A Set-Based Approach to the Decomposition of Linear Mixtures Using Quasi-Convex Programming

J. Verwaeren, M. Rademaker, B. De Baets
Ghent University, KERMIT,
Department of Mathematical Modelling,
Statistics and Bioinformatics
Jan.Verwaeren@UGent.be

A multivariate process can often be interpreted as a mixture of multiple source processes. Several applications require an estimate of the proportional contribution of at least one of these sources to such a mixture. To be able to estimate the proportional contribution of a source of interest to a given mixture, we typically need a formal representation of both the mixture and the sources. Additionally, we need to make assumptions concerning the relationship between these representations and the proportional contribution of the sources to the mixture. Let us illustrate this with a simple example :

Assume you have been given a cup of water with a salinity of 21 ppt (parts per thousand). Moreover, it is told that the water in this cup is a blend of fresh water (salinity of 2 ppt) and sea water (salinity of 40 ppt). Subsequently, you are asked to estimate the proportional amount of fresh water (x_1) in the cup. You might be tempted to give $x_1 = 50\%$ as an answer to this question.

In the above example, the representations of the mixture ($\mathbf{y}$), the source of fresh water ($\mathbf{c}_1 = 2$) and the source of sea water ($\mathbf{c}_2 = 40$) are their salinity. The following assumption

$$\mathbf{y} = x_1\mathbf{c}_1 + x_2\mathbf{c}_2\,, \qquad \text{s.t. } x_1, x_2 \geq 0 \text{ and } x_1 + x_2 = 1\,,$$

leads to the estimate of $x_1 = 50\%$. This assumption (relating the salinity to the proportional contributions) is generally known as the linear mixing model (LMM) assumption. In this example, where the representations of the mixture and the sources are scalars, one can readily obtain an estimate for x_1 and x_2. However, in most applications, the information about the sources is less explicit. For instance, we might only know that $\mathbf{c}_1 \in [1, 3]$ and $\mathbf{c}_2 \in [30, 45]$. In this setting, any solution to the system

$$\begin{aligned} \mathbf{y} &= x_1\mathbf{c}_1 + x_2\mathbf{c}_2\,, \\ 1 &= x_1 + x_2\,, \quad x_1 \geq 0\,, x_2 \geq 0 \\ \mathbf{c}_1 &\in [1, 3]\,, \quad \mathbf{c}_2 \in [30, 45]\,, \end{aligned}$$

can be considered as a possible estimate for x_1 and x_2. It can easily be shown that the solution set of x_1 (and x_2) is a closed interval. The minimal (resp. maximal) element of this interval can be found by solving the following minimization (resp. maximization) problem :

$$\min_{x_1,x_2,\mathbf{c}_1,\mathbf{c}_2} x_1 \qquad (\text{resp.} \max_{x_1,x_2,\mathbf{c}_1,\mathbf{c}_2} x_1) \tag{1}$$

s.t.

$$\mathbf{y} = x_1\mathbf{c}_1 + x_2\mathbf{c}_2\,, \tag{2}$$

$$1 = x_1 + x_2\,, \quad x_1 \geq 0\,, x_2 \geq 0 \tag{3}$$

$$\mathbf{c}_1 \in [1,3]\,, \quad \mathbf{c}_2 \in [30,45]\,. \tag{4}$$

In this toy example where $\mathbf{y}, \mathbf{c}_1, \mathbf{c}_2 \in \mathbb{R}$, there exists a simple, closed-form solution. However, the problem becomes non-trivial in the more general setting, where $\mathbf{y}, \mathbf{c}_1, \mathbf{c}_2 \in \mathbb{R}^d$; and (4) are replaced by $\mathbf{c}_1 \in \mathbf{C}_1$ and $\mathbf{c}_2 \in \mathbf{C}_2$ with $\mathbf{C}_1$ and $\mathbf{C}_2$ two convex subsets of $\mathbb{R}^d$. We formulate a quasi-convex optimization problem, that is shown to be equivalent to the original problem, and propose a procedure to efficiently solve it.

Examples of this setting occur in the detection of fraudulent adulteration of vegetable oils, the spectral unmixing of mixed pixels in remote sensing, and the analysis of differential gene expression experiments.