COMBINATORIAL MIXTURES OF MULTIPARAMETER DISTRIBUTIONS, WITH AN APPLICATION TO MICROARRAY DATA

Valeria Edefonti[1] and Giovanni Parmigiani[2,3]

[1]Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy
[2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA
[3]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA

The term *'combinatorial mixtures'* refers to a flexible class of models for inference on mixture distributions [4] whose components have multidimensional parameters. The idea behind it is to allow each element of component-specific parameter vectors to be shared by a subset of other components.
We develop Bayesian inference and computational approaches for this class of distributions. We define a structure for a general prior distribution where a positive probability is put on every possible combination of sharing patterns, whence the name combinatorial mixtures. This partial sharing allows for generality and flexibility in comparison with traditional approaches to mixture modeling, while still allowing to assign significant mass to models that are more parsimonious than the general mixture case in which no sharing takes place. This also unifies the inference on component-specific parameters with that on the number of components.
We illustrate our combinatorial mixtures in an application based on the normal model. We introduce normal mixture models for univariate and bivariate data, which are amenable to Markov Chain Monte Carlo computing. In the light of combinatorial mixtures, we assume a decomposition of the variance-covariance matrix proposed by Barnard et al. (2000) [1], which separates out standard deviations and correlations, and thus allows to model those parameters separately.
This development was originally motivated by applications in molecular biology, where one deals with continuous measures, such as RNA levels, or protein levels, that vary across unknown biological subtypes. In some cases, subtypes are characterized by an increase in the level of the marker measured, while in others they are characterized by variability in otherwise tightly controlled processes, or by the presence of otherwise weak correlations. Also, several mechanisms can coexist. It may also allow to model an interesting phenomenon observed in microarray analysis when two variables have the same mean and variance but opposite correlations in diseased and normal samples [2]. We use data on molecular classification of lung cancer from the web-based information supporting the published manuscript Garber et al. (2001) [3].

References

[1] Barnard, J, McCulloch, RE and Meng, XL (2000) Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. Statistica Sinica 10:1281-1311.
[2] Dettling, M, Gabrielson, E and Parmigiani, G (2005) Searching for differentially expressed gene combinations. Genome Biology 6(10): R88.
[3] Garber, ME, Troyanskaya, OG, Schluens K et al. (2001) Diversity of gene expression in adenocarcinoma of the lung. Proc. National Academy of Science USA 98(24):13784-13789
[4] Marin, JM, Mengersen, K and Robert, CP (2011) Bayesian modelling and inference on mixtures of distributions. In: *Essential Bayesian models*, Handbook of Statistics: Bayesian thinking - modeling and computation **25**, (eds. D. Dey and C. R. Rao) Elsevier- Sciences.