Consensus clustering for Bayesian mixture models

Stephen Coleman[1], Paul D.W. Kirk[2] and Chris Wallace[3]

[1] *MRC Biostatistics Unit, University of Cambridge, U.K., stephen.coleman@mrc-bsu.cam.ac.uk*
[2] *MRC Biostatistics Unit, University of Cambridge, U.K., paul.kirk@mrc-bsu.cam.ac.uk*
[3] *MRC Biostatistics Unit, University of Cambridge, U.K., cew54@cam.ac.uk*

Cluster analysis is an integral part of precision medicine and systems biology, used to define groups of patients or biomolecules. However, problems such as choosing the number of clusters and issues with high dimensional data arise consistently. An ensemble approach, such as consensus clustering, can overcome some of the difficulties associated with high dimensional data, frequently exploring more relevant clustering solutions than individual models. Another tool for cluster analysis, Bayesian mixture modelling, has alternative advantages, including the ability to infer the number of clusters present and extensibility. However, inference of these models is often performed using Markov-chain Monte Carlo (MCMC) methods which can suffer from problems such as poor exploration of the posterior distribution and long runtimes. This makes applying Bayesian mixture models and their extensions to 'omics data challenging. We apply consensus clustering [1] to Bayesian mixture models to address these problems.

Consensus clustering of Bayesian mixture models successfully finds the generating structure in our simulation study and captures multiple modes in the likelihood surface. This approach also offers significant reductions in runtime compared to traditional Bayesian inference when a parallel environment is available. We propose a heuristic to decide upon ensemble size and then apply consensus clustering to Multiple Dataset Integration [2], an extension of Bayesian mixture models for integrative analyses, showing consensus clustering can be applied to any MCMC-based clustering method, not just mixture models. We from an integrative analysis of three 'omics datasets for budding yeast and find clusters of co-expressed genes with shared regulatory proteins. We validate these clusters using data external to the analysis. These clusters can help assign likely function to understudied genes, for example *GAS3* clusters with histones active in S-phase, suggesting a role in DNA replication.

References

[1] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*.

[2] Paul Kirk, Jim E Griffin, Richard S Savage, Zoubin Ghahramani, and David L Wild (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*.

Exploring variance contributions in multiple high-dimensional data sources

Erica Ponzi[1], **Magne Thoresen** [1], **Kajsa Møllersen** [2]

[1] *Oslo Center for Biostatistics and Epidemiology, University of Oslo, Norway*
[2] *UiT, The Arctic University of Norway, Tromsø, Norway*
erica.ponzi@medisin.uio.no

The simultaneous analysis of multiple sources of high-dimensional data is nowadays a major challenge in several research areas. In cancer genomics, data collected on several omic platforms provide information both in form of individual patterns within each data source and of joint patterns that are shared among the different sources. Capturing these two components of variation can help provide a broader understanding of cancer genetics.

Several methods have been proposed to separate common and distinct components of variation in multiple data sources, based on different algorithms and frameworks. For instance, Joint and Individual Variation Explained (JIVE) [1] is based on an iterative algorithm to factorize the data matrix into low rank approximations that capture variation across and within data types. It has been widely used in integrative genomics, and several generalizations and improvements have been developed, such as the angle based JIVE (aJIVE) [2]. On the other hand, integrated PCA (iPCA) [3] is a model based generalization of principal components analysis that can be used in similar applications.

In this work, we compare these three methods and describe their application to a lung cancer case control study nested in the Norwegian Woman and Cancer (NOWAC) cohort study [4]. JIVE, aJIVE and iPCA are used to separate the joint and individual contributions of DNA methylation, miRNA and mRNA expression and to evaluate how this decomposition can be useful to predict the occurrence of lung cancer.

References

[1] Lock, E. F., Hoadley, K. A., Marron, J. S. and Nobel, A. B. (2013). Joint and Individual Variation Explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics*, **7**, 523 – 542.

[2] Feng, Q., Jiang, M., Hannig and J., Marron, J. S. (2018). Angle based joint and individual variation explained. *Journal of Multivariate Analysis*, **166**, 241 – 265.

[3] Tang, T. M. and Allen, G. I. (2018). Integrated principal components analysis. *arXiv*, 1810.00832.

[4] Lund, E., Dumeaux, V., Braaten, T., Hjartåker, A., Engeset, D., Skeie, G. and Kumle, M. (2008) Cohort profile: the Norwegian Women and Cancer Study—NOWAC—Kvinner og kreft. *International Journal of Epidemiology*, **37**, 36 – 41.

Genome-wide association studies with imbalanced binary responses

Pål Vegard Johnsen[1,2], Øyvind Bakke[1], Thea Bjørnland[1] and Mette Langaas[1]

[1] *NTNU, Trondheim, Norway*
[2] *SINTEF DIGITAL, Oslo, Norway*

In genome-wide association studies, one is searching for associations between specific genetic markers (SNPs) and for instance a disease. An association is screened for multiple genetic markers via a logistic regression model and by using the score test statistic. The score test statistic can be shown to follow an asymptotic normal distribution under the null hypothesis. However, practically this approximation is not sufficiently accurate under certain conditions when using binary responses. An improvement is proposed in [1] using saddlepoint approximation theory.

Saddlepoint approximation theory is a large domain within Statistics with many applications. In this work we will analyse different approaches on how to estimate the score statistic in GWAS using saddlepoint approximation theory, and how one should evaluate the accuracy in each approach. New features about the distribution of the score test statistic under the null hypothesis is revealed, and based on this we will conclude with what is the most robust saddlepoint approximation approach for GWAS with binary phenotypes. We will apply our methods on GWAS data from UK Biobank.

References

[1] Dey, R. et al. (2017). A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS *The American Journal of Human Genetics*, **101**, 37 – 49.