Tesis de Maestría en Estadística Aplicada

María Susana Vitelleschi[*]

Directora:

Marta Beatriz Quaglino

MODELOS PCA A PARTIR DE CONJUNTOS DE DATOS CON INFORMACIÓN FALTANTE. ¿SE AFECTAN SUS PROPIEDADES?[**]

Premio Provincial 2009 a Tesis de Maestría[***]

Resumen. En este trabajo se aborda la problemática de la construcción de modelos PCA *(Principal Component Analysis)* a partir de conjuntos de datos con información faltante. Se trabaja sobre tres situaciones diferentes con relación a la matriz de datos originales. En cada situación se generaron pérdidas a través de mecanismos aleatorios y no aleatorios, en diferentes porcentajes en una sola variable por vez, seleccionada mediante dos criterios: la que más contribuye y menos contribuye en la formación de la primera componente principal. A partir de cada conjunto de datos incompletos se construye el modelo PCA utilizando: Casos Completos, *Nonlinear Iterative Partial Least Squares* (NIPALS) y *Expectation Maximization* (EM). Se comparan los resultados con los obtenidos a través del conjunto de datos originales. Se definen una serie de medidas para estudiar cómo se afectan los resultados según la dimensión de la matriz de datos, el porcentaje y el mecanismo de pérdida, con relación a: bondad del ajuste, bondad de predicción, vectores cargas, ortonormalidad de la matriz de cargas y ortogonalidad de la matriz de *"scores".*

Palabras Clave: Mecanismos de Pérdidas; Algoritmo NIPALS; Algoritmo EM.

---

[*] Docente-Investigadora de la Facultad de Ciencias Económicas y Estadística. **Contacto**: mvitelle@fcecon.unr.edu.ar.

[**] Defendida en Rosario el 11 de diciembre de 2008.

[***] Otorgado por la Secretaría de Ciencia, Tecnología e Innovación del Gobierno de la Provincia de Santa Fe.

Abstract. This paper deals with the issue of building PCA (Principal Component Analysis) models from data sets with missing information. This Thesis worked on three different situations related to the original data set. In each situation, losses were generated through random and not random mechanisms, in different percentages in one variable at a time, selected by two criteria: the one that contributes the most and the one that contributes the least to the formation of the first principal component. With each set of incomplete data is built the PCA model using: Complete Cases, NIPALS algorithm and EM algorithm. The results are compared to those obtained from the original data set. It is examined how they are affected depending on the size of the data matrix data, the percentage of missing information and the missing data mechanism, in relation to: the goodness of fit, the goodness of prediction, loading vectors, the orthonormality of the loading matrix and the orthogonality of the score matrix. Measures are defined to study how these aspects are affected.

Key Words: Missing Data Mechanisms; NIPALS algorithm; EM algorithm.

1. Introducción

Dada la complejidad de los fenómenos que se investigan en, prácticamente, todas las ciencias experimentales es usual que se midan varias variables sobre muchas unidades de observación, dando origen a grandes volúmenes de datos. Los métodos estadísticos multivariados son apropiados en estas situaciones ya que analizan simultáneamente toda la información.

El Análisis de Componentes Principales es uno de los métodos estadísticos más utilizados para el análisis de datos multivariados, situación en la que es frecuente la aparición de información faltante. Esta problemática ha recibido mucha atención e importancia en los últimos años, dado que los esfuerzos del analista deben tender a no descartar las unidades con información incompleta.

El objetivo que se planteó en esta Tesis fue evaluar diferentes aspectos del modelo PCA que podrían ser afectados, según el método utilizado para la construcción del mismo, a partir de una variedad de situaciones que combinan distintas dimensiones de la matriz de datos originales, pérdidas en variables que influyen de forma diferente en las componentes principales, distintos mecanismos y porcentajes de pérdidas. Se propusieron varias medidas de comparación, con el fin de realizar un aporte a través de recomendaciones para el uso de uno u otro método.

2. Análisis de Componentes Principales

Análisis de Componentes Principales (Morrison, 2005; Sharma, 1996) es una técnica que permite explicar la variabilidad existente en un conjunto de datos multivariados con un gran número de variables altamente correlacionadas, a través de un número inferior de nuevas variables no correlacionadas llamadas componentes principales (CP) o variables latentes, construidas como combinaciones lineales de las originales de modo tal que su variancia decrece de la primera a la última.

A través de los trabajos de Wold et al. (1987), PCA se utilizó para construir un modelo explicativo de las variables originales en el que las CP intervienen como variables explicativas y puede ser utilizado para predecir valores futuros.

3. Construcción de modelos PCA a partir de conjuntos de datos con información faltante

Los métodos que se utilizaron para construir los modelos PCA (Wold et al., 1999) fueron: CC (Casos Completos en el que se eliminan las observaciones con información faltante), NIPALS *(Nonlinear Iterative Partial Least Squares*) es un método secuencial mediante el cual, en cada ciclo, se calcula una componente principal directamente de la matriz de datos (Geladi y Kowalski, 1986) y EM *(Expectation Maximization)* adopta una forma diferente a la conocida y en forma iterativa completa la matriz de datos (Little y Rubin, 2002). El conocimiento del mecanismo que produce que ciertos valores estén perdidos es un elemento fundamental para la elección del método más apropiado. Se trabajó con los siguientes mecanismos: MCAR *(Missing Completely At Random)*, MAR (*Missing At Random)* y NMAR *(Not Missing At Random).*

4. Estudio comparativo de estrategias para el tratamiento de información faltante en PCA

Se construyeron modelos PCA a partir de matrices de datos de diferentes dimensiones: rectangular con mayor número de individuos que de variables, cuadrada con igual número de individuos que de variables y rectangular con menor número de individuos que de variables. Estas matrices de datos provienen de distintas situaciones reales y en su expresión original tienen información completa. Sobre cada una de esas matrices de datos se generaron diferentes porcentajes (10%, 20%, 30%, 40% y 50%) de falta de información a través de mecanismos de pérdidas aleatorios (MCAR y MAR) y no aleatorios (NMAR). Las pérdidas se generaron en una sola variable, seleccionada a través de dos criterios diferentes, aplicados por separado. Uno de ellos fue elegir la variable que más contribuía en la formación de la primera CP y el otro consistió en seleccionar la que menos aportaba en su formación. En cada uno de los conjuntos de datos incompletos se construyó el modelo PCA (Walczak y Massart, 2001) utilizando: CC, EM y NIPALS. En todas las situaciones se comparan sus resultados con los obtenidos a partir de la matriz de datos originales.

Se definieron ocho medidas específicas con el propósito de evaluar distintos aspectos en los que pueden diferir los resultados: bondad del ajuste, bondad de predicción, ortonormalidad de la matriz de cargas y ortogonalidad de la matriz de scores.

Para visualizar la existencia de ciertas pautas de regularidad en el uso de uno u otro método de construcción del modelo PCA con información faltante, las comparaciones realizadas se resumieron de forma tal de evidenciar la existencia de: diferencias entre métodos, influencias del porcentaje y del mecanismo de pérdidas.

5. Conclusiones

Los resultados obtenidos a partir de los procesos de comparación abordados, permitieron reafirmar la importancia que tiene, frente a información faltante, el aplicar métodos de análisis estadísticos adecuados, que incorporen en la estimación de los modelos, toda la información disponible. A partir de las situaciones estudiadas se observó que EM, un método de concepción netamente estadístico, muestra sus ventajas en los casos clásicos de información multivariada con pocas variables medidas sobre muchos individuos. Mientras que un método como NIPALS, surgido como una aplicación particular de los modelos de regresión, es más adecuado en los casos extremos de información multivariada con pocos individuos y muchas variables. En ningún caso se muestra la conveniencia de desechar individuos por no contar con información completa para ellos.

Referencias Bibliográficas

Geladi, P. y Kowalski, B. R. (1986). Partial least squares regression: a tutorial. *Analytica Chimica Acta*, 185, 1-17.

Little, R. y Rubin, D. (2002). *Statistical analysis with missing data* (2nd Editon). New York: John Wiley and Sons.

Morrison, D. (2005). *Multivariate statistical methods* (4th Edition). New York: McGraw – Hill.

Sharma, S. (1996). *Applied multivariate techniques.* New York: John Wiley and Sons.

Walczak, B. y Massart, D. (2001). Dealing with missing data: Part II. *Chemometrics and Intelligent Laboratory Systems*, 58, 29-42.

Wold, S.; Esbensen, K y Geladi, P. (1987). Principal component analysis. *Chemometrics and Intellegent Laboratory Systems*, 2, 37-52.

Wold, S.; Eriksson, L.; Johansson, E. y Kettaneh-Wold, N. (1999). *Introduction to multi and megavariate data analysis using projection methods (PCA and PLS)*. *Umeå,* Sweden: Umetrics.