Generalized probabilistic principal component analysis

Weining Shen

UC Irvine
E-mail: weinings@uci.edu

Abstract: Principal component analysis (PCA) is a well-established tool in machine learning and data processing. The principal axes in PCA were shown to be equivalent to the maximum marginal likelihood estimator of the factor loading matrix in a latent factor model for the observed data, assuming that the latent factors are independently distributed as standard normal distributions. However, the independence assumption may be unrealistic for many scenarios such as modeling multiple time series, spatial processes, and functional data, where the outcomes are correlated. In this paper, we introduce the generalized probabilistic principal component analysis (GPPCA) to study the latent factor model for multiple correlated outcomes, where each factor is modeled by a Gaussian process. Our method generalizes the previous probabilistic formulation of PCA (PPCA) by providing the closed-form maximum marginal likelihood estimator of the factor loadings and other parameters. Based on the explicit expression of the precision matrix in the marginal likelihood that we derived, the number of the computational operations is linear to the number of output variables. Furthermore, we also provide the closed-form expression of the marginal likelihood when other covariates are included in the mean structure. We highlight the advantage of GPPCA in terms of the practical relevance, estimation accuracy and computational convenience. Numerical studies of simulated and real data confirm the excellent finite-sample performance of the proposed approach.