Genotype copy number variations using Gaussian mixture models

Chang-Yun Lin (林長鋆) Institute of Statistics, National Chung Hsing University

Abstract

Copy number variations (CNVs) are important in the disease association studies and are usually targeted by most recent microarray platforms developed for GWAS studies. However, the probes targeting the same CNV regions could vary greatly in performance, with some of the probes carrying little information more than pure noise. In this paper, we investigate how to best combine measurements of multiple probes to estimate copy numbers of individuals under the framework of Gaussian mixture model (GMM). First we show that under two regularity conditions and assume all the parameters except the mixing proportions are known, optimal weights can be obtained so that the univariate GMM based on the weighted average gives the exactly the same classification as the multivariate GMM does. We then developed an algorithm that iteratively estimates the parameters and obtains the optimal weights, and uses them for classification. The algorithm performs well on simulation data and two sets of real data, which shows clear advantage over classification based on the equal weighted average.