samsi
NSF·Duke·NCSU·UNC

Sixth Bayesian, Fiducial, and Frequentist (BFF6) Conference on Model Uncertainty
April 28, 2019 – May 1, 2019

SPEAKER TITLES/ABSTRACT

Mengyang Gu
Johns Hopkins University

"Generalized Probabilistic Principal Component Analysis of Correlated Data"

Abstract: Principal component analysis (PCA) is a well-established tool in machine learning and data processing. \cite{tipping1999probabilistic} proposed a probabilistic formulation of PCA (PPCA) by showing that the principal axes in PCA are 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 output variables are correlated.

In this paper, we introduce the generalized probabilistic principal component analysis (GPPCA) to study the latent factor model of multiple correlated outcomes, where each factor is modeled by a Gaussian process. The proposed method provides a probabilistic solution of the latent factor model with the scalable computation. In particular, we derive the maximum marginal likelihood estimator of the factor loading matrix and the predictive distribution of the output. Based on the explicit expression of the precision matrix in the marginal likelihood, the number of the computational operations is linear to the number of output variables. Moreover, with the use of the Mat{Ã©}rn covariance function, the number of the computational operations is also linear to the number of time points for modeling the multiple time series without any approximation to the likelihood function. We discuss the connection of the GPPCA with other approaches such as the PCA and PPCA, and highlight the advantage of GPPCA in terms of the practical relevance, estimation accuracy and computational convenience. Numerical studies confirm the excellent finite-sample performance of the proposed approach.