 ISBN: 9789073592421

IPS Paper

Flexible clustering for asymmetric data via mixtures of unrestricted skew normal factor analyzers

Author: Prof. Tsung-I Lin

Coauthors: Wan-Lun Wang

Submission ID: 1048

Reference Number: 1048

Brief Description

We propose a novel extension of the MFA model established by replacing the normality assumption for the component latent factors with the uMSN distribution, called the MuSNFA model, as a new model-based clustering tool for handling highdimensional data with asymmetrical characteristics and possibly missing values.

A computationally feasible ECM algorithm through the incorporation of two auxiliary indicator matrices enables certain degree of convenience in its implementation.

Experimental results demonstrate that the proposed MuSNFA model may outperform the traditional approaches on providing a better fit, improved classification performance, and more accurate prediction for missing values.

Abstract

Mixtures of factor analyzers (MFA) based on the restricted skew normal distribution (rMSN) has been shown to be a flexible tool to handle asymmetrical high-dimensional data with heterogeneity. However, the rMSN distribution is oft-criticized a lack of sufficient ability to accommodate potential skewness arisen from more than one feature space. This paper presents an alternative extension of MFA by assuming the unrestricted skew normal (uMSN) distribution for the component factors. In particular, the proposed mixtures of unrestricted skew normal factor analyzers (MuSNFA) can simultaneously capture multiple directions of skewness and deal with the occurrence of missing values or nonresponses. Under the missing at random (MAR) mechanism, we develop a computationally feasible expectation conditional maximization (ECM) algorithm for computing the maximum likelihood estimates of model parameters. Practical aspects related to model-based clustering, prediction of factor scores and missing values are also discussed. The utility of the proposed methodology is illustrated with the analysis of simulated data and the Pima Indian women diabetes data containing genuine missing values.