Abstract

Today's computerized processes generate massive amounts of streaming data. In many applications, data is collected for modeling the processes. The process model is hoped to drive objectives such as decision support, data visualization, business intelligence, automation and control, pattern recognition and classification, etc. However, we face significant challenges in data-driven modeling of processes. Apart from the errors, outliers and noise in the data measurements, the main challenge is due to a large dimensionality, which is the number of variables each data sample measures. The samples often form a long temporal sequence called a multivariate time series where any one sample is influenced by the others. We wish to build a model that will ensure robust generation, reviewing, and representation of new multivariate time series that are consistent with the underlying process.

In this thesis, we adopt a modeling framework to extract characteristics from multivariate time series that correspond to dynamic variation-covariation common to the measured variables across all the samples. Those characteristics of a multivariate time series are named its 'commonalities' and a suitable measure for them is defined. What makes the multivariate time series model versatile is the assumption regarding the existence of a latent time series of known or presumed characteristics and much lower dimensionality than the measured time series; the result is the well-known 'dynamic factor model'. Original variants of existing methods for estimating the dynamic factor model are developed: The estimation is performed using the frequency-domain equivalent of the dynamic factor model named the 'spectral factor model'. To estimate the spectral factor model, ideas are sought from the asymptotic theory of spectral estimates. This theory is used to attain a probabilistic formulation, which provides maximum likelihood estimates for the spectral factor model parameters. Then, maximum likelihood parameters are developed with all the analysis entirely in the spectral-domain such that the dynamically transformed latent time series inherits the commonalities maximally.

The main contribution of this thesis is a learning framework using the spectral factor model. We term learning as the ability of a computational model of a process to robustly characterize the data the process generates for purposes of pattern matching, classification and prediction. Hence, the spectral factor model could be claimed to have learned a multivariate time series if the latent time series when dynamically transformed extracts the commonalities reliably and maximally. The spectral factor model will be used for mainly two multivariate time series learning applications: First, real-world streaming datasets obtained from various processes are to be classified; in this exercise, human brain magnetoencephalography signals obtained during various cognitive and physical tasks are classified. Second, the commonalities are put to test by asking for reliable prediction of a multivariate time series given its past evolution; share prices in a portfolio are forecasted as part of this challenge.

For both spectral factor modeling and learning, an analytical solution as well as an iterative solution are developed. While the analytical solution is based on low-rank approximation of the spectral density function, the iterative solution is based on the expectation-maximization algorithm. For the human brain signal classification exercise, a strategy for comparing similarities between the commonalities for various classes of multivariate time series processes is developed. For the share price prediction problem, a vector autoregressive model whose parameters are enriched with the maximum likelihood commonalities is designed. In both these learning problems, the spectral factor model gives commendable performance with respect to competing approaches.