Imputation for economic data under linear restrictions

Discussion paper 04002

Caren Tempelman

*Ton Steerneman**

The authors would like to thank Ton de Waal and Jeroen Pannekoek for their helpful comments and suggestions.

* Department of econometrics, University of Groningen

Statistics Netherlands

Voorburg/Heerlen, March 2004

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2003–2004	= 2003 to 2004 inclusive
2003/2004	= average of 2003 up to and including 2004
2003/'04	= crop year, financial year, school year etc. beginning in 2003 and ending in 2004

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg
The Netherlands

Printed by
Statistics Netherlands - Facility Services

Cover design
WAT ontwerpers, Utrecht

Information
E-mail: infoservice@cbs.nl

Where to order
E-mail: verkoop@cbs.nl

Internet
http://www.cbs.nl

ISSN: 1572-0314
Key figure: X-10
Production code: 6008304002

Statistics Netherlands

IMPUTATION FOR ECONOMIC DATA UNDER LINEAR RESTRICTIONS

Abstract: In this paper we propose a new approach to impute data under linear restrictions. If the data are normally distributed we will use simulations from the standard normal distribution. If the data are not normally distributed we study the use of the Dirichlet distribution.

Keywords: Imputation, linear restrictions, normal distribution, Dirichlet distribution, EM algorithm

1 Introduction

Missing data is a prevalent problem in survey analysis. At Statistics Netherlands imputation is often used to estimate and fill in missing data items, since missing values result in less efficient estimates due to the reduction of the sample size of the dataset. Besides, if the nonrespondents are judged to be significantly different from the respondents on the basis of auxiliary information, then imputation reduces the nonresponse bias. Finally, imputation is applied because standard complete data methods such as regression analysis cannot immediately be used to analyse data when items are missing.

Several imputation methods have been developed, see for an overview of the methods that are frequently used, for example, Little and Rubin (2002), Rubin (1987) and Kalton and Kasprzyk (1986). Imputation methods can be either deterministic or stochastic. Deterministic methods determine imputed values uniquely, this means that when the imputation process is repeated the same value will be imputed. Stochastic methods depend on some sort of randomness, which means that when the process is repeated, other values may be imputed. Deterministic imputation methods avoid the loss in precision associated with the added randomness as opposed to stochastic methods. Therefore these methods are well suited to estimate means or totals. However, the variance will be underestimated and the shape of the distribution will be distorted. So for the creation of general purpose datasets, stochastic imputation is preferred.

A property of economic data is that there are many logical constraints on the data items, such as the fact that company profits must equal turnover minus expenses. Commonly used imputation methods such as hot deck and (random) regression imputation mostly do not provide imputations that satisfy these linear restrictions.

In this paper we therefore suggest the use of other imputation schemes to obtain imputations that will satisfy these linear restrictions. In section 2 consistent

imputation is discussed in the case of normally distributed data. In section 3 consistent imputation is discussed when the data are not normally distributed. The distribution used to model these data is the Dirichlet distribution. Random number generation and parameter estimation will also be treated as well as the EM algorithm. We also touch upon the subject of more complex edit structures in a discussion on directions for future research.

2 Imputation under restrictions when the data are multivariate normally distributed

In general there are two types of linear edit restrictions: balance and inequality edit restrictions. They can be written as follows

$$c_1X_1 + \cdots + c_{k-1}X_{k-1} = X_k \quad (1)$$

$$c_1X_1 + \cdots + c_{k-1}X_{k-1} \leq X_k \quad (2)$$

We will first discuss balance edit restrictions. Consider a data vector $\mathbf{X}$, where $\mathbf{X} = (X_1\ X_2\ \cdots\ X_k)'$. Assume that the c_j, $j = 1, \ldots, k-1$ are all equal to one, so $X_k = \sum_{j=1}^{k-1} X_j$. It is assumed that $\mathbf{X}$ is k-variate normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Since $X_k = X_1 + \cdots + X_{k-1}$, the covariance matrix $\boldsymbol{\Sigma}$ will be singular. Note that $\boldsymbol{\Sigma}$ will be singular for any value of c_j, $j = 1, \ldots, k-1$. By means of an eigenvalue decomposition $\boldsymbol{\Sigma}$ can be decomposed into $\mathbf{C}\boldsymbol{\Lambda}\mathbf{C}'$, where $\mathbf{C}$ is the orthogonal matrix of eigenvectors and $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{\Sigma}$ on the main diagonal. Since $\boldsymbol{\Sigma}$ is singular, one or more eigenvalues will be equal to zero. In this case only one eigenvalue will be equal to zero because there is only one linear balance constraint. Thus $\boldsymbol{\Lambda}$= diag$\{\lambda_1, \ldots, \lambda_{k-1}, 0\}$ and we will assume $\lambda_1 \geq \cdots \geq \lambda_{k-1} > 0$.
Transformation of $\mathbf{X}$ will lead to the following:

$$\mathbf{C}'(\mathbf{X} - \boldsymbol{\mu}) \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Lambda})$$

Now create $\mathbf{Z}$ of order $k \times 1$, which consists of drawings from the standard normal distribution for $j = 1, \ldots, k-1$ and zero for the kth entry. Calculate $\tilde{\mathbf{Z}} = \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{Z}$, with $\boldsymbol{\Lambda}^{\frac{1}{2}}$=diag$\{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_{k-1}}, 0\}$. Note that

$$\tilde{\mathbf{Z}} \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Lambda})$$

So finally we calculate

$$\mathbf{X} = \mathbf{C}\tilde{\mathbf{Z}} + \boldsymbol{\mu}$$

We are now able to generate values from the multivariate singular normal distribution of $\mathbf{X}$ that satisfy the linear balance edit restriction, using this singularity. However, there are probably also some available item values. For instance, when

only X_1 and X_2 are missing and the other X's are observed. We then want to estimate the values of X_1 and X_2 taking the values of the other observed X's into account. For this purpose we need to make use of the conditional distribution of X_1 and X_2 given $X_3, \ldots, X_k$.
Due to the fact that $\mathbf{X}$ is multivariate normally distributed, the following holds. Let us partition $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as $\mathbf{X} = (\mathbf{Y}_1' \ \mathbf{Y}_2')'$, where

$$\mathbf{Y}_1 = \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix} \text{ and } \mathbf{Y}_2 = \begin{pmatrix} X_{m+1} \\ \vdots \\ X_k \end{pmatrix}$$

If we partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ accordingly we have

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \sim \mathcal{N}_k(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix})$$

It is well known that

$$\mathbf{Y}_1 \mid \mathbf{Y}_2 \ \sim \ \mathcal{N}_m(\boldsymbol{\mu}_{1.2}, \boldsymbol{\Sigma}_{11.2})$$

where

$$\begin{aligned} \boldsymbol{\mu}_{1.2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{Y}_2 - \boldsymbol{\mu}_2), \\ \boldsymbol{\Sigma}_{11.2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{aligned}$$

For a detailed proof see e.g. Anderson (1984). Due to the fact that $\sum_{j=1}^{m} X_j = X_k - \sum_{j=m+1}^{k-1} X_j$, the covariance matrix $\boldsymbol{\Sigma}_{11.2}$ is singular. This means that again we can decompose this matrix by means of an eigenvalue decomposition: $\boldsymbol{\Sigma}_{11.2} = \tilde{\mathbf{C}}\tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{C}}'$, where $\tilde{\boldsymbol{\Lambda}} = \text{diag}\{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{m-1}, 0\}$ with $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_{m-1} > 0$. Therefore

$$\begin{aligned} \mathbf{Y}_1 \mid \mathbf{Y}_2 \ &\sim \ \mathcal{N}_m(\boldsymbol{\mu}_{1.2}, \tilde{\boldsymbol{C}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{C}}' \ ) \\ \tilde{\mathbf{C}}'(\mathbf{Y}_1 - \boldsymbol{\mu}_{1.2}) \mid \mathbf{Y}_2 \ &\sim \ \mathcal{N}_m(\mathbf{0}, \tilde{\boldsymbol{\Lambda}}) \end{aligned}$$

Next generate $Z_j \sim \mathcal{N}(0,1)$ for $j = 1, \ldots, m-1$, and take $Z_m = 0$, now calculate $\tilde{\mathbf{Z}} = \tilde{\boldsymbol{\Lambda}}^{\frac{1}{2}}\mathbf{Z}$. So

$$\tilde{\mathbf{Z}} \sim \mathcal{N}_m(\mathbf{0}, \tilde{\boldsymbol{\Lambda}})$$

Thus

$$\begin{aligned} \tilde{\mathbf{C}}'(\mathbf{Y}_1 - \boldsymbol{\mu}_{1.2}) \mid \mathbf{Y}_2 &= \tilde{\mathbf{Z}} \\ \mathbf{Y}_1 \mid \mathbf{Y}_2 &= \tilde{\mathbf{C}}\tilde{\mathbf{Z}} + \boldsymbol{\mu}_{1.2} \end{aligned}$$

2.1 Estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

The next issue we need to deal with is the fact that we use $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to generate imputations, but the presence of missing data complicates the estimation of $\boldsymbol{\mu}$

and $\boldsymbol{\Sigma}$.

The Expectation-Maximization (EM) algorithm has been developed by Dempster, Laird and Rubin (1977) to calculate maximum likelihood estimates in the presence of missing data. The EM algorithm is a popular tool in statistics and is widely used. For a detailed description of the EM algorithm for multivariate normally distributed data see Little and Rubin (2002) and Schafer (1997).

Let $\mathbf{X}$ denote the complete data matrix. Denote the observed part of $\mathbf{X}$ by $\mathbf{X}_{obs}$ and the missing part by $\mathbf{X}_{mis}$, so that $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$. The intuition behind the EM algorithm is that the missing data matrix $\mathbf{X}_{mis}$ and the estimates for the parameters, in this case $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, are interdependent. The (function of) $\mathbf{X}_{mis}$ appearing in the loglikelihood function will be estimated based on an initial estimate of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and the observed data $\mathbf{X}_{obs}$. Next we re-estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ based on $\mathbf{X}_{obs}$ and the values that have been filled in for (the function of) $\mathbf{X}_{mis}$ and iterate until the estimates converge. The idea behind the algorithm is that we would like to maximize the complete data likelihood but since we do not know it, we maximize its expectation instead. The EM algorithm consists of two steps, an expectation step and a maximization step.

Expectation step.

For each iteration t compute $\ell_t(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathrm{E}[\ell(\boldsymbol{\mu},\boldsymbol{\Sigma}) \mid \mathbf{X}_{obs}, \boldsymbol{\mu}^{t-1},\boldsymbol{\Sigma}^{t-1}]$, where $\ell(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the complete data loglikelihood and the expectation is taken with respect to the conditional distribution of the missing data given the observed data and the parameter estimates $\boldsymbol{\mu}^{t-1}$ and $\boldsymbol{\Sigma}^{t-1}$.

In the case of distributions from a regular exponential family we only need to calculate the expected sufficient statistics in this step.

Maximization step.

Once we have $\ell_t(\boldsymbol{\mu},\boldsymbol{\Sigma})$, we can calculate the maximum likelihood estimates based on the complete data loglikelihood, and thus re-estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

The main disadvantage of EM is its slow convergence rate, which is shown to be linear, especially compared to the quadratic convergence rate of the Newton-Raphson algorithm. Nevertheless, the convergence of EM is monotonic and the algorithm is assured to converge to a local optimum, that is: $\ell(\boldsymbol{\mu}^{t+1},\boldsymbol{\Sigma}^{t+1}) \geq \ell(\boldsymbol{\mu}^{t},\boldsymbol{\Sigma}^{t})$, see Dempster, Laird and Rubin (1977) for a detailed description and proof of these properties.

The EM algorithm can be applied to normally distributed data as follows. Suppose that $\mathbf{X}_i,\ i = 1, \ldots, n$ of order $k \times 1$ contains the item responses for record i. So the complete data matrix is $\mathbf{X} = (X_{ij})_{i=1,\ldots,n,\ j=1,\ldots,k}$. For each record i subdivide the items into an observed part and a missing part, thus $\mathbf{X}_i' = (\mathbf{X}_{i,obs}', \mathbf{X}_{i,mis}')$. Let $\boldsymbol{\mu}^t$ and $\boldsymbol{\Sigma}^t$ denote the parameter estimates at iteration t.

In the E-step the complete data loglikelihood is calculated by taking the expectation of the distribution of the missing data given the observed data $\mathbf{X}_{obs}$ and the current parameter estimates $\boldsymbol{\mu}^t$ and $\boldsymbol{\Sigma}^t$. Since the normal distribution is an exponential family we only need to calculate the expectations of the sufficient statistics.

Recall that the sufficient statistics of the normal distribution are

$$\begin{aligned} T_1(\mathbf{X}) &= \sum_{i=1}^{n} X_{ij}, \qquad j = 1, \ldots, k \\ T_2(\mathbf{X}) &= \sum_{i=1}^{n} X_{ij} X_{il}, \quad j, l = 1, \ldots, k \end{aligned}$$

The expectation step of the algorithm consists of calculating the expectations of the sufficient statistics which means that $\mathrm{E}[X_{ij} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t]$ and $\mathrm{E}[X_{ij}X_{il} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t]$, for $i = 1, \ldots, n$, and $j, l = 1, \ldots, k$ need to be calculated.

$$\mathrm{E}[X_{ij}] = \begin{cases} X_{ij} & \text{if } X_{ij} \text{ is observed} \\ X_{ij}^* & \text{if } X_{ij} \text{ is missing} \end{cases}$$

and

$$\mathrm{E}[X_{ij}X_{il}] = \begin{cases} X_{ij}X_{il} & \text{if } X_{ij} \text{ and } X_{il} \text{ are both observed} \\ X_{ij}^*X_{il} & \text{if } X_{ij} \text{ is missing and } X_{il} \text{ is observed} \\ X_{ij}X_{il}^* & \text{if } X_{ij} \text{ is observed and } X_{il} \text{ is missing} \\ X_{ij}^*X_{il}^* + \mathrm{Var}(X_{ij}^*, X_{il}^*) & \text{if } X_{ij} \text{ and } X_{il} \text{ are both missing} \end{cases}$$

where X_{ij}^* are the elements of $\mathbf{X}_{i,mis}$ calculated by

$$\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t] = \boldsymbol{\mu}_{mis}^t + \boldsymbol{\Sigma}_{mis,obs}^t (\boldsymbol{\Sigma}_{obs,obs}^t)^{-1} (\mathbf{X}_{i,obs} - \boldsymbol{\mu}_{obs}^t)$$

and $\mathrm{Var}(X_{ij}^*, X_{il}^*)$ are the elements of the covariance matrix of the missing items calculated by

$$\mathrm{Var}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t) = \boldsymbol{\Sigma}_{mis,mis}^t - \boldsymbol{\Sigma}_{mis,obs}^t (\boldsymbol{\Sigma}_{obs,obs}^t)^{-1} \boldsymbol{\Sigma}_{obs,mis}^t$$

Note that $\mathrm{Var}(X_{ij}, X_{il})$, $\mathrm{Var}(X_{ij}^*, X_{il})$ and $\mathrm{Var}(X_{ij}, X_{il}^*)$ all are equal to zero since one of the X's is observed and thus regarded as fixed. Also note that X_{ij}^* and $\mathrm{Var}(X_{ij}^*, X_{il}^*)$ are dependent on t, which is left out for ease of notation. Now define

$$\mathrm{V}^t(\mathbf{X}_{i,mis})_{jl} = \begin{cases} \mathrm{Var}(X_{ij}^*, X_{il}^*) & \text{if } X_{ij} \text{ and } X_{il} \text{ are both missing} \\ 0 & \text{otherwise} \end{cases}$$

For the M-step these estimated complete data sufficient statistics are used to calculate the maximum likelihood estimates, and thus re-estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. So $\mathbf{X}_i$ will become $\mathbf{X}_i' = (\mathbf{X}_{i,obs}', \mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t]')$ and the new estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\boldsymbol{\mu}^{t+1} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$$

$$\begin{aligned}\boldsymbol{\Sigma}^{t+1} &= \frac{1}{n}\sum_{i=1}^{n}\Big((\mathbf{X}_i-\boldsymbol{\mu}^{t+1})(\mathbf{X}_i-\boldsymbol{\mu}^{t+1})'+\mathrm{V}^t(\mathbf{X}_{i,mis})\Big)\\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbf{X}_i\mathbf{X}_i'+\mathrm{V}^t(\mathbf{X}_{i,mis})\Big)-\frac{1}{n}\boldsymbol{\mu}^{t+1}(\boldsymbol{\mu}^{t+1})'\end{aligned}$$

2.2 Why the linear balance restriction is always satisfied

It is crucial that the estimates generated by the EM algorithm satisfy the linear balance restriction on the data. In this section we will show that they do. If $\mathbf{A}'\boldsymbol{\mu}^t = 0$ with $\mathbf{A}$ a $k \times 1$ vector representing the restriction, it should hold that $\mathbf{A}'\boldsymbol{\mu}^{t+1} = 0$. We know that this restriction holds for the observed complete cases of the data matrix $\mathbf{X}$, so $\mathbf{A}'\mathbf{X} = \mathbf{0}$. This means that for each nonrespondent i the following holds for the respondents with no missing items: $\mathbf{A}'_{i,mis}\mathbf{X}_{h,mis} = -\mathbf{A}'_{i,obs}\mathbf{X}_{h,obs}$, where h is a respondent without any item nonresponse and $\mathbf{A}_{i,mis}$ is the vector with the restriction on the missing items for the nonrespondent. Of course it also holds that $\mathbf{A}'_{i,mis}\boldsymbol{\mu}_{mis} = -\mathbf{A}'_{i,obs}\boldsymbol{\mu}_{obs}$ for each missing data pattern.

Now consider a record $\mathbf{X}_i$ which contains missing values. Subdivide $\mathbf{X}_i$ into an observed and a missing part, thus $\mathbf{X}'_i = (\mathbf{X}'_{i,obs}, \mathbf{X}'_{i,mis})$. The set of respondents without item nonresponse is denoted by $\mathcal{R}$ and the number of respondents without item nonresponse by r.

In the EM algorithm the missing items for nonrespondent i will be replaced by their expected values:

$$\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t] = \boldsymbol{\mu}^t_{mis} + \boldsymbol{\Sigma}^t_{mis,obs}(\boldsymbol{\Sigma}^t_{obs,obs})^{-1}(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^t_{obs})$$

We will start at $t = 0$. Assuming there are sufficient complete cases, the complete cases mean and covariance matrix will be used as starting values, which means that $\mathbf{A}'\boldsymbol{\mu}^0 = 0$.

The imputed items need to satisfy the restriction, this means that the following has to hold

$$\mathbf{A}'_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0] = -\mathbf{A}'_{i,obs}\mathbf{X}_{i,obs}$$

Now calculate

$$\begin{aligned}\mathbf{A}'_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0] &= \mathbf{A}'_{i,mis}\boldsymbol{\mu}^0_{mis} + \\ &\quad \mathbf{A}'_{i,mis}\boldsymbol{\Sigma}^0_{mis,obs}(\boldsymbol{\Sigma}^0_{obs,obs})^{-1}(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^0_{obs})\end{aligned}$$

where

$$\begin{aligned}\mathbf{A}'_{i,mis}\boldsymbol{\Sigma}^0_{mis,obs} &= \mathbf{A}'_{i,mis}\frac{1}{r}\sum_{h\in\mathcal{R}}\Big((\mathbf{X}_{h,mis}-\boldsymbol{\mu}^0_{mis})(\mathbf{X}_{h,obs}-\boldsymbol{\mu}^0_{obs})'\Big)\\ &= \frac{1}{r}\sum_{h\in\mathcal{R}}\Big((\mathbf{A}'_{i,mis}\mathbf{X}_{h,mis}-\mathbf{A}'_{i,mis}\boldsymbol{\mu}^0_{mis})(\mathbf{X}_{h,obs}-\boldsymbol{\mu}^0_{obs})'\Big)\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{r}\sum_{h\in\mathcal{R}}\Big((-\mathbf{A}'_{i,obs}\mathbf{X}_{h,obs}+\mathbf{A}'_{i,obs}\boldsymbol{\mu}^0_{obs})(\mathbf{X}_{h,obs}-\boldsymbol{\mu}^0_{obs})'\Big)\\
&= -\mathbf{A}'_{i,obs}\frac{1}{r}\sum_{h\in\mathcal{R}}\Big((\mathbf{X}_{h,obs}-\boldsymbol{\mu}^0_{obs})(\mathbf{X}_{h,obs}-\boldsymbol{\mu}^0_{obs})'\Big)\\
&= -\mathbf{A}'_{i,obs}\boldsymbol{\Sigma}^0_{obs,obs}
\end{aligned}
$$

So

$$
\begin{aligned}
\mathbf{A}'_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis}\mid\mathbf{X}_{i,obs},\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0] &= \mathbf{A}'_{i,mis}\boldsymbol{\mu}^0_{mis}-\mathbf{A}'_{i,obs}\mathbf{X}_{i,obs}+\mathbf{A}'_{i,obs}\boldsymbol{\mu}^0_{obs}\\
&= -\mathbf{A}'_{i,obs}\mathbf{X}_{i,obs}
\end{aligned}
$$

The mean, $\boldsymbol{\mu}^1$ is calculated based on the observed and estimated data in the previous step and since $\mathbf{A}'_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis}\mid\mathbf{X}_{i,obs},\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0]=-\mathbf{A}'_{i,obs}\mathbf{X}_{i,obs}$ for the imputed data, it will also hold that $\mathbf{A}'_{i,mis}\boldsymbol{\mu}^1_{mis}=-\mathbf{A}'_{i,obs}\boldsymbol{\mu}^1_{obs}$.
Now $\boldsymbol{\Sigma}$ will be re-estimated by

$$
\boldsymbol{\Sigma}^1 = \frac{1}{n}\sum_{i=1}^{n}\Big((\mathbf{X}_i-\boldsymbol{\mu}^1)(\mathbf{X}_i-\boldsymbol{\mu}^1)'+\mathrm{V}^0(\mathbf{X}_{i,mis})\Big)
$$

For each record i it holds that $\mathbf{A}'\mathrm{V}^0(\mathbf{X}_{i,mis})=\mathbf{0}$. To show that this holds subdivide $\mathbf{A}$ into a missing and an observed part and subdivide the matrix $\mathrm{V}^0(\mathbf{X}_{i,mis})$ similarly. Now

$$
\begin{aligned}
\mathbf{A}'\mathrm{V}^0(\mathbf{X}_{i,mis}) &= (\mathbf{A}'_{i,mis}\ \mathbf{A}'_{i,obs})\begin{pmatrix}\mathrm{Var}(\mathbf{X}_{i,mis}\mid\mathbf{X}_{i,obs},\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0) & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\\
&= (\mathbf{A}'_{i,mis}\mathrm{Var}(\mathbf{X}_{i,mis}\mid\mathbf{X}_{i,obs},\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0),0,\ldots,0)
\end{aligned}
$$

where

$$
\mathrm{Var}(\mathbf{X}_{i,mis}\mid\mathbf{X}_{i,obs},\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0)=\boldsymbol{\Sigma}^0_{mis,mis}-\boldsymbol{\Sigma}^0_{mis,obs}(\boldsymbol{\Sigma}^0_{obs,obs})^{-1}\boldsymbol{\Sigma}^0_{obs,mis}
$$

So

$$
\begin{aligned}
\mathbf{A}'_{i,mis}\mathrm{Var}(\mathbf{X}_{i,mis}\mid\mathbf{X}_{i,obs},\boldsymbol{\mu}^0,\boldsymbol{\Sigma}^0) &= \mathbf{A}'_{i,mis}\boldsymbol{\Sigma}^0_{mis,mis}-\\
&\quad\ \mathbf{A}'_{i,mis}\boldsymbol{\Sigma}^0_{mis,obs}(\boldsymbol{\Sigma}^0_{obs,obs})^{-1}\boldsymbol{\Sigma}^0_{obs,mis}\\
&= -\mathbf{A}'_{i,obs}\boldsymbol{\Sigma}^0_{obs,mis}+\mathbf{A}'_{i,obs}\boldsymbol{\Sigma}^0_{obs,mis}=\mathbf{0}
\end{aligned}
$$

using the fact that $\mathbf{A}'_{i,mis}\boldsymbol{\Sigma}^0_{mis,mis}=-\mathbf{A}'_{i,obs}\boldsymbol{\Sigma}^0_{obs,mis}$ and $\mathbf{A}'_{i,mis}\boldsymbol{\Sigma}^0_{mis,obs}=-\mathbf{A}'_{i,obs}\boldsymbol{\Sigma}^0_{obs,obs}$. Therefore $\mathbf{A}'\mathrm{V}^0(\mathbf{X}_{i,mis})=\mathbf{0}$.
Next consider the case where we are at $t=1$. Impute for nonrespondent i once again, but now based on the new estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

$$
\mathrm{E}[\mathbf{X}_{i,mis}\mid\mathbf{X}_{i,obs},\boldsymbol{\mu}^1,\boldsymbol{\Sigma}^1]=\boldsymbol{\mu}^1_{mis}+\boldsymbol{\Sigma}^1_{mis,obs}(\boldsymbol{\Sigma}^1_{obs,obs})^{-1}(\mathbf{X}_{i,obs}-\boldsymbol{\mu}^1_{obs})
$$

Then

$$
\begin{aligned}
\mathbf{A}'_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis}\mid\mathbf{X}_{i,obs},\boldsymbol{\mu}^1,\boldsymbol{\Sigma}^1] &= \mathbf{A}'_{i,mis}\boldsymbol{\mu}^1_{mis}+\\
&\quad\ \mathbf{A}'_{i,mis}\boldsymbol{\Sigma}^1_{mis,obs}(\boldsymbol{\Sigma}^1_{obs,obs})^{-1}(\mathbf{X}_{i,obs}-\boldsymbol{\mu}^1_{obs})
\end{aligned}
$$

where the product $\mathbf{A}'_{i,mis}\boldsymbol{\Sigma}^1_{mis,obs}$ equals

$$\mathbf{A}'_{i,mis}\frac{1}{n}\sum_{h=1}^{n}\Big((\mathbf{X}_{h,mis}-\boldsymbol{\mu}^1_{mis})(\mathbf{X}_{h,obs}-\boldsymbol{\mu}^1_{obs})'+\mathrm{V}^0(\mathbf{X}_{h,mis})_{mis,obs}\Big) \qquad (3)$$

Again we will use the fact that $\mathbf{A}'_{i,mis}\mathbf{X}_{h,mis}=-\mathbf{A}'_{i,obs}\mathbf{X}_{h,obs}$ and $\mathbf{A}'_{i,mis}\boldsymbol{\mu}^1_{mis}=-\mathbf{A}'_{i,obs}\boldsymbol{\mu}^1_{obs}$. But now we will also use the property $\mathbf{A}'\mathrm{V}^0(\mathbf{X}_{i,mis})=\mathbf{0}$, that was given above, from which it can be derived that $\mathbf{A}'_{i,mis}\mathrm{V}^0(\mathbf{X}_{h,mis})_{mis,obs}=-\mathbf{A}'_{i,obs}\mathrm{V}^0(\mathbf{X}_{h,mis})_{obs,obs}$. This means that equation (3) becomes

$$-\mathbf{A}'_{i,obs}\frac{1}{n}\sum_{h=1}^{n}\Big((\mathbf{X}_{h,obs}-\boldsymbol{\mu}^1_{obs})(\mathbf{X}_{h,obs}-\boldsymbol{\mu}^1_{obs})'+\mathrm{V}^0(\mathbf{X}_{h,mis})_{obs,obs}\Big)$$

which equals $-\mathbf{A}'_{i,obs}\boldsymbol{\Sigma}^1_{obs,obs}$. So

$$\mathbf{A}'_{i,mis}\boldsymbol{\Sigma}^1_{mis,obs}=-\mathbf{A}'_{i,obs}\boldsymbol{\Sigma}^1_{obs,obs}$$

and therefore

$$\begin{aligned}\mathbf{A}'_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis}\mid\mathbf{X}_{i,obs},\boldsymbol{\mu}^1,\boldsymbol{\Sigma}^1] &= \mathbf{A}'_{i,mis}\boldsymbol{\mu}^1_{mis}-\mathbf{A}'_{i,obs}\mathbf{X}_{i,obs}+\mathbf{A}'_{i,obs}\boldsymbol{\mu}^1_{obs}\\ &= -\mathbf{A}'_{i,obs}\mathbf{X}_{i,obs}\end{aligned}$$

In a similar way it can be shown that this holds for the following iterations $t=2,3,\ldots$. So at each iteration the linear balance restriction will be satisfied.

2.3 Starting values

The EM algorithm requires a starting value for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for the first iteration. The mean vector and covariance matrix can, for example, be calculated from the completely observed data or by available cases. Using the complete cases provides consistent estimates of the parameters if the data are MCAR and if there are at least $k+1$ observations. In general the choice of the starting value is not crucial, unless the fraction of missing data is very high (Schafer (1997)). In our case, however, the choice of the starting value is very crucial in the sense that if the starting value does not satisfy the linear restriction, $\mathbf{A}'\boldsymbol{\mu}^0=\mathbf{0}$, the final estimates of the parameters will not either. This rules out the use of available cases estimates.

2.4 Conclusions

This imputation method will preserve the distribution of the data and may therefore be very useful. Another advantage of this method is that we do not have to specify the type of edit constraint (as long as it concerns a balance constraint). The information about the edit constraint is automatically incorporated in the singular covariance matrix. Unfortunately, this method cannot be used for inequality constraints as the data cannot be normally distributed

then. Another point of concern is the dependence on the normal distribution, which is not very realistic with regard to economic data. We therefore have developed another method that may work well on data that are not normally distributed.

3 Imputation under restrictions when the data are not normally distributed

3.1 The statistical distribution of economic data

The method described above will work well for data that are normally distributed. However, examining realistic data from Statistics Netherlands leads to the conclusion that the data are only rarely normally distributed and that they are mostly very skew. Consequently, we will try to fit another distribution on our data. Distributions for models are often chosen on the basis of the range of the random variable. For a variable constrained between zero and one the beta distribution has proved useful.
The beta distribution is defined by the probability density function (pdf)

$$f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},\ 0 < x < 1,\ \alpha, \beta > 0$$

where $\Gamma(\cdot)$ is the gamma function defined by $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$.
This functional form is extremely flexible in the shapes it will accommodate. It is symmetric if $\alpha = \beta$ and asymmetric otherwise. Besides it can be hump-shaped ($\alpha, \beta > 1$) or U-shaped ($\alpha, \beta < 1$). Note that it reduces to the uniform distribution if $\alpha = \beta = 1$. Also note that if $X \sim \text{beta}(\alpha, \beta)$ then $1 - X \sim \text{beta}(\beta, \alpha)$.
An extension of the beta distribution is the so-called Dirichlet distribution, also referred to as the multivariate beta. Its pdf is

$$f(x_1, \ldots, x_k \mid \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k x_j^{\alpha_j - 1} \tag{4}$$

where

$$x_j \geq 0,\ \alpha_j > 0,\ j = 1, \ldots, k,\ \sum_{j=1}^k x_j = 1$$

Note that because $\sum_{j=1}^k x_j = 1$, this is actually a $(k-1)$-dimensional distribution. One variable can be obtained from the other ones. Consequently, the pdf is sometimes written as

$$f(x_1, \ldots, x_{k-1} \mid \alpha_1, \ldots, \alpha_{k-1}; \alpha_k) = \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^{k-1} x_j^{\alpha_j - 1} \big(1 - \sum_{j=1}^{k-1} x_j\big)^{\alpha_k - 1} \tag{5}$$

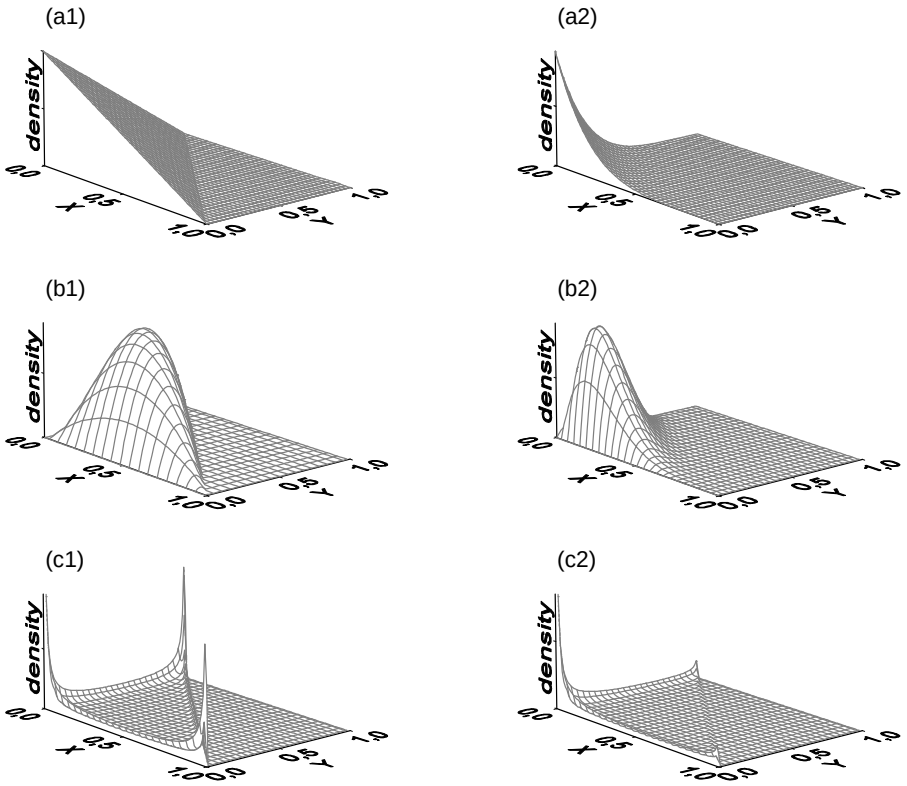

Figure 1. *Some bivariate Dirichlet density plots for (a1)* $\boldsymbol{\alpha} = (1,1,2)$*, (a2)* $\boldsymbol{\alpha} = (1,1,6)$*, (b1)* $\boldsymbol{\alpha} = (2,2,2)$*, (b2)* $\boldsymbol{\alpha} = (2,2,6)$*, (c1)* $\boldsymbol{\alpha} = (0.2,0.2,0.2)$ *and (c2)* $\boldsymbol{\alpha} = (0.2,0.2,0.6)$.

The Dirichlet is a convenient distribution on the simplex. The family of Dirichlet distributions is an exponential family and has finitely many complete sufficient statistics. We will refer to the Dirichlet distribution given by (4) with $\mathrm{Dir}_{k-1}(\alpha_1,\ldots,\alpha_k)$ and the Dirichlet distribution given by (5) with $\mathrm{Dir}_{k-1}(\alpha_1, \ldots,\alpha_{k-1};\alpha_k)$. Note that in the case of $k = 2$ the Dirichlet reduces to the beta distribution.

In Figure 1 some examples of Dirichlet densities ($k = 3$) are shown. Figure 1 (a1) shows the density for $\boldsymbol{\alpha} = (1,1,2)$ and (a2) shows the density for $\boldsymbol{\alpha} = (1,1,6)$. Two humpshaped densities are shown in Figure 1 (b1) and (b2) with the parameters $\boldsymbol{\alpha} = (2,2,2)$ and $\boldsymbol{\alpha} = (2,2,6)$ respectively. Finally in Figure 1 (c1) and (c2) two U-shaped densities are shown with parameters $\boldsymbol{\alpha} = (0.2,0.2,0.2)$ and $\boldsymbol{\alpha} = (0.2,0.2,0.6)$, respectively.

The first and second order moments of the Dirichlet distribution are

$$\begin{aligned} \mathrm{E}X_j &= \frac{\alpha_j}{\alpha}, & j = 1,\ldots,k \\ \mathrm{Var}X_j &= \frac{\alpha_j(\alpha-\alpha_j)}{\alpha^2(\alpha+1)}, & j = 1,\ldots,k \end{aligned}$$

where $\alpha = \sum_{j=1}^{k} \alpha_j$.

The covariances of the X's are

$$\mathrm{Cov}(X_j, X_l) = \frac{-\alpha_j \alpha_l}{\alpha^2(\alpha+1)} \quad j, l = 1, \ldots, k,\ j \neq l$$

Notice that if the means are held constant but α is allowed to increase, the variances and covariances decrease. For this reason α can be regarded as some sort of precision parameter: as α increases the distribution becomes more tightly concentrated about the mean.
The following theorems apply (for a derivation of these theorems see Wilks (1962)).

Theorem 1 (Marginal Dirichlet)
If $(X_1, \ldots, X_k)$ *is a random variable vector having the* $(k-1)$*-variate Dirichlet distribution* $\mathrm{Dir}_{k-1}(\alpha_1, \ldots, \alpha_k)$*, then the marginal distribution of* $(X_1, \ldots, X_{k_1})$*,* $k_1 < k$ *is the* (k_1-1)*-variate Dirichlet distribution* $\mathrm{Dir}_{k_1-1}(\alpha_1, \ldots, \alpha_{k_1-1}; \alpha_{k_1} + \cdots + \alpha_k)$.

Theorem 2 (Conditional Dirichlet)
If $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')' \sim \mathrm{Dir}_{k-1}(\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')$ *where* $\mathbf{X}_1$ *and* $\boldsymbol{\alpha}_1$ *consist of* r *elements and* $\mathbf{X}_2$ *and* $\boldsymbol{\alpha}_2$ *consist of* s *elements and* $k = r + s$*, then*

$$(1 - \mathbf{X}_2'\mathbf{1})^{-1}\mathbf{X}_1 \mid \mathbf{X}_2 \sim \mathrm{Dir}_{r-1}(\boldsymbol{\alpha}_1')$$

3.2 Imputation of multivariate missing item values

3.2.1 The edit constraints

Again we will first consider a balance edit constraint. We believe that we can impute the missing items immediately satisfying the edit constraint and preserving the distribution of the data by making use of the Dirichlet distribution. Consider edit rule (1) and transform it by dividing the different parts by the total X_k, this is done in order to restrict the domain of the variables $\tilde{X}_1, \ldots, \tilde{X}_{k-1}$ to the simplex; we take $\tilde{X}_j = c_j X_j / X_k,\ j = 1, \ldots, k-1$.

$$\begin{array}{ccccc} \frac{c_1 X_1}{X_k} & + \cdots + & \frac{c_{k-1} X_{k-1}}{X_k} & = 1, & X_k > 0 \\ \tilde{X}_1 & + \cdots + & \tilde{X}_{k-1} & = 1 & \end{array}$$

Note that we assume that the total, X_k, is known. This is done for two reasons. First of all since X_k is an aggregate the nonresponse rate will probably be low. And secondly if it is indeed missing we expect to be able to estimate this value very well based on the other variables in the survey, whereas the different subtotals are far more difficult to estimate this way.

3.2.2 Imputation

A special case arises when only one $\tilde{X}_j,\ j = 1, \ldots, k-1$ is missing. In this instance deductive imputation can be used. Deductive imputation means that the value of the missing item can be established with certainty based on the other items in the survey.
If all items of the data vector $\tilde{\mathbf{X}}$ are missing, we can obtain imputations by drawing from $\text{Dir}_{k-2}(\alpha_1, \ldots, \alpha_{k-1})$. However, as we mentioned earlier, a common circumstance is that a few item values are missing and the others are observed. In this case one needs to draw values from the conditional distribution of the missing items given the observed ones, which is also a Dirichlet distribution as was established in Theorem 2.
Partition $\tilde{\mathbf{X}}$ into $\tilde{\mathbf{X}}_{mis}$ and $\tilde{\mathbf{X}}_{obs}$, where $\tilde{\mathbf{X}}_{mis}$ represents the missing items and $\tilde{\mathbf{X}}_{obs}$ represents the observed items. The vector with missings, $\tilde{\mathbf{X}}_{mis}$, consists of m elements and $\tilde{\mathbf{X}}_{obs}$ consists of o elements, which is the number of observed items and $m + o = k - 1$. Partition $\boldsymbol{\alpha}$ accordingly. Then it holds that

$$(1 - \mathbf{1}'\tilde{\mathbf{X}}_{obs})^{-1}\tilde{\mathbf{X}}_{mis} \mid \tilde{\mathbf{X}}_{obs} \sim \text{Dir}_{m-1}(\boldsymbol{\alpha}_{mis}) \tag{6}$$

Thus imputations for missing items can be obtained by drawing from the conditional Dirichlet distribution mentioned in (6).

3.3 Random number generation

In order to impute missing values we need to generate random values from the Dirichlet distribution. This can be done as follows.
Recall that if $U_1 \sim \text{gamma}(\alpha, \lambda)$ and $U_2 \sim \text{gamma}(\beta, \lambda)$ then $Z = \frac{U_1}{U_1+U_2} \sim \text{beta}(\alpha, \beta)$. This can be generalised to the Dirichlet distribution, see for example Wilks (1962).
Suppose $U_1, \ldots, U_{k-1}$ are independent random variables having gamma distributions $\text{gamma}(\alpha_1, \lambda), \ldots, \text{gamma}(\alpha_{k-1}, \lambda)$. For $j = 1, \ldots, k-1$, let $Z_j = \frac{U_j}{U_1+\cdots+U_{k-1}}$, then $(Z_1, \ldots, Z_{k-1})$ has the $(k-2)$-variate Dirichlet distribution $\text{Dir}_{k-2}(\alpha_1, \ldots, \alpha_{k-1})$. Thus random values from the Dirichlet distribution can be obtained by drawing independently from gamma distributions.

3.4 Parameter estimation

3.4.1 The method of moments estimator

The parameters $\alpha_1, \ldots, \alpha_{k-1}$ can be estimated by a method of moments estimator. The method of moments estimator is consistent.
The first and second order moments of the Dirichlet distribution are

$$\mu_j = \frac{\alpha_j}{\alpha}, \qquad j = 1, \ldots, k-1 \tag{7}$$

$$\sigma_j^2 \quad = \quad \frac{\alpha_j(\alpha - \alpha_j)}{\alpha^2(\alpha+1)}, \quad j = 1, \ldots, k-1 \tag{8}$$

where $\alpha = \sum_{j=1}^{k-1} \alpha_j$. Rewrite (7) as $\alpha = \frac{\alpha_j}{\mu_j}$ and substitute this in equation (8). Then

$$\begin{aligned} \sigma_j^2 &= \frac{\alpha_j\big(\frac{\alpha_j}{\mu_j} - \alpha_j\big)}{\big(\frac{\alpha_j}{\mu_j}\big)^2\big(\frac{\alpha_j}{\mu_j}+1\big)} \\ \big(\frac{\alpha_j}{\mu_j}+1\big)\sigma_j^2 &= (1-\mu_j)\mu_j \end{aligned}$$

Solving for α_j gives

$$\alpha_j \quad = \quad \mu_j\big(\frac{\mu_j}{\sigma_j^2}(1-\mu_j) - 1\big), \quad j = 1, \ldots, k-1$$

So

$$\hat{\alpha}_{MM,j} \quad = \quad \hat{\mu}_j\big(\frac{\hat{\mu}_j}{\hat{\sigma}_j^2}(1-\hat{\mu}_j) - 1\big), \quad j = 1, \ldots, k-1$$

where $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$, $j = 1, \ldots, k-1$.
Although the method of moments is straightforward, estimation based on the method of moments generally is not statistically efficient. That is, the asymptotic variance-covariance matrix of the estimators is usually larger than the inverse of the information matrix. However, estimation based on the method of moments can serve as an excellent initial guess to start iterations in the Newton-Raphson algorithm, which we use to maximize the likelihood.

3.4.2 Maximum likelihood estimation

In order to find a consistent estimator, that is statistically efficient, maximum likelihood estimation can be applied. When sampling from a distribution that is a member from an exponential family of distributions, the maximum likelihood estimators will be a function of the sufficient statistics.
The likelihood is defined as follows

$$L(\boldsymbol{\alpha} \mid \mathbf{X}) = \prod_{i=1}^{n} f(\mathbf{X}_i, \boldsymbol{\alpha})$$

Since the joint density of $X_1, \ldots, X_{k-1}$ is the Dirichlet density, the likelihood will be

$$L(\boldsymbol{\alpha} \mid \mathbf{X}) = \frac{\Gamma^n(\sum_{j=1}^{k-1}\alpha_j)}{\prod_{j=1}^{k-1}\Gamma^n(\alpha_j)} \prod_{i=1}^{n}\prod_{j=1}^{k-1} X_{ij}^{\alpha_j - 1}$$

Taking the natural logarithm leads to the following loglikelihood function

$$\ell(\boldsymbol{\alpha} \mid \mathbf{X}) = \ln L(\boldsymbol{\alpha} \mid \mathbf{X}) \quad = \quad n\ln\Gamma\big(\sum_{j=1}^{k-1}\alpha_j\big) - n\sum_{j=1}^{k-1}\ln\Gamma(\alpha_j) +$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k-1}(\alpha_j - 1)\ln X_{ij}$$

Taking the first derivative results in

$$\begin{aligned}\ell'_p(\boldsymbol{\alpha} \mid \mathbf{X}) = \frac{\partial \ell(\boldsymbol{\alpha} \mid \mathbf{X})}{\partial \alpha_p} &= n\frac{\partial \ln\Gamma(\sum_{j=1}^{k-1}\alpha_j)}{\partial \alpha_p} - n\frac{\partial \ln\Gamma(\alpha_p)}{\partial \alpha_p} + \\ &\sum_{i=1}^{n}\ln X_{ip}, \qquad p = 1, \ldots, k-1\end{aligned}$$

This leads to

$$\ell'_p(\boldsymbol{\alpha} \mid \mathbf{X}) = n\Psi(\sum_{j=1}^{k-1}\alpha_j) - n\Psi(\alpha_p) + \sum_{i=1}^{n}\ln X_{ip},\ p = 1, \ldots, k-1 \tag{9}$$

where $\Psi(\cdot)$ is the digamma function defined by $\Psi(x) = \frac{\partial \ln\Gamma(x)}{\partial x} = \frac{\Gamma'(x)}{\Gamma(x)}$.
We need some iterative scheme to solve the equation $\ell'_p(\boldsymbol{\alpha} \mid \mathbf{X}) = 0,\ p = 1, \ldots, k-1$. A commonly used method is the Newton-Raphson method. Determine an initial value for $\boldsymbol{\alpha}$, for example by means of the method of moments estimator. To find $\hat{\boldsymbol{\alpha}}_{MLE}$ that solves $\ell'(\boldsymbol{\alpha} \mid \mathbf{X}) = \mathbf{0}$, calculate by iteration and until convergence

$$\boldsymbol{\alpha}^{new} = \boldsymbol{\alpha}^{old} - \mathbf{H}^{-1}\ell'$$

where $\mathbf{H}$ is the Hessian, the matrix of second-derivatives of ℓ given by

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\alpha} \mid \mathbf{X})}{\partial \alpha_p^2} &= n\Psi'(\sum_{j=1}^{k-1}\alpha_j) - n\Psi'(\alpha_p), \qquad p = 1, \ldots, k-1 \\ \frac{\partial^2 \ell(\boldsymbol{\alpha} \mid \mathbf{X})}{\partial \alpha_p \partial \alpha_q} &= n\Psi'(\sum_{j=1}^{k-1}\alpha_j), \qquad p, q = 1, \ldots, k-1,\ p \neq q\end{aligned}$$

where Ψ' is known as the trigamma function.
The Hessian is therefore

$$\begin{aligned}\mathbf{H} &= -n[\mathrm{diag}\{\Psi'(\alpha_1), \ldots, \Psi'(\alpha_{k-1})\} - \Psi'(\sum_{j=1}^{k-1}\alpha_j)\boldsymbol{\iota}_{k-1}\boldsymbol{\iota}'_{k-1}] \\ &= -n[\mathbf{D} - c\boldsymbol{\iota}_{k-1}\boldsymbol{\iota}'_{k-1}]\end{aligned}$$

where $\boldsymbol{\iota}_{k-1}$ denotes a vector of ones of length $k-1$. The inverse of $\mathbf{H}$ can be easily calculated using a well-known matrix inversion lemma, also referred to as the Sherman-Morrison formula. So

$$\mathbf{H}^{-1} = -\frac{1}{n}\Big[\mathbf{D}^{-1} + \frac{c\mathbf{D}^{-1}\boldsymbol{\iota}_{k-1}\boldsymbol{\iota}'_{k-1}\mathbf{D}^{-1}}{1 - c\boldsymbol{\iota}'_{k-1}\mathbf{D}^{-1}\boldsymbol{\iota}_{k-1}}\Big]$$

Under some regularity conditions the likelihood function is strictly concave for exponential families, and the MLE exists and is unique. Since the Dirichlet distribution is an exponential family this holds true for the Dirichlet distribution.

A direct proof has been given by Ronning (1989).
However, when we are faced with missing item values $\mathbf{X}$ is not completely observed and the loglikelihood cannot be calculated directly. In order to estimate the maximum likelihood estimates when some of the variables involved are not observed the EM algorithm has been developed.

3.4.3 The Expectation-Maximization algorithm

As we already discussed in section 2.1 the EM algorithm calculates maximum likelihood estimates in the presence of missing data.
In order to apply the EM algorithm when the data are Dirichlet distributed the expected sufficient statistics need to be calculated, since the Dirichlet distribution is an exponential family.
These values can be easily computed from the natural parameterisation of the exponential family representation of the Dirichlet distribution. The density function of the Dirichlet for $\mathbf{X} = (X_1, \ldots, X_{k-1})'$ can be written in this form as follows

$$p(\mathbf{X} \mid \boldsymbol{\alpha}) = \exp\{\ln\Gamma(\sum_{j=1}^{k-1}\alpha_j) - \sum_{j=1}^{k-1}\ln\Gamma(\alpha_j) + \sum_{j=1}^{k-1}(\alpha_j - 1)\ln X_j\}$$

The natural parameter of the Dirichlet is therefore $\theta_j = \alpha_j - 1$ and the sufficient statistic is $T(X_j) = \ln X_j$, $j = 1, \ldots, k-1$. Using the general fact that the derivative of the natural logarithm of the normalization factor with respect to the natural parameter is equal to the expectation of the sufficient statistic, we obtain

$$\mathrm{E}[\ln X_j \mid \boldsymbol{\alpha}] = \Psi(\alpha_j) - \Psi(\sum_{j=1}^{k-1}\alpha_j), \quad j = 1, \ldots, k-1$$

The expectation of the sufficient statistic conditional on the observed values can also be easily calculated since the conditional distribution of Dirichlet distributed variables is also a Dirichlet (see Theorem 2).
Consider $\mathbf{X} = (X_1, \ldots, X_{k-1})'$ which is $(k-2)$-variate Dirichlet distributed with parameters $\alpha_1, \ldots, \alpha_{k-1}$. For each record i, $i = 1, \ldots, n$, the missing X's are present in the vector $\mathbf{X}_{i,mis} = (X_1, \ldots, X_{m_i})'$ and the observed X's are represented by $\mathbf{X}_{i,obs} = (X_{m_i+1}, \ldots, X_{k-1})'$. The number of missing values for each record is thus denoted by m_i. Then

$$\hat{\mathbf{X}}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\alpha} \sim \mathrm{Dir}_{m_i-1}(\alpha_1, \ldots, \alpha_{m_i})$$

where

$$\hat{\mathbf{X}}_{i,mis} = (1 - \mathbf{1}'\mathbf{X}_{i,obs})^{-1}\mathbf{X}_{i,mis}$$

It follows that for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$

$$\mathrm{E}[\ln \hat{X}_{ij} \mid \mathbf{X}_{i,obs}, \boldsymbol{\alpha}] \quad = \quad \Psi(\alpha_j) - \Psi(\sum_{p=1}^{m_i}\alpha_p)$$

$$\begin{aligned} \mathrm{E}[\ln \frac{X_{ij}}{1-\mathbf{1}'\mathbf{X}_{i,obs}} \mid \mathbf{X}_{i,obs}, \boldsymbol{\alpha}] &= \Psi(\alpha_j) - \Psi(\sum_{p=1}^{m_i} \alpha_p) \\ \mathrm{E}[\ln X_{ij} \mid \mathbf{X}_{i,obs}, \boldsymbol{\alpha}] &= \ln(1-\mathbf{1}'\mathbf{X}_{i,obs}) + \Psi(\alpha_j) - \Psi(\sum_{p=1}^{m_i} \alpha_p) \end{aligned}$$

Plug this value into equation (9) and re-estimate the parameters $\boldsymbol{\alpha}$. Repeat this procedure until the estimates converge.

3.5 Linear inequality edit restrictions

This procedure can be straightforwardly extended to handle an inequality edit restriction as mentioned in (2). Transform this edit similarly by dividing by the aggregate X_k. This results in

$$\tilde{X}_1 + \cdots + \tilde{X}_{k-1} \leq 1$$

We will now use the Dirichlet distribution as described in equation (5) of dimension $k-1$. This means that a variable equal to $1 - \sum_{j=1}^{k-1} \tilde{X}_j$ needs to be added in order to reduce this inequality to an equality restriction. So in order to impute the missing items we need to draw from the $(k-1)$-variate Dirichlet distribution $\mathrm{Dir}(\alpha_1, \ldots, \alpha_{k-1}; \alpha_k)$.
Note that deductive imputation cannot be used in the case of one missing $\tilde{X}_j$, since we do not now the exact sum of the subtotals but only the maximum value.
Furthermore all the methods described in the previous sections can be applied to impute these items.

4 Discussion

In the techniques described above for both normally and Dirichlet distributed data we have only considered one balancing edit constraint, whereas in real life there are of course several. This also means that variables can be present in more than one constraint, which seriously complicates the imputation process, since the imputed values need to satisfy all constraints at once.
Moreover, the inequality edit constraints can also be of a less useful form than the one mentioned in (2). For example, when they consist of just two variables, where one variable is the upper bound of the other.
The need for an imputation method that simultaneously imputes all missing items immediately satisfying all linear constraints, balance as well as inequality restrictions, while preserving the distribution of the data arises. Future research will be focussed on this subject.

References

[1] Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis*, Second Edition, New York: Wiley Series.

[2] Dempster, A.P., Laird, N.M. and D.B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion)," *Journal of the Royal Statistical Society B*, 39, 1-38.

[3] Kalton, G. and D. Kasprzyk (1986), "The Treatment of Missing Survey Data," *Survey Methodology*, 12, 1-16.

[4] Little, R.J.A. and Rubin, D.B. (2002), *Statistical Analysis with Missing Data*, Second Edition, New York: Wiley Series.

[5] Ronning, G. (1989), "Maximum Likelihood Estimation of Dirichlet Distributions," *Journal of Statistical Computation and Simulation*, 32, 215-221.

[6] Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley Series.

[7] Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

[8] Wilks, S.S. (1962), *Mathematical Statistics*, New York: Wiley Series.

Contributors

Authors

Caren Tempelman
Department of Methods and Informatics, Statistics Netherlands
PO Box 4000
2270 JM Voorburg
The Netherlands
e-mail: DTMN@CBS.nl

Ton Steerneman
Department of Econometrics, University of Groningen
PO Box 800
9700 AV Groningen
The Netherlands
e-mail: A.G.M.Steerneman@eco.rug.nl