Imputation of Missing Data Items under Linear Restrictions

Caren Tempelman

Statistics Netherlands

Problem definition

Economic data consist of many linear constraints, such as:

$$c_1 X_1 + c_2 X_2 + \cdots + c_k X_k = X_{k+1}$$

If several X_i's are missing, how do we impute these items immediately satisfying the linear constraint and preserving the distribution of the data?

Approach

The linear constraint can be transformed as follows in order to restrict the domain of the variables to the simplex:

$$\frac{c_1 X_1}{X_{k+1}} + \frac{c_2 X_2}{X_{k+1}} + \cdots + \frac{c_k X_k}{X_{k+1}} = 1 \quad \Longrightarrow$$

$$\tilde{X}_1 + \tilde{X}_2 + \cdots + \tilde{X}_k = 1$$

The Dirichlet distribution and its advantages

$$f(x_1,\ldots,x_k \mid \alpha_1,\ldots,\alpha_k) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}x_i^{\alpha_i-1}$$

where $x_i \geq 0, \quad i = 1,...,k, \text{ and } x_1 + \cdots + x_k = 1$

with parameters $\alpha_i > 0, \quad i = 1,..., k.$

Advantages

- Extremely flexible in the shapes it will accommodate
- On the simplex