THE UNIVERSITY  OF HONG KONG

Institute of Mathematical Research
Department of Mathematics

PROBABILITY AND INFORMATION THEORY SEMINAR

High-Dimensional Model Selection with Latent Variables

Professor Vincent Tan
National University of Singapore

Abstract:

Nowadays enormous amounts of high-dimensional data are constantly being generated by various sources such as high-throughput sequencing technologies and automated clinical databases, which are useful for predicting disease susceptibility. To perform inference on such large datasets, the scientist needs to learn the underlying structure of the data by identifying a statistical model that captures the inter-dependencies among large number of variables. In many important problems, the structure among variables can be well represented by sparse graphical models or low-rank matrices.

I will analyze the error rates of algorithms for learning sparse, in particular tree-structured, graphical models. I will demonstrate that for Gaussian graphical models, among the class of trees, stars are the hardest to learn and Markov chains are the easiest. I will then consider latent tree models where only a subset of variables are observed and the rest are hidden. Algorithms with low computational and sample complexity are derived to infer the latent tree structure. I prove theoretical guarantees in the high-dimensional setting where the number of variables far exceeds the number of samples. Experiments on monthly stock returns of the companies in the S&P 100 reveal that companies of the same nature (computer, retail) are probabilistically clustered in a hierarchical fashion.

For modeling data using low-rank matrices, I will discuss estimating the latent dimensionality in nonnegative matrix factorization (NMF), which serves to decompose complicated objects into their constituent parts (e.g., a music piece into its various tones). Uncovering the correct model order (the number of parts) is important to strike the right balance between data fidelity and overfitting. Such questions had not previously been studied for NMF. I will propose a Bayesian graphical model for NMF and will show that the estimation of the model order and its parameters can be reliably accomplished. Finally, I will demonstrate the efficacy and robustness of our algorithms on a music decomposition example and show that the number of tones can be correctly uncovered.

Biography: Vincent Tan is an Assistant Professor in the Department of Electrical and Computer Engineering (ECE) and the Department of Mathematics at the National University of Singapore (NUS). He received the B.A. and M.Eng. degrees in Electrical and Information Sciences from Cambridge University in 2005. He received the Ph.D. degree in Electrical Engineering and Computer Science (EECS) from the Massachusetts Institute of Technology in 2011. He was a postdoctoral researcher at the University of Wisconsin-Madison and following that, a scientist at the Institute for Infocomm Research (I2R), A*STAR, Singapore. His research interests include information theory, machine learning and statistical signal processing.

Dr. Tan has received several awards including the MIT EECS Jin-Au Kong outstanding doctoral thesis prize in 2011 and the NUS Young Investigator Award in 2014. He is a member of the IEEE "Machine Learning for Signal Processing" Technical Committee and an Associate Editor for Coding and Communication Theory for the IEEE Transactions on Communications.

Date:	June 12, 2015 (Friday)
Time:	4:00 – 5:00pm
Place:	Room 210, Run Run Shaw Bldg., HKU

All are welcome