Chapter 3

Imputation of Data Subject to One Balance Restriction

Common well-known imputation methods provide imputations that are likely not to satisfy the linear balance restrictions on economic data. Adjusting these values in order to satisfy the restrictions will lead to a distortion of the distribution of the imputed data and consequently to a distortion of the distribution of the completed data. Besides, valuable information provided by the balance restrictions is not used. In this chapter we suggest an imputation procedure that generates imputations for non-negative data items subject to one linear balance restriction. Our suggestion is to use the Dirichlet distribution to model the data.

3.1 Introduction

In chapter 1 an elaborate overview was given of frequently used imputation methods. We established that imputation methods can be either deterministic or stochastic. Deterministic methods determine imputed values uniquely, which 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. We prefer to have
an imputation method that can be both deterministic as well as stochastic so
that they are suitable for both situations.

A property of economic data is that there are many logical constraints on the
data items, such as the fact that total operating expenses must be the added
total of all operating expenses reported, such as housing costs, depreciation,
personnel costs and so on. Commonly used imputation methods such as hot
deck and (random) regression imputation mostly do not provide imputations
that satisfy these linear balance restrictions. This means that imputations need
to be adjusted in order to satisfy the balance restrictions on the data. Although
this can be done rather straightforwardly using linear optimisation techniques
such as the simplex algorithm, the adjustment of imputed values inadvertently
leads to a distortion of the distribution of the imputed values and therefore to a
distortion of the distribution of the final, completed dataset. Besides, these bal-
ance restrictions provide the imputer with valuable information on what values
(not) to impute and therefore it would be desirable to incorporate the restric-
tions in the imputation process.

In this chapter we suggest the use of an imputation scheme to obtain im-
putations for data items that are subject to one linear balance restriction and
that need to be non-negative. In section 3.2 the edit constraints are described
and in section 3.3 the Dirichlet distribution and some of its properties are dis-
cussed. Subsequently parameter estimation and the Expectation-Maximisation
(EM) algorithm that is used to obtain the maximum likelihood estimates in
the presence of nonresponse are treated in sections 3.4 and 3.5. In section 3.6
imputation is discussed and in section 3.7 some results are presented on the
performance of this imputation method, compared to other missing data pro-
cedures, on empirical data. Finally, in section 3.8 some concluding remarks will
be made.

3.2 The edit constraint

Consider a balance edit restriction of the following shape

$$c'X = X_k,$$

where $X$ is a vector of order $(k - 1) \times 1$ which contains solely non-negative
elements and the $c_j$’s, $j = 1, \ldots, k - 1$, are known, positive constants. This
restriction, for example, could refer to different types of operating expenses
$(X_1, \ldots, X_{k-1})$ that need to add up to the total operating expenses ($X_k$), where
3.3. A statistical distribution of economic data

\( c_j = 1, \text{ for } j = 1, \ldots, k - 1. \) This restriction can be transformed by dividing the different parts by the total \( X_k \) in order to restrict the domain of the resulting variables \( Y_1, \ldots, Y_{k-1} \) to the simplex:

\[
\frac{c_1 X_1}{X_k} + \cdots + \frac{c_{k-1} X_{k-1}}{X_k} = 1, \quad X_k > 0 \\
Y_1 + \cdots + Y_{k-1} = 1.
\] (3.1)

Note that we assume that the total, \( X_k \), is known. This is plausible 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 quite accurately based on the other variables in the survey, whereas subtotals are far more difficult to estimate this way. For notational convenience define \( l = k - 1 \), then the data that need to be modelled and imputed are the ratios \( Y_1, \ldots, Y_l \).

A well-known and popular imputation scheme is the hot deck method. In the current situation the traditional hot deck method needs to be slightly altered in order to make sure the imputed items satisfy the linear balance restriction. In the adjusted hot deck method not the actual responses of a donor but their ratios with respect to the total \( X_k \) are used as imputations. We refer to this method as the ratio hot deck method, see e.g. Pannekoek and de Waal (2005). A disadvantage of this method is the fact that we need to find a suitable donor, which can be complicated for economic data, especially if the nonresponse rates are substantial. One can either select a donor randomly or use some sort of distance function to find the donor that is most similar to the record with missing data items. The latter procedure will probably lead to better results but is also more difficult as one needs to establish what variables should be present in the distance function, i.e. on what variables should the donor search be based. Furthermore, these variables need to be standardised in order to have equal weights. Finally, for large datasets this method can become quite slow as distances need to be calculated between the record with missing items and all available donors.

The method we suggest based on the Dirichlet distribution offers a flexible, model-based, alternative to ratio hot deck methods.

### 3.3 A statistical distribution of economic data

Models for distributions 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
\[
f(y | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad 0 < y < 1, \quad \alpha, \beta > 0,
\]
where \(\Gamma(\cdot)\) is the gamma function defined by \(\Gamma(y) = \int_0^\infty u^{y-1}e^{-u}du\).

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 \(Y \sim \text{beta}(\alpha, \beta)\) then \(1 - Y \sim \text{beta}(\beta, \alpha)\).

An extension of the beta distribution is the so-called Dirichlet distribution (see Johnson and Kotz, 1972), also referred to as the multivariate beta. Its pdf is
\[
f(y_1, \ldots, y_l \mid \alpha_1, \ldots, \alpha_l) = \frac{\Gamma(\sum_{j=1}^l \alpha_j)}{\prod_{j=1}^l \Gamma(\alpha_j)} \prod_{j=1}^l y_j^{\alpha_j-1}, \quad (3.2)
\]
where
\[
y_j \geq 0, \quad \alpha_j > 0, \quad j = 1, \ldots, l, \quad \sum_{j=1}^l y_j = 1.
\]
Note that since \(\sum_{j=1}^l y_j = 1\), this is actually a \((l-1)\)-dimensional distribution; one variable can be obtained with certainty from the others. Consequently, the pdf is sometimes written as
\[
f(y_1, \ldots, y_{l-1} \mid \alpha_1, \ldots, \alpha_{l-1}; \alpha_l) = \frac{\Gamma(\sum_{j=1}^{l-1} \alpha_j)}{\prod_{j=1}^{l-1} \Gamma(\alpha_j)} \prod_{j=1}^{l-1} y_j^{\alpha_j-1}(1 - \sum_{j=1}^{l-1} y_j)^{\alpha_l-1}. \quad (3.3)
\]
We will refer to the Dirichlet distribution given by (3.2) with \(\text{Dir}_{l-1}(\alpha_1, \ldots, \alpha_l)\) and the Dirichlet distribution given by (3.3) with \(\text{Dir}_{l-1}(\alpha_1, \ldots, \alpha_{l-1}; \alpha_l)\). Note that in the case of \(l = 2\) the Dirichlet reduces to the beta distribution. The Dirichlet is a convenient distribution on the simplex, as the family of Dirichlet distributions is an exponential family and has complete sufficient statistics of the form \(\ln Y_j, \quad j = 1, \ldots, l\).

The restriction that the data of interest lie in the simplex appears in several disciplines, such as geology, medicine and biology as well as in economics. This type of data is often referred to as compositional data. For a detailed description of applications and statistical analysis methods for compositional data, see
Aitchison (1986). Aitchison and Shen (1980) introduce the logistic normal distribution as a framework for the analysis of compositional data as an alternative to the Dirichlet distribution. This technique assumes multivariate normality of additive log-ratio transformed data, which means that the inference tools for multivariate normal data can be applied to the transformed compositions. The additive log-ratio transformation of a vector $Y$ in the $(l-1)$-dimensional simplex to $\mathbb{R}^{l-1}$ is

$$W_j = \ln \left( \frac{Y_j}{Y_l} \right), \quad j = 1, \ldots, l - 1.$$ 

The (unrestricted) vector $W$ is then modelled with a $(l - 1)$-dimensional multivariate normal distribution. However, if we are confronted with missing data, the fact that $W$ is unrestricted will most likely result in imputations that do not satisfy the composition after transforming $W$ back to $Y$: $Y_j = \exp(W_j)Y_l$. Furthermore, although Aitchison (1986) shows that all statistical procedures are invariant to the choice of component used as the denominator, this poses difficulties in an imputation context as this component needs to be observed for all records, which is highly unlikely. Additionally, this method is not readily applicable if some of the components are zero. Because of this and the fact that draws from the Dirichlet yield compositions immediately, the Dirichlet distribution seems more appropriate for the imputation of compositional data. For different applications of the Dirichlet distribution to compositional data see, for example, Jeuland et al. (1980), DeSarbo et al. (1993) and Haas and Formery (2002).

In Figure 3.1 the flexibility of the Dirichlet distribution is illustrated by some examples of Dirichlet density plots for the parameterisation given in (3.3) and $l = 3$. Figure 1 (a1) shows the density for $\alpha = (1, 1; 2)$ and (a2) shows the density for $\alpha = (1, 1; 6)$. Two lumpshaped densities are shown in Figure 1 (b1) and (b2) with the parameters $\alpha = (2, 2; 2)$ and $\alpha = (2, 2; 6)$ respectively. Finally in Figure 1 (c1) and (c2) two U-shaped densities are shown with parameters $\alpha = (0.2, 0.2; 0.2)$ and $\alpha = (0.2, 0.2; 0.6)$, respectively.

The first and second order moments of the Dirichlet distribution are

$$E[Y_j] = \frac{\alpha_j}{\alpha}, \quad j = 1, \ldots, l$$

$$\text{Var}(Y_j) = \frac{\alpha_j(\alpha - \alpha_j)}{\alpha^2(\alpha + 1)}, \quad j = 1, \ldots, l.$$
Figure 3.1: Some bivariate Dirichlet density plots for (a1) $\alpha = (1, 1; 2)$, (a2) $\alpha = (1, 1; 6)$, (b1) $\alpha = (2, 2; 2)$, (b2) $\alpha = (2, 2; 6)$, (c1) $\alpha = (0.2, 0.2; 0.2)$ and (c2) $\alpha = (0.2, 0.2; 0.6)$.

where $\alpha = \sum_{j=1}^{l} \alpha_j$ and the covariances between $Y_j$ and $Y_h$ are

$$\text{Cov}(Y_j, Y_h) = -\frac{\alpha_j \alpha_h}{\alpha^2(\alpha + 1)}, \quad j, h = 1, \ldots, l, \ j \neq h.$$ 

Notice that if the means are held constant but $\alpha$ is allowed to increase, the variances and covariances decrease. For this reason $\alpha$ can be regarded as some sort of precision parameter; as $\alpha$ increases the distribution becomes more tightly concentrated about the mean.

The following theorems apply (see Wilks, 1962, for a derivation of these theorems).
Theorem 1 (Marginal Dirichlet)
If $\mathbf{Y} = (Y_1, \ldots, Y_l)$ is a random variable vector having the $(l-1)$-variate Dirichlet distribution $\text{Dir}_{l-1}(\alpha_1, \ldots, \alpha_{l-1}; \alpha_l)$, then the marginal distribution of $Y_1 = (Y_1, \ldots, Y_{l_1})$, with $l_1 < l$ is the $(l_1-1)$-variate Dirichlet distribution $\text{Dir}_{l_1-1}(\alpha_1, \ldots, \alpha_{l_1-1}; \alpha_{l_1} + \cdots + \alpha_l)$.

Theorem 2 (Conditional Dirichlet)
If $\mathbf{Y} = (Y_1', Y_2')' \sim \text{Dir}_{l-1}(\alpha'_1, \alpha'_2)$ where $Y_1'$ and $\alpha_1$ consist of $l_1$ elements and $Y_2'$ and $\alpha_2$ consist of $l_2$ elements and $l = l_1 + l_2$, then

$$Y_1^* | Y_2 \sim \text{Dir}_{l_1-1}(\alpha_1'),$$

where $Y_1^* = (1 - Y_2' (1_{l_2}))^{-1} Y_1'$, with $1_{l_2}$ a vector of ones of length $l_2$. In this way the vector $Y_1$ is rescaled such that the elements of $Y_1^*$ sum up to one.

This second theorem is useful in the context of imputation as the missing data, that need to be imputed, conditional on the observed data still follow a Dirichlet distribution.

### 3.4 Parameter estimation

#### 3.4.1 The method of moments estimator

The parameters $\alpha_1, \ldots, \alpha_l$ of the Dirichlet distribution can be estimated by a method of moments (MM) estimator, which is consistent. Recall that the first and second order moments of the Dirichlet distribution are

$$\mu_j = \frac{\alpha_j}{\alpha}, \quad j = 1, \ldots, l \quad (3.4)$$

$$\sigma^2_j = \frac{\alpha_j (\alpha - \alpha_j)}{\alpha^2 (\alpha + 1)}, \quad j = 1, \ldots, l, \quad (3.5)$$

where $\alpha = \sum_{j=1}^{l} \alpha_j$. Rewrite (3.4) as $\alpha = \frac{\alpha_j}{\mu_j}$ and substitute this in equation (3.5). Then

$$\sigma^2_j = \frac{\alpha_j (\frac{\alpha_j}{\mu_j} - \alpha_j)}{\left(\frac{\alpha_j}{\mu_j}\right)^2 (\frac{\alpha_j}{\mu_j} + 1)}$$

$$(\frac{\alpha_j}{\mu_j} + 1) \sigma^2_j = (1 - \mu_j) \mu_j.$$
Solving for $\alpha_j$ gives the moments estimator
\[ \hat{\alpha}_{MM,j} = \hat{\mu}_j \left( \frac{\hat{\mu}_j}{\sigma_j^2} \right) (1 - \hat{\mu}_j) - 1, \quad j = 1, \ldots, l, \]
where $\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} Y_{ij}$, $j = 1, \ldots, l$.

Although the method of moments is straightforward, estimation based on the method of moments generally is not statistically efficient. That is, the asymptotic 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 maximise the likelihood function.

### 3.4.2 Maximum likelihood estimation

In order to find a consistent estimator for $\alpha$, that is statistically efficient, maximum likelihood estimation can be applied, see e.g. Narayanan (1991) and Ronning (1989). The likelihood function for the $(l-1)$-variate Dirichlet distribution is
\[ L(\alpha \mid y) = \frac{\Gamma^n \left( \sum_{j=1}^{l} \alpha_j \right) \prod_{j=1}^{l} \prod_{i=1}^{n} y_{ij}^{\alpha_j - 1}}{\prod_{j=1}^{l} \Gamma(\alpha_j)}. \]

Taking the natural logarithm leads to the following loglikelihood function
\[ \ell(\alpha \mid y) = n \ln \Gamma \left( \sum_{j=1}^{l} \alpha_j \right) - n \sum_{j=1}^{l} \ln \Gamma(\alpha_j) + \sum_{i=1}^{n} \sum_{j=1}^{l} (\alpha_j - 1) \ln y_{ij}. \]

In order to obtain the maximum likelihood estimates we need to set the first order derivatives of the (log)likelihood equal to zero. These first order derivatives, for $j = 1, \ldots, l$, are
\[ S_j(\alpha) = \frac{\partial \ell(\alpha \mid y)}{\partial \alpha_j} = n \Psi \left( \sum_{h=1}^{l} \alpha_h \right) - n \Psi(\alpha_j) + \sum_{i=1}^{n} \ln y_{ij}, \quad (3.6) \]
where $\Psi(\cdot)$ is the digamma function defined by $\Psi(y) = \frac{\partial \ln \Gamma(y)}{\partial y} = \frac{\Gamma'(y)}{\Gamma(y)}$.

As these equations cannot be solved analytically, we need some iterative scheme to find the optimal parameter values. Commonly used methods are the Newton-Raphson method and the Fisher scoring technique. In case of the Dirichlet distribution the negative of the Hessian matrix of second derivatives is
equal to the expected information matrix, which we will show momentarily, and therefore the Newton-Raphson and Fisher scoring algorithms are the same, thus

$$\alpha^{(t+1)} = \alpha^{(t)} + I^{-1}(\alpha)|_{\alpha=\alpha^{(t)}} S(\alpha)|_{\alpha=\alpha^{(t)}}, \quad t = 1, 2, \ldots,$$

where $I(\alpha)$ is the observed (or expected) information matrix and $S(\alpha)$ is the score vector, which contains the first order derivatives of the loglikelihood with respect to $\alpha$. The Hessian, $H(\alpha)$, is the matrix of second order derivatives of $\ell(\alpha \mid y)$, with elements

$$\frac{\partial^2 \ell(\alpha \mid y)}{\partial \alpha_j^2} = n \Psi'\left(\sum_{h=1}^l \alpha_h\right) - n \Psi'(\alpha_j), \quad j = 1, \ldots, l$$

$$\frac{\partial^2 \ell(\alpha \mid y)}{\partial \alpha_j \partial \alpha_p} = n \Psi'\left(\sum_{h=1}^l \alpha_h\right), \quad j, p = 1, \ldots, l, \quad j \neq p,$$

where $\Psi'(\cdot)$ is known as the trigamma function. The observed information matrix is therefore

$$I(\alpha) = -H(\alpha) = n[\text{diag}\{\Psi'(\alpha_1), \ldots, \Psi'(\alpha_l)\} - \Psi'(\sum_{j=1}^l \alpha_j) \ell \ell']$$

$$= n[D - \alpha \ell \ell'],$$

where $\ell$ denotes a vector of ones of length $l$. Clearly the observed information is independent of the random variable $Y$, meaning that $E[I(\alpha)] = I(\alpha)$, so in this case the expected information equals the observed information matrix.

The inverse of $I(\alpha)$ can be easily calculated using a well-known matrix inversion lemma, also referred to as the Sherman-Morrison formula (see Sherman and Morrison, 1950). Then

$$I^{-1}(\alpha) = \frac{1}{n} \left[D^{-1} + \frac{eD^{-1} \ell \ell' D^{-1}}{1 - \alpha \ell' D^{-1} \ell}\right].$$

Under some regularity conditions the likelihood function for exponential families is strictly concave and the maximum likelihood estimate exists and is unique. Since the Dirichlet distribution belongs to an exponential family and these regularity conditions are met, this also holds true for the Dirichlet distribution. A direct proof has been given by Ronning (1989).

Unfortunately, when we encounter missing item values, $Y$ is not completely
observed and the loglikelihood cannot be maximised directly. In order to obtain the maximum likelihood estimates in the presence of nonresponse the EM (Expectation Maximisation) algorithm was developed by Dempster, Laird and Rubin (1977).

3.5 The EM algorithm

As the Dirichlet distribution is a member of an exponential family, the EM algorithm consists of iteratively calculating the expected sufficient statistics and using these quantities to calculate the maximum likelihood estimates until convergence, as was shown in chapter 2.

The expected sufficient statistics can be easily computed from the natural parameterisation of the exponential family representation of the Dirichlet distribution. The density function of the Dirichlet for \( \mathbf{Y} = (Y_1, \ldots, Y_l)' \) can be written in this form as follows

\[
f(\mathbf{y} \mid \alpha) = \exp\left\{ \sum_{j=1}^{l} \ln \Gamma(\alpha_j) - \sum_{j=1}^{l} \ln \Gamma(\alpha_j) + \sum_{j=1}^{l} (\alpha_j - 1) \ln y_j \right\}.
\]

The natural parameter of the Dirichlet is \( \eta_j = \alpha_j - 1 \) and the sufficient statistic is \( t_j(\mathbf{y}) = \ln y_j, \; j = 1, \ldots, l \). In section 2.3.2 in chapter 2 we established that the negative of the derivative of the function \( g(\mathbf{\eta}) \) with respect to the natural parameter \( \mathbf{\eta} \) is equal to the expectation of the sufficient statistic, i.e.

\[
E[t_j(\mathbf{Y})] = -\frac{\partial g(\mathbf{\eta})}{\partial \eta_j}, \quad j = 1, \ldots, l.
\]

In case of the Dirichlet distribution the function \( g(\mathbf{\eta}) \) is

\[
g(\mathbf{\eta}) = \ln \Gamma(\sum_{j=1}^{l} \eta_j + l) - \sum_{j=1}^{l} \ln (\eta_j + 1).
\]

So, for \( j = 1, \ldots, l \),

\[
E[\ln Y_j \mid \mathbf{\alpha}] = \Psi(\eta_j + 1) - \Psi(\sum_{h=1}^{l} \eta_h + l) = \Psi(\alpha_j) - \Psi(\sum_{h=1}^{l} \alpha_h).
\]

The expectation of the sufficient statistic of the missing data items conditional on the observed values can be easily calculated as the conditional distribution of Dirichlet distributed variables is also a Dirichlet (see Theorem 2).
Consider a data vector $\mathbf{Y}$ that is distributed according to a Dirichlet distribution: $\mathbf{Y} \sim \text{Dir}_i(\alpha_1, \ldots, \alpha_i)$. For each record $i$ partition $\mathbf{Y}_i$ into a missing and an observed part, $\mathbf{Y}_i' = (\mathbf{Y}'_{i, \text{mis}}, \mathbf{Y}'_{i, \text{obs}})$. Let $m_i$ denote the number of missing items in record $i$. Then
\[
\mathbf{Y}'_{i, \text{mis}} | \mathbf{Y}_{i, \text{obs}} = \mathbf{y}_{i, \text{obs}}, \boldsymbol{\alpha} \sim \text{Dir}_{m_i - 1}(\alpha_1, \ldots, \alpha_{m_i}),
\]
where
\[
\mathbf{Y}'_{i, \text{mis}} = (1 - \mathbf{y}_{i, \text{obs}}' \mathbf{t}_{l - m_i})^{-1} \mathbf{Y}_{i, \text{mis}}.
\]

The expectation step of the EM algorithm now consists of calculating the expected sufficient statistics for $i = 1, \ldots, n$ and $j = 1, \ldots, l$, which are
\[
\hat{E}[\ln Y_{ij} | \mathbf{Y}_{i, \text{obs}} = \mathbf{y}_{i, \text{obs}}, \boldsymbol{\alpha}] = \begin{cases} 
\ln y_{ij} & \text{if } Y_{ij} \text{ is observed} \\
E[\ln Y_{ij} | \mathbf{y}_{i, \text{obs}}, \boldsymbol{\alpha}] & \text{if } Y_{ij} \text{ is missing}
\end{cases},
\]
\[\tag{3.7}
\]
We know that
\[
E[\ln Y_{ij}^* | \mathbf{Y}_{i, \text{obs}} = \mathbf{y}_{i, \text{obs}}, \boldsymbol{\alpha}] = \Psi(\alpha_j) - \Psi(\sum_{h=1}^{m_i} \alpha_h).
\]

From this it follows that
\[
E[\ln (1 - \mathbf{y}_{i, \text{obs}}' \mathbf{t}_{l - m_i})^{-1} Y_{ij} | \mathbf{Y}_{i, \text{obs}} = \mathbf{y}_{i, \text{obs}}, \boldsymbol{\alpha}] = \Psi(\alpha_j) - \Psi(\sum_{h=1}^{m_i} \alpha_h)
\]
and therefore
\[
E[\ln Y_{ij} | \mathbf{Y}_{i, \text{obs}} = \mathbf{y}_{i, \text{obs}}, \boldsymbol{\alpha}] = \ln (1 - \mathbf{y}_{i, \text{obs}}' \mathbf{t}_{l - m_i}) + \Psi(\alpha_j) - \Psi(\sum_{h=1}^{m_i} \alpha_h).
\]

Now calculate $\hat{E}[\ln Y_{ij} | \mathbf{Y}_{i, \text{obs}} = \mathbf{y}_{i, \text{obs}}, \boldsymbol{\alpha}]$ given in equation (3.7) and plug these values into equation (3.6). The maximisation step then consists of finding the parameter value for $\boldsymbol{\alpha}$ for which the expected loglikelihood is maximal. Next these updated parameter values are used to re-estimate the expectation and so on. This process is iterated until the estimates converge.

Several criteria can be used to assess whether the algorithm has converged. For instance, when the changes in the parameter estimates are sufficiently small: $\max |\alpha^{(t)} - \alpha^{(t-1)}| < \varepsilon$, or when the first order conditions are sufficiently close to zero: $\max |S(\alpha)|_{\alpha=\alpha^{(t)}} < \varepsilon$. 

Chapter 3. Imputation of Data Subject to One Balance Restriction

3.6 Imputation of missing data items

Now recall the edit constraint given in equation (3.1), where the data items $Y_1, \ldots, Y_l$ are restricted to the simplex. Assume that $Y$ is Dirichlet distributed with parameters $\alpha_1, \ldots, \alpha_l$. Partition the data vector for respondent $i$ into a missing and an observed part: $Y'_i = (Y'_{i,\text{mis}}, Y'_{i,\text{obs}})$ and partition $\alpha$ accordingly. Again $m_i$ denotes the number of missing items for respondent $i$. It holds that

$$Y'_{i,\text{mis}} | Y_{i,\text{obs}} = y_{i,\text{obs}} \sim \text{Dir}_{m_i-1}(\alpha_{i,\text{mis}}),$$

for $Y'_{i,\text{mis}} = (1 - y'_{i,\text{obs}}^l)^{-1} Y_{i,\text{mis}}$. Imputations for $Y_{i,\text{mis}}$ can be obtained either deterministically or stochastically.

Deterministic imputation can be done by using the expected values for the missing items given the observed items, based on the parameter estimates from the EM algorithm, as imputations. Note that, as opposed to distributions where the log likelihood is linear in the data, the calculations in the E-step cannot be used for imputation as $E[\ln Y] \neq \ln E[Y]$. We will therefore use the first order moment to calculate the expected values

$$E[(1 - y'_{i,\text{obs}}^l)^{-1} Y_{i,\text{mis}} | Y_{i,\text{obs}} = y_{i,\text{obs}}, \alpha_{i,\text{mis}}] = \frac{\alpha_{i,\text{mis}}}{\alpha_i},$$

where $\alpha_i = \sum_{j=1}^{m_i} \alpha_j$ and thus

$$Y_{i,\text{imp}} = (1 - y'_{i,\text{obs}}^l)^{-1} \frac{\alpha_{i,\text{mis}}}{\alpha_i}.$$ 

Stochastic imputation can be done by generating random draws from the Dirichlet distribution given in equation (3.1). 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 that $U_1, \ldots, U_l$ are independent random variables having gamma distributions $\text{gamma}(\alpha_1, \lambda), \ldots, \text{gamma}(\alpha_l, \lambda)$. For $j = 1, \ldots, l$, let $Z_j = \frac{U_j}{U_1 + \ldots + U_l}$, then $Z$ has the $(l-1)$-variate Dirichlet distribution $\text{Dir}_{l-1}(\alpha_1, \ldots, \alpha_l)$. Thus random values from the Dirichlet distribution can be obtained by drawing independently from gamma distributions. Let $V$ denote a random draw from (3.1), then

$$Y_{i,\text{imp}} = (1 - y'_{i,\text{obs}}^l)^{-1} V.$$ 

Finally, the imputed proportions can be transformed to imputations for the original data $X_i$ by multiplying them with the total for that respondent, $X_{ik}$:

$$X_{i,\text{imp}} = Y_{i,\text{imp}} x_{ik}.$$
3.7 Imputation performance

In order to assess the performance of this imputation method on empirical data, we will use it to impute data that have been gathered by Statistics Netherlands on a part of the wholesale industry for businesses with more than 10 employees. The effects on estimation of population parameters as well as the imputation performance with respect to individual values will be assessed.

3.7.1 Description of the data

The questioned businesses provide information about their company profits, sales and expenses. Additionally, data is provided on employees, such as the number of employees, the number of temporary employees, their wages and so on.

The first dataset that will be used concerns labour costs, which consists of the following variables

\[ X_{11} = \text{gross wages and salaries} \]
\[ X_{12} = \text{social security costs} \]
\[ X_{13} = \text{pension charges} \]
\[ X_{14} = \text{other social costs} \]
\[ X_{1t} = \text{total labour costs}, \]

which are all non-negative and for which the following balance restriction holds

\[ X_{11} + X_{12} + X_{13} + X_{14} = X_{1t}. \]

The second dataset concerns costs of third party rendering of services, consisting of the variables

\[ X_{21} = \text{costs of banking} \]
\[ X_{22} = \text{insurance premiums} \]
\[ X_{23} = \text{costs of accounting, tax or legal advice} \]
\[ X_{24} = \text{costs of automisation} \]
\[ X_{25} = \text{costs of waste processing} \]
\[ X_{26} = \text{other costs of third party services} \]
\[ X_{2t} = \text{total costs of third party rendering of services}. \]
These variables again are all non-negative and satisfy the balance restriction

\[ X_{21} + X_{22} + X_{23} + X_{24} + X_{25} + X_{26} = X_{21}. \]

### 3.7.2 Estimation of population parameters

As we already mentioned in chapter 1, estimates of population parameters are potentially subject to the following sources of variation: sampling variance and nonresponse variance. Sampling variance is the variance that occurs due to surveying only a subset of the population and would vanish if the whole population was sampled. Nonresponse variance is introduced by the nonresponse in the sample, so if the sample would be completely observed no nonresponse variance would arise. We are currently mostly interested in the nonresponse component of the total variance of the population estimate. A simulation study will be conducted to assess the magnitude of this nonresponse variance for different imputation and incomplete data procedures.

### 3.7.3 Generation of missing data items

In order to be able to assess the effects of the imputation procedures we need to have completely observed data, i.e. data without any missing items, at our disposal. The cases with missing items will therefore be removed from the datasets on labour and third party rendering of services costs, and only the complete cases will be used for analysis as sufficiently large datasets will still remain. Next data items can be removed from these completely observed datasets, which means that for each dataset we will have completely observed data and corresponding data that contain missing values. In order to assess the variance that arises due to nonresponse and imputation several samples will be taken from these complete datasets, with randomly assigned missing data items. These different samples containing missing values will subsequently be used for parameter estimation.

The missing data items will be generated assuming that the missing data are MCAR (missing completely at random). This means that the probability that an item is missing does not depend on the value of that item or other data items in the survey. The missing data are generated using Bernoulli draws with parameter \( p \). The value of \( p \) is chosen such that the percentage of generated nonresponse is similar to the percentage of nonresponse that was observed in the original survey. Although MCAR may be a somewhat unrealistic assumption it does give us insight in the performance of the different imputation methods.
3.7. Imputation performance

regarding point estimation of the parameters and the sensitivity of the outcomes with respect to the realised set of missing items. After generation of the missing data, we will first employ deductive imputation to obtain the imputations that can be derived with certainty based on the observed items in the survey. The remaining datasets are used for imputation.

3.7.4 The effects of imputation on parameter estimation

The main target of this survey is to estimate the population parameters on mean and dispersion: $\mu$ and $\sigma$. To assess both the ability of imputation methods to produce accurate point estimates as well as the effects of nonresponse and imputation on the nonresponse variance component of the total variance of these population parameter estimates we will compare parameter estimates based on the imputed datasets as well as parameter estimates based on the incomplete data with the true parameter estimates for several randomly realised sets of missing data.

Procedures employed for parameter estimation using the incomplete data are the complete cases (CC) approach, which discards all records that contain missing responses, and the available cases (AC) method, which removes missing items on a variable-by-variable basis. The datasets with missing items will consequently be imputed after which the parameters are estimated as well. First of all we will use the Dirichlet approach, where either the expected values (Dir) or random draws (DirR) are used as imputations. Secondly, a ratio hot deck (RHD) approach is used where donors are found at random. Finally, ratios of the nearest neighbours (RNN) are used for imputation. The distance between records is calculated as the difference between the totals ($X_{1t}$ for the first dataset and $X_{2t}$ for the second dataset). Note that the Dirichlet method with expectations imputed and the nearest neighbour ratio imputation method are both deterministic methods, meaning that given a certain set of missing items these methods will always generate the same imputations. The Dirichlet method with random draws imputed and the random ratio hot deck method on the other hand are stochastic procedures, where the imputed values will vary given a set of missing values. These latter two procedures are developed to obtain better point estimates for $\sigma$ and to preserve the distribution of the data better. This will, however, create a larger nonresponse variance. The process of generating and imputing missing items and subsequently estimating the population parameters $\mu$ and $\sigma$ is iterated 100 times.

In Table 3.1 the results on the generation of missing item values are presented for both datasets, the standard errors are shown between brackets. The
Table 3.1: Generation of missing items.

<table>
<thead>
<tr>
<th>Labour costs</th>
<th>Third party rendering of services</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample size</td>
<td>597</td>
</tr>
<tr>
<td>663</td>
<td></td>
</tr>
<tr>
<td># of complete cases</td>
<td>286 (13)</td>
</tr>
<tr>
<td># of missing cases</td>
<td>$X_{11}$ 186 (10)</td>
</tr>
<tr>
<td>$X_{12}$ 185 (11)</td>
<td>$X_{22}$ 214 (12)</td>
</tr>
<tr>
<td>$X_{13}$ 187 (13)</td>
<td>$X_{23}$ 221 (11)</td>
</tr>
<tr>
<td>$X_{14}$ 186 (13)</td>
<td>$X_{24}$ 207 (11)</td>
</tr>
<tr>
<td>$X_{25}$ 201 (12)</td>
<td></td>
</tr>
<tr>
<td>$X_{26}$ 203 (12)</td>
<td></td>
</tr>
</tbody>
</table>

Parameter estimates and their 95% confidence intervals, which represent the nonresponse variance, for the dataset on labour costs are presented in Figures 3.2 and 3.3 and the parameter estimates and 95% confidence intervals for the dataset on the costs of third party rendering of services are presented in Figures 3.4 and 3.5. The solid line refers to the true parameter estimate, which is calculated based on the completely observed data.

Although the missing data mechanism is MCAR, using the complete cases estimates will lead to inaccurate parameter estimates. This is due to the high rate of nonresponse, which leads to a relatively small complete sample, especially for the dataset on third party rendering of services as can be seen in Table 3.1. Using the available cases results in nonresponse rates of about 30% for each variable in both datasets and therefore leads to somewhat more accurate point estimates, which do not satisfy the linear balance restriction however. In both cases the 95% confidence intervals are unacceptably large most of the time, meaning that the nonresponse variance is quite substantial. Consequently the accuracy of the parameter estimates will be highly dependent on the realised set of missing values, which is undesirable. Besides, both methods seriously underestimate the standard deviations in both datasets (Figures 3.3 and 3.5). So, if we are dealing with a considerable number of missing records the complete and available cases estimates cannot be employed.

For the parameter estimates that are based on the imputed datasets concerning labour costs (Figures 3.2 and 3.3) we observe the following. The nearest neighbour method produces accurate point estimates, but sometimes the confid-
ference intervals remain quite large, which reduces precision. The Dirichlet method, where the expectation is imputed, generates acceptable point estimates as well, with higher precision. As was expected the Dirichlet method with random imputations yields similar point estimates, but with less precision due to the added randomness. Beforehand we believed that stochastic imputation methods are necessary to provide non-biased point estimates for the standard deviations. From Figure 3.3 it becomes clear that this is not always the case as reasonable estimates for the standard deviations, with small confidence intervals, are obtained using the Dirichlet method with expectations imputed. This is probably due to the fact that one conditions on the observed data, which show quite some variation and therefore the imputed values will vary as well. Besides, the model parameters $\alpha$ of the Dirichlet distribution need to be estimated, introducing even more variation. Apparently, these two sources of variation result in a good approximation of the actual variation. Furthermore, using a stochastic imputa-
tion procedure results in considerably more nonresponse variance. Especially the random ratio hot deck method produces point estimates with quite large confidence intervals, in particular for the variable $X_{14}$. This method therefore seems to be the worst imputation method for this type of data if one is mainly interested in aggregate values.

The effects of nonresponse and imputation on parameter estimation of $\mu$ and $\sigma$ for the data on third party rendering of services (Figures 3.4 and 3.5) are less clear-cut. First of all we find that all the imputation methods yield similar estimates for $\mu_{21}$, $\mu_{22}$, $\mu_{24}$ and $\mu_{25}$. The nearest neighbour method produces much more accurate point estimates for $\mu_{23}$ and $\mu_{26}$, however. With respect to the nonresponse variance, on average, the nearest neighbour method and the random Dirichlet variance, on average, the nearest neighbour method and the random Dirichlet display the largest variation and the Dirichlet method with expectations imputed results in the smallest nonresponse variance. The estimates of $\sigma$ show a somewhat different pattern. Again, all imputation methods
yield similar estimates for \( \sigma_{12} \) and \( \sigma_{22} \), but for \( \sigma_{24} \) the random Dirichlet and the random hot deck method and for \( \sigma_{25} \) the random Dirichlet together with the nearest neighbour approach clearly perform worse than the other methods. Furthermore, as opposed to the differences in performances for the estimation of \( \mu_{23} \), all methods result in similar deviation of the estimates of \( \sigma_{23} \) from the true value. Finally, for the estimation of \( \sigma_{26} \) the nearest neighbour method remains the best choice. In general, similar to the results obtained from the previous dataset, the nearest neighbour method provides quite reasonable point estimates for both \( \mu \) and \( \sigma \) for almost all variables. The confidence intervals, however, can be substantial especially for the estimates of \( \sigma \), indicating a considerable nonresponse variance. The Dirichlet method with expectations imputed
produces less accurate but mostly quite acceptable point estimates with higher precision, meaning that this method is much less dependent on the realised set of missing data. Again it is also observed that no stochastic imputation procedures are needed to obtain reasonable estimates for \( \sigma \).

In conclusion, it is difficult to state which imputation method performs best, as their performances differ across variables and datasets. Furthermore, a balance needs to be established between required accuracy and precision. A natural candidate to measure the performance of imputation methods, that incorporates both the bias and the variance of an estimate, is the mean squared error (MSE) of an estimate. Unfortunately, the MSE only provides a measure to compare imputation methods within variables and not across variables. Consequently, a
procedure remains to be developed that combines information from the MSE’s for each method across variables in order to determine which method performs best.

An important result that can be derived from Figures 3.2, 3.3, 3.4 and 3.5 is the fact that some methods produce much larger confidence intervals in comparison to the other methods. As we mentioned earlier, large confidence intervals indicate a high nonresponse variance, which means that the imputation or incomplete data method in question is highly dependent on the actual set of missing data. In reality only one of the possible sets of missing data actually is drawn and it is therefore desirable to use a procedure that will result in similar parameter estimates, irrespective of the realised set of missing data. In this sense, the Dirichlet method appears to be a promising alternative with respect to ratio hot deck imputation techniques.

### 3.7.5 The performance of the imputation methods on item level

If the main interest of the imputer is to obtain accurate population estimates as well as a general purpose dataset, other desirable properties arise. Ideally, we would like the results of any statistical analysis on the imputed data to lead to the same conclusions as the same analysis on the true data. Obviously it is impossible to define performance measures that can assess this. The imputation methods will therefore be judged based on their ability to preserve the true values (predictive accuracy) and the distribution of the true values (distributional accuracy). With respect to the preservation of the true values, the average absolute deviation of the imputed data from the true data is calculated. To assess whether the distribution of the imputed data with respect to the true data is preserved, the Kolmogorov-Smirnov statistic, which measures the maximal distance between the empirical distribution functions of the imputed and the true data, is calculated. The results are presented in Table 3.2 for the dataset concerning labour costs and in Table 3.3 for the dataset on costs of third party rendering of services, standard errors are given between brackets.

In Table 3.2 we find that with regard to the predictive accuracy of the imputation methods, the Dirichlet method with expectations imputed performs best. This is not surprising as the expectation seems to be the best prediction. With respect to the preservation of the distribution, this Dirichlet method performs worse than the other methods, particularly for the variable $X_{14}$. Note, however that the Dirichlet method did preserve $\mu_{14}$ and $\sigma_{14}$ well. In this case it seems that nonparametric methods are much better at preserving the dis-
tribution of the variable. This is probably due to the fact that $X_{14}$ is a semi-continuous variable with a high proportion of records with a fixed value (at zero) and a small continuous part. The donor methods will impute much more zero item values in this case, which the Dirichlet with expectations imputed obviously will not. In chapter 6 models are discussed and employed to deal with this particular type of variable. For the other three variables the Kolmogorov-Smirnov distance is relatively small for all methods, indicating that the imputed and the true data are from equal distributions. The random hot deck and nearest neighbour methods, however, do preserve the distributions better as the Kolmogorov-Smirnov distance is smaller for these methods. This means that if the imputer is solely interested in the distribution of the data, the random hot deck and nearest neighbour method are preferred for this dataset.

The data on third party rendering of services yield similar results, which are presented in Table 3.3. The Dirichlet method, with expectations imputed, outperforms the other methods with respect to predictive accuracy, but is outperformed with regard to distributional accuracy. As this dataset consists of large proportions of zeroes in most of these variables, the nonparametric methods perform much better with respect to distributional accuracy. Again, this will be further investigated in chapter 6.

Finally, it is observed that the Dirichlet method with expectations imputed results in relatively small standard errors for both datasets. This means that also with respect to predictive and distributional accuracy the Dirichlet method displays a small variation due to nonresponse.

### 3.8 Concluding remarks

In this chapter we have developed an imputation method that models the data through a Dirichlet distribution in order to impute non-negative data items that are part of a balance restriction from which the total is assumed to be known. As a part of this imputation procedure the EM algorithm for the Dirichlet distribution is derived such that maximum likelihood estimates can be obtained in the presence of nonresponse.

The results in section 3.7 on empirical data show that the choice of an appropriate imputation method strongly depends on the main interest of the data user. If preserving aggregates such as averages, totals and variances is most important to the data user, the Dirichlet method with expectations imputed seems a good approach with high precision. Another good option would be to use the nearest neighbour imputation method. When the main interest of the data user
Table 3.2: Predictive and distributional accuracy of the imputed data on labour costs.

<table>
<thead>
<tr>
<th></th>
<th>Average deviation</th>
<th>Kolmogorov-Smirnov distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dir</td>
<td>DirR</td>
</tr>
<tr>
<td>$X_{11}$</td>
<td>110</td>
<td>266</td>
</tr>
<tr>
<td></td>
<td>(36)</td>
<td>(198)</td>
</tr>
<tr>
<td>$X_{12}$</td>
<td>95</td>
<td>228</td>
</tr>
<tr>
<td></td>
<td>(27)</td>
<td>(144)</td>
</tr>
<tr>
<td>$X_{13}$</td>
<td>110</td>
<td>186</td>
</tr>
<tr>
<td></td>
<td>(37)</td>
<td>(126)</td>
</tr>
<tr>
<td>$X_{14}$</td>
<td>46</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>(15)</td>
<td>(58)</td>
</tr>
</tbody>
</table>

Table 3.3: Predictive and distributional accuracy of the imputed data on third party rendering of services costs.

<table>
<thead>
<tr>
<th></th>
<th>Average deviation</th>
<th>Kolmogorov-Smirnov distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dir</td>
<td>DirR</td>
</tr>
<tr>
<td>$X_{21}$</td>
<td>51</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>(22)</td>
<td>(41)</td>
</tr>
<tr>
<td>$X_{22}$</td>
<td>66</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>(35)</td>
<td>(47)</td>
</tr>
<tr>
<td>$X_{23}$</td>
<td>78</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>(34)</td>
<td>(53)</td>
</tr>
<tr>
<td>$X_{24}$</td>
<td>68</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>(18)</td>
<td>(46)</td>
</tr>
<tr>
<td>$X_{25}$</td>
<td>53</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>(23)</td>
<td>(43)</td>
</tr>
<tr>
<td>$X_{26}$</td>
<td>113</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>(44)</td>
<td>(53)</td>
</tr>
</tbody>
</table>

is to preserve the true values of the data this Dirichlet method seems to be the best approach, but if the data user is mainly interested in the distribution of the data using the random hot deck and nearest neighbour method may be wiser,
especially if the dataset contains a large number of semi-continuous variables.

Another important issue is the fact that the nonresponse variance varies strongly between imputation methods. This means that some imputation methods are much more dependent on the actual set of missing data that is realised. It is preferred to use an imputation method that produces as little nonresponse variance as possible, while still producing accurate parameter estimates, imputed values and so on. In this sense, the Dirichlet method appears to be the best option.

In this chapter the Dirichlet approach has been applied to two relatively small datasets concerning company expenses gathered by a certain business survey. Note, however, that the procedure is applicable to all datasets that consist of continuous non-negative variables that add up to a given total. The Dirichlet method can therefore also be used to impute data on employment or company turnover variables. Furthermore, this model can be straightforwardly extended to other business surveys as almost all business surveys at Statistics Netherlands are based on a similar structure. In social surveys this method may be of use as well, for example for the imputation of household income variables.

Further research is needed with respect to the linear restrictions that can be handled by the imputation procedure. The advantage that the Dirichlet method, as well as the ratio imputation methods, satisfies non-negativity restrictions also has a downside. Some financial variables, such as profit, financial result or exceptional result are allowed to be negative and therefore cannot be imputed using this approach. Besides, economic data usually need to satisfy many linear restrictions. As this method deals with data that are subject to non-negativity constraints and one linear balance restriction only, there is still a need for an imputation method that can take several linear balance and other types of inequality restrictions into account. In the next chapter we will therefore develop an imputation method that can cope with multiple linear balance restrictions.