

## **Combinatorial Mixtures of Multiparameter Distributions: an Application to Bivariate Data**

Valeria Edefonti<sup>1</sup> and Giovanni Parmigiani<sup>2</sup>

<sup>1</sup> Istituto di Statistica Medica e Biometria "G. A. Maccacaro", Università degli Studi di Milano, ITALY

<sup>2</sup> The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD, U.S.A.

*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 computation approaches for this class of distributions. We define a general prior distribution structure where a positive probability is put on every possible combination of sharing patterns, whence the name combinatorial mixtures. This partial sharing allows for greater 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.

We illustrate our combinatorial mixtures in an application based on the normal model. We propose models for univariate and bivariate data. In the light of combinatorial mixtures, we adopt the decomposition of variance-covariance matrix proposed by Barnard et al. (2000) [1]. It separates the standard deviations and correlations, thus allowing to model those parameters separately. We develop a customized strategy to sample the correlation coefficient for each group. This work was originally motivated by the analysis of cancer subtypes: in terms of biological measures of interest, subtypes may be characterized by differences in location, scale, correlations or any of the combinations. It may also allow to model an interesting phenomenon observed in microarray analysis when two variables have the same mean and variances 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 (2005) Bayesian modelling and inference on mixtures of distributions. In: *Handbook of Statistics*, (eds. D. Dey and C. R. Rao) Elsevier-Sciences, 459-507