A protocol to adopt the mixture model by Zhu et al. for the analysis of population stratification on the data with missing allele calls

Suenori Chiku
Mizuho Information & Research Institute, Inc.
Kimio Yoshimura
Keio University School of Medicine
Teruhiko Yoshida
National Cancer Center Research Institute

Method Article

Keywords: population stratification, genome wide association study, principal component analysis, expectation-maximization algorithm

Posted Date: July 10th, 2008

DOI: https://doi.org/10.1038/nprot.2008.129

Abstract

Introduction

There are two kinds of applications of principal component analysis \(PCA) to analyze population substructures of genetic polymorphism data. One application is for an individual covariance matrix, and the other application is for a marker covariance matrix. The former method is already implemented in EIGENSTRAT \[1]; the latter method, however, is not common because it cannot be applied, if data include missing typing data \(allele call). Here, we describe some modification of a Mixture Model \[2] \(MM), so that it can handle data with missing allele calls \(we call it a compensated mixture model \(CMM) protocol). MM applies PCA to a marker covariance matrix before applying the normal-distribution mixture model.

Equipment

1. Genotype data file on markers \(e.g. SNPs in our GWAS on gastric cancer), which were selected so that the marker loci would be independent each other \(an example of such selection criteria is given below for the analysis shown in Table 1 and Figure 1). 2. CMM program module \(please contact us if you want to use our in-house software which is written by C++)

Procedure

The calculation procedures for CMM are as follows: 1. Calculate allele frequencies for each locus. 2. Sample genotype randomly based on the allele frequencies at the missing-data loci for each of the subjects showing missing allele calls of the loci. 3. Calculate M times M marker covariance matrix \(M is the number of marker loci). 4. Calculate eigenvectors up to the 3rd or 4th largest eigenvalues of the covariance matrix. 5. Calculate Bayesian information criterions \(BICs) of the principle components, assuming K normal-distributions mixture models \(K corresponds to the number of subpopulations). 6. Count the inferred subpopulation number K based on minimum BIC. 7. Iterate the above steps from 2 to 6 \(we iterated this procedure 200 times in our paper). The result on the 5,197 SNP typing data on the Chinese and Japanese population of the HapMap project \(SNPs were selected by the following criteria: physical distances among the SNPs are more than 500kbp, minor allele frequency more than 3%, and missing genotype call rate less than 5%) are shown in Table 1 and Figure 1.

References

\[1] Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. & Reich, D. Principal components analysis corrects for stratification in genome-wide association studies, _Nature Genetics_ **38**, 904-909 \(2006). \[2] Zhu, X., Zhang, S., Zhao, H. & Cooper, R.S. Association mapping, using a mixture model for complex traits. _Genet. Epidemiol._ **23**, 181-196 \(2002).

Acknowledgements

This work was supported in Japan by the program for promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation \(NiBio).

Figures

K=1	K=2	K=3	K=4
0	193	7	0

Figure 1

Table 1 The number of counts of the inferred subpopulation number based on Bayesian information criterion for the HapMap Chinese and Japanese data on the 5,197 SNPs.

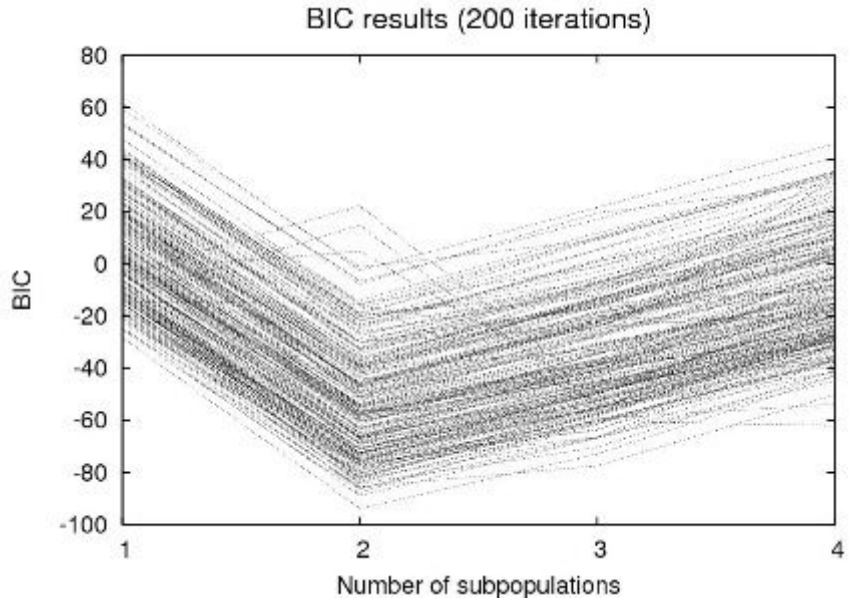

Figure 2

Figure 1 Bayesian information criterion values of the 5,197 SNPs of the HapMap Chinese and Japanese data. A result of 200 iterations of CMM is shown.