DENIS TALBOT, Université du Québec à Montréal

Bayesian Adjustment for Confounding: A Look at its Theoretical Basis

It is well known that causal inference and predictive inference are two different concepts. Wang and his collaborators proposed a data-driven method for model selection in a causal inference framework that accounts for uncertainty associated with model selection (Bayesian Adjustment for Confounding). We will discuss this method and examine its theoretical basis. We also show that the method can be justified more rigorously using the graphical framework advocated by Pearl and deriving sufficient conditions to avoid confounding.

SHINPEI IMORI, Hiroshima University, Japan

Theoretical Comparisons with AIC and TIC

In many fields, Akaike's information criterion (AIC) and Takeuchi's information criterion (TIC) are frequently used for selecting the best model among the candidate models. The AIC was derived under the assumption that the candidate model is correctly specified, and the TIC was developed under the assumption that the candidate model is not necessarily correctly specified. In the present study, we discuss the theoretical difference between AIC and TIC through the asymptotic probabilities of selecting the true model by AIC and TIC. Especially, we focus on selections of mean and covariance structures.

AHMED FARAHAT, University of Waterloo

An Efficient Greedy Method for Unsupervised Variable Selection

The selection of relevant variables is a crucial task in data analysis. This work proposes a novel method for unsupervised variable selection, which efficiently selects variables in a greedy manner. We first define an effective criterion for unsupervised variable selection which measures the reconstruction error of the data matrix based on the selected variables, and then present a novel algorithm for greedily minimizing the reconstruction error based on the variables selected so far. Experiments on real datasets demonstrate the effectiveness of the proposed algorithm in comparison to the state-of-the-art methods for unsupervised variable selection.

CHRISTIAN LÉGER, Université de Montréal

A Law of the Single Logarithm for Weighted Sums of Arrays Applied to Bootstrap Model Selection in Regression

We generalize a law of the single logarithm obtained by Qi (1994) and Li et al. (1995) to the case of weighted sums of triangular arrays of random variables. We apply this result to bootstrapping the all-subsets model selection problem in regression, where we show that the popular Bayesian Information Criterion of Schwarz (1978) is no longer asymptotically consistent. Indeed, at the bootstrap level, the weighted sum of random variables used to select the variables is of order \((n \log n)^{1/2}\) instead of \((n \log \log n)^{1/2}\) as with the original variables.

HIROKAZU YANAGIHARA, Hiroshima University, Japan

A Consistency Property of AIC for Multivariate Linear Models when the Dimension and the Sample Size are Large

It is common knowledge that Akaike's information criterion (AIC) is not consistent in the model selection. This fact has been confirmed from an asymptotic selection-probability based on a large-sample asymptotic framework. However, we note that when a high-dimensional asymptotic framework such that the dimension and the sample size are large is used for evaluating a selection-probability, a consistency property of AIC for selecting variables in multivariate linear models can be proved. This
means that a selection-probability of selecting the true model by AIC goes to 1 as the sample size and the dimension are approaching to infinity simultaneously.

MARIKO YAMAMURA, Hiroshima University, Japan

Variable Selection by $C_p$ Statistic in Multiple Responses Regression with Fewer Sample Size than the Dimension

We consider the problem of selecting $q$ explanatory variables out of $k$ ($q \leq k$), when the dimension $p$ of the response variables is larger than the sample size $n$ in the multiple responses regression. We consider $C_p$ statistic which is an estimator of the sum of standardized mean squared errors. The standardization uses the inverse of the variance-covariance matrix of $p$ response variables and thus the estimator of the inverse of the sample variance-covariance matrix. However, since $n < p$, such an inverse matrix cannot be used. Thus, we use the Moore-Penrose inverse and define the $C_p$ statistic.