

Small Sample Issues in Microarray Studies: Sample Size and Error Rate Estimation

Wenjiang Fu PhD

Department of Statistics

Texas A & M University

Abstract:

Microarray technology has gained increasing popularity. It provides great opportunities to screen thousands of genes simultaneously through a small number of samples but also poses great challenges, such as sample size determination, misclassification error rate estimation, and gene selection, due to the special data structure of small sample size and high dimensionality.

In this presentation, I will address several aspects of this general small sample problem. The first topic is the determination of sample size, where conventional sample size calculations may not apply. I will introduce a novel sequential approach, which allows large enough sample size to make sound decisions and yet small enough sample size to make the studies affordable. The second topic is the estimation of misclassification error, where current available methods, such as cross-validation, leave-one-out bootstrap, .632 bootstrap (Efron 1983) and .632+ bootstrap (Efron and Tibshirani 1997), suffer from large variability or high bias. I will propose a novel bootstrap cross-validation (BCV) method of estimating misclassification error with small samples. I will demonstrate the above methods through Monte Carlo simulations and applications to microarray data, although our methods also apply to other types of data, such as clinical diagnosis in medical research. I will also briefly mention my most recent results in Bayesian gene selection through hierarchical models with a novel prior family, the Laplacian – Gaussian mixture (LGM) prior.