DM STAT-1 CONSULTING BRUCE RATNER, PhD

574 Flanders Drive North Woodmere, NY 11581 br@dmstat1.com
516.791.3544 ~ fax 516.791.5075 www.dmstat1.com

Principal Component Analysis of Yesterday and Today

Bruce Ratner, Ph.D.

ABSTRACT

Principal component analysis (PCA), invented in 1901 by Karl Pearson as a data reduction technique, uncovers the interrelationship among many variables by creating linear combinations of the many original variables into a few new variables such that most of the variation among the many original variables is accounted for or retained by the few new uncorrelated variables. The literature is sparse on PCA used as a reexpression, not a reduction, technique. I posit the latter distinction for repositioning PCA as an EDA (exploratory data analysis) technique. EDA is a force for identifying structure: PCA is a classical data reduction technique of the 1900s; PCA is a reexpression method of 1997, a very good year for the statistics community as Tukey's seminal EDA book was released then; and PCA is a data mining method of today, as if the voguish term data mining can replace EDA proper. In this article, I put PCA in its appropriate place in the EDA reexpression paradigm. I illustrate PCA as a statistical data mining technique capable of serving in a common application with expected solution and illustrate PCA in an uncommon application, yielding a reliable and robust solution.

In addition, I provide an original and valuable use of PCA as a method of variable selection. Finding the best possible subset of variables to put in a model has been a frustrating exercise. Many methods of variable selection exist, but none of them is perfect. Moreover, they do not create new variables, which would enhance the predictive power of the original variables themselves. I illustrate PCA as a statistical data mining technique that can compete with the popular and suboptimal stepwise variable selection and the mighty "seven-step cycle of statistical analysis" posited by Tukey.

For more information about this article, contact the author.