, it is observed that the synthetic crime indicator behaves as an arithmetic ratio that measures "how many more robberies there are than injuries, thefts and common danger, or vice versa". That is, despite the fact that the citizens of the Province of Callao perceive that there is a strong preponderance between injuries against the body, life and health (injuries), crimes against property (theft and robbery) and crimes against public

safety (common danger), the synthetic indicator of crime pertinently shows the different intensities in terms of differences between the crime of robbery and the crime of injury, theft and common danger.

	BetaEst	CoefCorr Z vs Xj
Lesiones	-0.24278441	-0.86007337
Hurto	-0.55808532	-0.80219413
Robo	0.77578054	0.91728579
Peligro Comun	-0.16007128	-0.77429856
MicroComerc Drogas	0.04624118	0.06684911

A graphic visualization of this difference in intensity of the synthetic crime indicator is shown in Figure 2.For example, the Ventanilla (VE) district presents more robberies than injuries, thefts and common dangers; On the other hand, the district of La Punta (PU) has a greater intensity of cases of injuries, thefts and common danger, regarding the crime of robbery.

5 Conclusions

The proposed SPCA solution is pertinent to the creation of a Synthetic Crime Indicator that measures different intensities between the districts that make up the Constitutional Province of Callao. Furthermore, it is appropriate to display the results through an Intensity Map providing an easy to perceive and interpret visualization, especially when making decisions on educational policies.

The data matrix X corresponds to only a part of the total of crimes that have occurred, due to the perception that the victims have regarding that they should do justice. However, the proposed solution using the SPCA is appropriate to any context of recorded information.

References

[1] E. Y. Atanu. Analysis of nigeria's crime data: A principal component approach using correlation matrix. *International Journal of Scientific and Research Publications*, 2019.

[2] F d'Aspremont; Bach and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *Journal of Machine Learning*, pages 1269–1294, 2008.

[3] Trendafilov N. T. Jolliffe, I. T. and M. Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistic*, pages 531–547, 2003.

[4] Western J.S. McGuffog, I. and P. Mullins. Exploratory spatial data analysis techniques for examining urban crime. *British Journal of criminology*, pages 309–329, 2001.

[5] Johnson R. and Wichern D. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall/Pearson Education, Inc, 2007.

[6] H.G. Dikko Shehu U. Gulumbe and Yusuf Bello. Analysis of crime data using principal component analysis: A case study of katsina state. *CBN Journal of Applied Statistics*, 2011.

[7] Hastie T. Zou H. and Tibshirani R. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, pages 265–286, 2006.