

7° Simposio Internacional de Investigación Multidisciplinaria / Ciencia y Tecnología
7th Internationalm Symposium on Multidisciplinary Research / Sciences and Technology

CT-12 Aplicación de rutinas en lenguaje R para análisis de componentes principales

Application of R-language routines for principal components analysis

Ezequiel A. López-Bautista^{1*}, Byron H. González-Ramírez²

¹Sub-área de Métodos de Cuantificación e Investigación y ²Centro de Telemática,
Facultad de Agronomía, Universidad de San Carlos de Guatemala,

*Autor al que se dirige la correspondencia: elbautis@gmail.com

Resumen

Cuando se dispone de gran cantidad de observaciones para cada individuo, los análisis univariados tienen la limitación de no contemplar las interrelaciones entre todas las variables, ni cómo esas interrelaciones afectan al conjunto de individuos. Como alternativa fueron desarrolladas las técnicas multivariadas, que brindan descripciones de estas interrelaciones, tomando las variables en su conjunto. El análisis de componentes principales (ACP) es la esencia del análisis estadístico multivariado y tiene como objetivo reducir el número de variables, a través de la generación de nuevas variables, que son combinaciones lineales de las variables originales. El objetivo de este trabajo es presentar una serie de rutinas computacionales en el lenguaje R v. 3.10 para el ACP. Este lenguaje es de uso mundial, por ser libre, flexible y con soporte de estadísticos de reconocido prestigio. Para ilustrar su aplicación, se tomó una base de datos referente a los precios de cinco tipos de alimentos (en centavos de dólar): pan, hamburguesa, leche, naranjas y tomates en 23 ciudades de los Estados Unidos. Los datos fueron analizados exploratoriamente seguido del análisis de correlación entre pares de variables. Utilizando la función PRINCOMP y el paquete RCOMMANDER se realizó el ACP y la construcción del biplot. Además fueron utilizadas las bibliotecas: Hmisc, Plotrix, Ellipse, Corrplot, Ggplot2, Devtools y Vqv/ggbiplot. Los resultados fueron comparados con los obtenidos en otros programas, no encontrando diferencias. Se concluye que el lenguaje R es una herramienta poderosa, debiendo divulgar su uso en la docencia de la estadística a nivel universitario.

Palabras claves: Análisis de componentes principales, estadística universitaria, análisis multivariado

Abstract

When a great deal of observations are available for each individual, univariate analyses have limitation to consider the interactions between all variables, and how these interrelations affect the group of individuals. As an alternative, multivariate techniques were developed, which provide descriptions of these interactions, taking the variables as a whole. Principal component analysis (PCA) is considered the essence of multivariate statistical analysis, and its objective is to reduce the number of variables, through the generation of new variables, which are linear combinations of the original variables. The purpose of this research is to present a series of computational routines in the R language v. 3.10 for the PCA. This language used worldwide, because it is free, flexible and supported by statisticians of recognized prestige. To illustrate its application, a database was taken from the prices of five types of food (in dollar cents): bread, hamburger, milk, oranges and tomatoes in 23 cities in the United States. The data were analyzed by exploratory analysis and then by correlation analysis between pairs of variables. The PCA and the biplot construction were performed by using the PRINCOMP function and the RCOMMANDER package. In addition the libraries: Hmisc, Plotrix, Ellipse, Corrplot, Ggplot2, Devtools and Vqv / ggbiplot were consulted. The results were compared with those obtained in other softwares finding no differences. Concluding that the R language is a powerful tool, and its use should be disseminated in the teaching of statistics at university education.

Keywords: Principal component analysis, university statistics, multivariate analysis