Comparación de métodos para el tratamiento de información faltante en un análisis de componentes principales sobre datos biológicos.

RECIBIDO: 03/06/2008
ACEPTADO: 19/03/2009

Quaglino, M. B. • Vitelleschi, M. S.

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Facultad de Ciencias Económicas y Estadística. Universidad Nacional de Rosario. Bvrd. Oroño 1261. (2000) Rosario. E-mail: mvitelle@fcecon.unr.edu.ar

RESUMEN: El Análisis de Componentes Principales (ACP) es una técnica para analizar datos multivariados, utilizada frecuentemente con el propósito de sintetizar y explorar realidades complejas. Clásicamente, su aplicación requiere información completa para todas las observaciones. Si se presentan matrices de datos incompletas, es usual descartar los individuos con información faltante, acarreando probablemente, pérdida de eficiencia.
El algoritmo Nonlinear Iterative Partial Least Squares (NIPALS), propuesto inicialmente en aplicaciones del área química, obtiene las componentes principales directamente a partir de las matrices de datos, aún existiendo pérdidas.
En este trabajo se analizan datos provenientes de niños con Leucemia Linfoblástica Aguda, sobre quienes se registraron mediciones fisiológicas y bioquímicas. Se comparan críticamente los resultados del ACP sobre la matriz original sin información faltante y sobre matrices obtenidas simulando pérdidas al azar, utilizando algoritmo NIPALS y el método de Casos Completos.
PALABRAS CLAVE: Análisis de Componentes Principales, algoritmo NIPALS, información faltante.

SUMMARY: *Comparison of methods for dealing with missing information in principal component analysis on biological data.*
The Principal Component Analysis (PCA) is a technique for analyzing multivariate data. It is often used to explore and synthesize complex realities.
Its implementation requires complete information for all observations. In the presence of incomplete data, the usual alternative is to exclude the individuals with missing information, deriving probably in a loss of efficiency. The Non-linear Iterative Partial Least Squares (NIPALS) algorithm, which was initially proposed for applications in the chemical area, obtains the principal components directly from data matrixes, even with missing information. This work analyzes information

on children with Acute Lymphoblastic Leukemia, where physiological and biochemical measurements were carried out. PCA results obtained by NIPALS algorithm and the Complete Case method are critically compared utilizing the original data matrix and matrixes constructed with missing information at random.
KEYWORDS: Principal Component Analysis, NIPALS algorithm, Missing information.