A latent Gaussian model for compositional data with zeroes

Adam Butler and Chris Glasbey†

Biomathematics & Statistics Scotland, Edinburgh, UK

Summary. Compositional data record the relative proportions of different components within a mixture, and arise frequently in many fields. Standard statistical techniques for the analysis of such data assume the absence of proportions which are genuinely zero. However, real data can contain a substantial number of zero values. We present a latent Gaussian model for the analysis of compositional data which contain zero values, which is based on assuming that the data arise from a (deterministic) Euclidean projection of a multivariate Gaussian random variable onto the unit simplex. We propose an iterative algorithm to simulate values from this model, and apply the model to data on the proportions of fat, protein and carbohydrate in different groups of food products. Finally, evaluation of the likelihood involves the calculation of difficult integrals if the number of components is more than three, so we present a hybrid Gibbs-rejection sampling sampling scheme that can be used to draw inferences about the parameters of the model when the number of components is arbitrarily large.

Keywords: Unit sum constraint; Nutrition; Dietary composition; Gibbs sampling; Approximate Bayesian Computation

1. Introduction

This paper is concerned with the analysis of compositional data that contain zero values. Motivation is provided by data on the proportions of protein, fat and carbohydrate that are contained in a range of different food products. The proportions of the nutrients vary widely between products: for example, "boiled leeks" contain 9.4% fat, 2.3% protein and 88.3% carbohydrate, whereas "raw swordfish" contain 83.2% fat, 16.8% protein and no carbohydrate. Proportions for four particular food groups are plotted in Figure 1 using a ternary diagram - this is an equilateral triangle, whose vertices represent the three elements of the composition; points which lie close to a vertex have high proportions of the element represented by that vertex, whilst points lying in the centre of the triangle have equal proportions of all three elements. Our interest is in attempting to describe the properties of the data for individual groups, and in comparing the nutritional properties of food products in different groups; in particular, to test whether there are significant differences in the composition of the four groups of meat products (beef, pork, poultry and lamb), and between meat and fish products.

The data are compositional, in that they record information about the relative frequencies associated with different components of a system - in this case the proportions associated with different nutrients. Compositional data routinely arise in disciplines such

†*Address for correspondence:* Adam Butler, Biomathematics & Statistics Scotland, James Clerk Maxwell Building, The King's Building, Edinburgh EH9 3JZ, United Kingdom
E-mail: adam@bioss.ac.uk

Fig. 1. Proportion of protein and fat in products belonging to four food groups that contain a relatively large number of zero values. Data are taken from the Nutrient Dataset for Standard Reference (Agricultural Research Service, US Department of Agriculture, 2006). Compositional data are constrained to lie within the unit simplex (shown in grey).
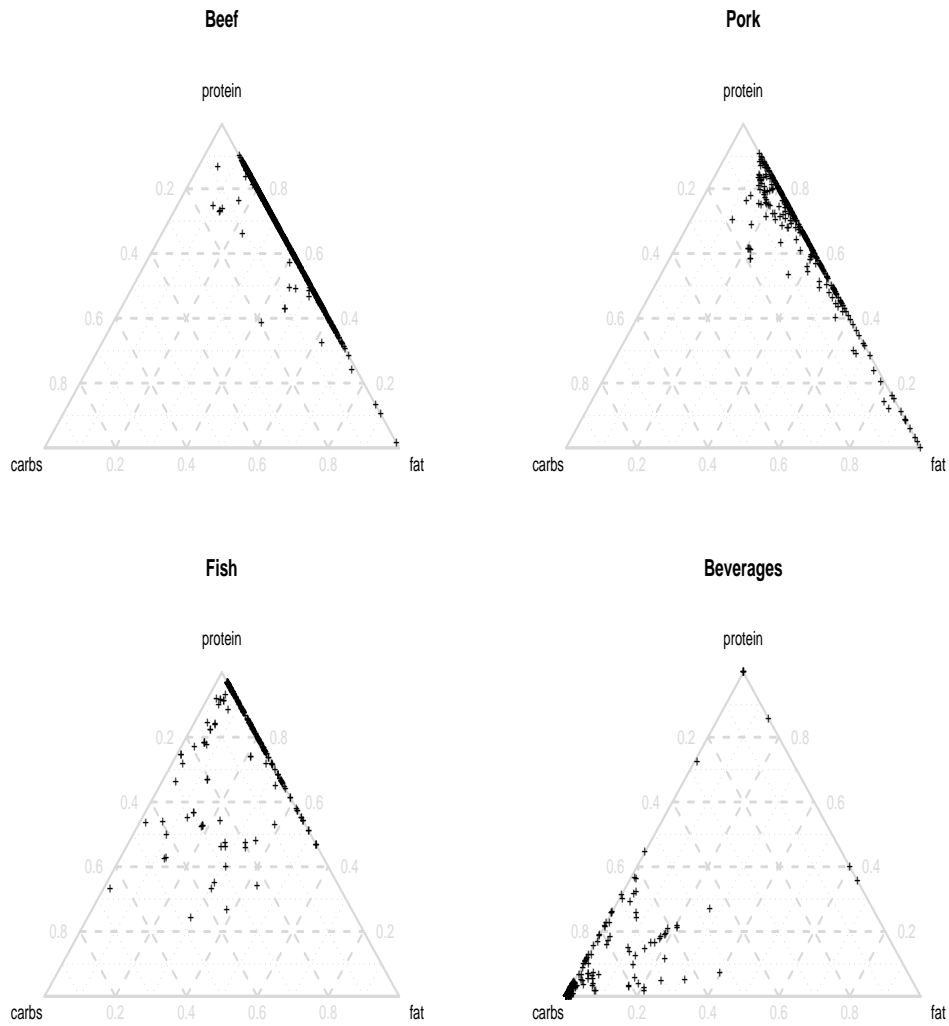
as geology, economics and ecology, and it has long been recognised that they should not be analysed using standard statistical methods because of the intrinsic constraints that (a) the proportions associated with each component must lie between zero and one and (b) the proportions associated with the various components must sum to one. In geometrical terms, a D-dimensional random variable $\mathbf{Y}$ is compositional if it is constrained to lie in the unit simplex,

$$S_D = \{\mathbf{y} \in \mathbb{R}^D : \mathbf{y}^T \mathbf{1} = 1, \mathbf{y} \in [0,1]^D\}.$$

Aitchison (1986) demonstrated that compositional data could be treated in a statistically rigorous fashion by analysing the log-ratios of the proportions, rather than the proportions themselves, and proposed a suite of statistical methods based upon assuming a multivariate Gaussian distribution for these log-ratios. These ideas have subsequently been developed into a comprehensive set of statistical tools (Aitchison and Egozcue, 2005), and the theoretical and practical aspects of compositional data analysis (CODA) remain active fields of methodological research (e.g. Egozcue et al., 2006, von Eynatten et al., 2002).

Certain food groups within the nutrition data contain a high proportion of products for which one or more of the nutrients is absent - for example, the proportion of fat in "dried egg white" is zero, and the proportions of protein and carbohydrate in "bacon grease" are both zero. Similar issues arise in data on seabird dietary intake (e.g. Bull et al., 2004), in pollen data (Salter-Townshend and Haslett, 2006), and in a wide range of other applications. The approach proposed by Aitchison breaks down when the proportions associated with some or all of the components may be zero, because it is no longer possible to calculate the log ratios between the different proportions. Rigorous methods for dealing with the situation in which zero values arise solely through the rounding off of small values have been developed (Elston et al., 1996; Fry et al., 2000; Martín-Fernández et al., 2000; Martín-Fernández et al., 2003; Martín-Fernández et al., 2003; Fry and Chong, 2005; Palarea-Albaladejo et al., 2007a,b). Models to deal with the possibility that zero proportions in the data may correspond to genuine absences of the component concerned ('structural zeroes') have also been proposed (Aitchison and Kay, 2003; Bull et al., 2004), and, in a limited number of cases, used (Bacon-Shone, 2003; Adolph, 2004). There is currently, however, no established methodology for dealing with structural zeroes.

Latent Gaussian models provide a general approach for the analysis of data that are constrained to lie on a subset of the real space $\mathbb{R}^D$. They are based on the idea that the observed data $\mathbf{Y}$ are related, via a transformation g, to a latent variable $\mathbf{Z}$ that has a (univariate or multivariate) Gaussian distribution with unknown parameters. The form of the function g may or may not depend upon the values of additional unknown parameters. Latent Gaussian model have been successfully used to deal with the presence of zero values in a range of applications (Tobin, 1958; Allcroft and Glasbey, 2003a,b), and in this paper we argue that an analogous approach can be used for the analysis of compositional data that contain zero values.

We introduce our proposed model in Section 2, outline an iterative algorithm for simulating realisations from this model, and present the likelihood function associated with the model. In Section 3 we apply the model to data on the proportion of fat, protein and carbohydrate in different food products. We show (Section 4) how Bayesian methods can be used to draw inferences about the parameters of the model when the number of components D is relatively large. In the discussion (Section 5) we outline some of the theoretical limitations of our model, and discuss situations in which the latent Gaussian approach will or will not be appropriate, before outlining some possible ways in which this work could be

extended.

2.   Proposed latent Gaussian model

We propose to model the distribution of $\mathbf{Y}$ using a latent Gaussian model. More specifically, we assume that

$$\mathbf{Y} = g(\mathbf{Z}),$$

where $\mathbf{Z}$ is a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ, and where g is a deterministic function.

The components of $\mathbf{Y}$ must sum to one, and we therefore impose the constraint that the components of the latent variable $\mathbf{Z}$ must also sum to one. In geometric terms, this implies that $\mathbf{Z}$ must lie on the unit hyperplane

$$H_D = \{\mathbf{z} \in \mathbb{R}^D : \mathbf{z}^T \mathbf{1} = 1\}.$$

Imposing such a constraint on the random variable $\mathbf{Z}$ is equivalent to imposing sum constraints on the parameters μ and Σ, such that $\mu^T \mathbf{1} = 1$ and $\Sigma \mathbf{1} = \mathbf{0}$. It follows that the proposed model contains $(D-1)$ unknown parameters within the mean vector μ, and a further $(D-1)D/2$ unknown parameters within the covariance matrix Σ.

2.1.   Choice of transformation function

The function g determines the relationship between a value $\mathbf{z}$ of the latent variable $\mathbf{Z}$, and a value $\mathbf{y}$ of the observed variable $\mathbf{Y}$. Throughout this paper, we take g to be the function which minimises the (squared) Euclidean distance $(\mathbf{y} - \mathbf{z})^T(\mathbf{y} - \mathbf{z})$ between $\mathbf{z}$ and $g(\mathbf{z}) = \mathbf{y}$, subject to the constraints $\mathbf{y}^T \mathbf{1} = 1$ and $\mathbf{y} > 0$. In geometric terms, g performs a Euclidean projection from the unit hyperplane H_D onto the unit simplex S_D. This constrained optimisation problem can be solved by introducing Lagrangian multipliers, λ and γ, and seeking to minimise the Lagrangian

$$L = \frac{1}{2}(\mathbf{y} - \mathbf{z})^T(\mathbf{y} - \mathbf{z}) - (\mathbf{y}^T \mathbf{1} - 1)\lambda - \mathbf{y}^T \gamma$$

with respect to $\mathbf{y}$. Differentiation gives

$$\frac{\partial L}{\partial \mathbf{y}^T} = (\mathbf{y} - \mathbf{z}) - \mathbf{1}\lambda - \gamma,$$

which will be equal to zero if $\mathbf{y} = \mathbf{z} + \mathbf{1}\lambda + \gamma$. By Theorem 9.4.2 of Fletcher (1981) this will be a solution to the constrained optimisation problem under conditions of primal feasibility ($\mathbf{y}^T \mathbf{1} = 1$, $\mathbf{y} \geq 0$), dual feasibility ($\gamma \geq 0$) and complementary slackness ($\mathbf{y}^T \gamma = 0$).

If we can find a set $J \subseteq \{1, \ldots, D\}$ which satisfies

$$j \in J \text{ if and only if } z_j + \lambda < 0, \text{ where } \lambda = \frac{\sum_{k \in J} z_k}{D - |J|}, \tag{1}$$

then these conditions will be satisfied by taking the elements of $\mathbf{y}$ and γ to be

$$y_j = \begin{cases} 0 & \text{if } j \in J \\ z_j + \lambda & \text{if } j \notin J, \end{cases}, \; \gamma_j = \begin{cases} -z_j - \lambda & \text{if } j \in J \\ 0 & \text{if } j \notin J. \end{cases} \tag{2}$$

Note that if the observed variable $\mathbf{y}$ lies on the interior of the unit simplex - that is, if none of the components of $\mathbf{y}$ is equal to zero - then it is trivial to show that the latent variable must be equal to the observed variable (so that $\mathbf{z} = \mathbf{y}$).

2.2. Simulation of data

Simulation of data $\mathbf{y}$ from the proposed model involves (a) simulation of $\mathbf{z}$ from a multivariate Gaussian distribution with mean μ and variance matrix Σ using standard methods, (b) identification of the set J, and (c) calculating the elements of $\mathbf{y}$ using Equation 2.

A simple algorithm can be used to evaluate the set J, which indexes the values of $\mathbf{z}$ that will be transformed by g to be zero. Let $z^{(k)}$ denote the element of $\mathbf{z}$ with rank k, and assume that $z^{(1)} < 0$ (since otherwise J is null and $\mathbf{y} = \mathbf{z}$). Then

(0) Let $k = 2$;

(1) If $z^{(k)} + \frac{1}{D-k+1} \sum_{l=1}^{k-1} z^{(l)} \geq 0$ then set $J = \{j : z_j < z^{(k)}\}$;

(2) Otherwise set $k = k + 1$ and return to step (1).

It can be shown that this leads to a set which satisfies the condition given in Equation 1. If $\phi(j,k) = z^{(j)} + \frac{1}{D-k+1} \sum_{l=1}^{k-1} z^{(l)}$ then we need to show that $\phi(j,k) < 0$ for $j < k$ and $\phi(j,k) \geq 0$ for $j \geq k$. We know that $\phi(k,k) \geq 0$, and is true by definition that $\phi(j+1,k) \geq \phi(j,k)$ for all j, so it is sufficient to show that $\phi(k-1,k) < 0$. Now

$$\phi(k-1,k) = \frac{D-k+2}{D-k+1} z^{(k-1)} + \frac{1}{D-k+1} \sum_{l=1}^{k-2} z^{(l)} = \frac{D-k+2}{D-k+1} \phi(k-1,k-1).$$

We know that $\phi(k-1,k-1) < 0$, since the algorithm failed to terminated at the $(k-1)$-th step, and it follows that $\phi(k-1,k) < 0$, as required. Note that $z^{(D)} + \frac{1}{D-D+1} \sum_{l=1}^{D-1} z^{(l)} = 1$, so the algorithm will always terminate for some $k \leq D$.

2.3. Special cases

If $D = 2$ then the proportion associated with the second component depends in a deterministic way upon the proportion associated with the first component, and our model reduces to assuming that we have a univariate random variable $Y \in [0,1]$ that is defined by

$$Y = \begin{cases} 0 & \text{if} \quad Z < 0 \\ Z & \text{if} \quad 0 \leq Z \leq 1 \\ 1 & \text{if} \quad Z > 1, \end{cases}$$

where $Z \sim N(\mu, \sigma^2)$.

Now consider the case $D = 3$. We can assume, without loss of generality, that the elements of $\mathbf{Z}$ are arranged in ascending order, so that $Z_1 \leq Z_2 \leq Z_3$. Then

$$\mathbf{Y} = \begin{cases} \mathbf{Z} & \text{if } Z_1 \geq 0 \\ (0, Z_2 + Z_1/2, Z_3 + Z_1/2) & \text{if } Z_2 + Z_1/2 \geq 0 \text{ and } Z_1 < 0 \\ (0, 0, 1) & \text{otherwise}, \end{cases}$$

where $\mathbf{Z} \sim \text{MVN}_3(\mu, \Sigma)$ and $\mathbf{Z}^T \mathbf{1} = 1$.

2.4. Likelihood

Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ denote n iid realisations from a random variable $\mathbf{Y}$ whose distribution is given by the proposed latent Gaussian model with unknown parameters $\theta = (\mu, \Sigma)$. The log-likelihood function is of the form

$$
\begin{aligned}
\log \mathbb{P}(\mathbf{Y} = \mathbf{y}|\theta) &= \sum_{i=1}^{n} \log \mathbb{P}(\mathbf{Y} = \mathbf{y}_i; \mu, \Sigma) \\
&= \sum_{i=1}^{n} \log \mathbb{P}(\mathbf{Z} \in \{\mathbf{z}_i \in H_D : g(\mathbf{z}_i) = \mathbf{y}_i\}; \mu, \Sigma) \\
&= \sum_{i=1}^{n} \log \int_{\{\mathbf{z}_i \in H_D : g(\mathbf{z}_i) = \mathbf{y}_i\}} \frac{\exp\left(-\frac{1}{2}(\mathbf{z}_i - \mu)^T \Sigma^{-1}(\mathbf{z}_i - \mu)\right)}{\sqrt{(2\pi)^n |\Sigma|}} d\mathbf{z}_i.
\end{aligned}
$$

If $D = 2$ then the log-likelihood associated with a set of iid observations $y_1, \ldots, y_n$ is

$$
|L| \log \Phi\left(\frac{-\mu}{\sigma}\right) + |U| \log\left[1 - \Phi\left(\frac{1-\mu}{\sigma}\right)\right] - |M|\left(\log \sigma + \frac{\log 2\pi}{2}\right) - \frac{1}{2}\sum_{i \in M}\left(\frac{y_i - \mu}{\sigma}\right)^2,
$$

where $L = \{i : y_i = 0\}$, $U = \{i : y_i = 1\}$ and $M = \{i : 0 < y_i < 1\}$, and where Φ denotes the cumulative distribution function of the standard Gaussian distribution. The log-likelihood for the case $D = 3$ is given in Appendix A.

3. Nutritional composition of food products

Publicly available data from the Agricultural Research Service, US Department of Agriculture (2006) record the nutritional compositional of 7270 individual food types or products, classified into 25 different food groups. The database aims to provide a relatively comprehensive coverage of currently available food products, and to achieve this by collating information derived from the widest possible range of published and unpublished sources. The selection of food types is consequently dependent primarily on data availability, rather than on any systematic attempt to sample from the set of all food products. Sampling methodologies for the collection of nutrient data also differ between food types, leading to heterogeneity in data quality and precision.

We do not address these issues here, but focus on comparing the relative proportions of protein, fat and carbohydrate between different food groups. We concentrate on the eight groups - beef, beverages, ethnic, fats, fish, lamb, pork and poultry - that contain zero values in at least one nutritional component for more than 25% of the foods within that group; the remaining food groups contain either a low proportion of zero values (e.g. spices, baby food) or contain no zero values whatsoever (e.g. vegetables, snacks).

We estimate the parameters of the model by numerical maximisation of the likelihood function using the Nelder-Mead algorithm, and use the inverse of the observed information matrix as an approximation to the associated covariance matrix. The observed information is found by evaluation of the Hessian matrix at the maximum likelihood estimates. All computations are performed in R (R Development Core Team, 2006), Version 2.1.1, and optimisation is performed using the `optim` function. For the five groups that are of most interest (the meat and fish groups) we have verified that the results obtained by maximum likelihood are close to those obtained using the Bayesian approach that we will introduce in Section 4.

Table 1. Maximum likelihood estimates for the five parameters of the latent Gaussian model, as fitted (separately) to data for eight food groups in the USDA dataset. Standard errors are also given, in brackets.

Food group	μ_1	μ_2	Σ_{11}	Σ_{22}	Σ_{12}
Beef	.833 (.032)	.505 (.032)	.0337 (.0042)	.0292 (.0034)	-.0159 (.0030)
Lamb	.844 (.038)	.473 (.039)	.0309 (.0050)	.0507 (.0090)	-.0228 (.0045)
Pork	.669 (.011)	.358 (.011)	.0319 (.0008)	.0365 (.0010)	-.0318 (.0002)
Poultry	.771 (.013)	.334 (.011)	.0410 (.0037)	.0239 (.0021)	-.0214 (.0022)
Fish	.930 (.025)	.302 (.022)	.0623 (.0090)	.0261 (.0045)	.0086 (.0058)
Fats	.010 (.005)	1.088 (.037)	.0012 (.0002)	.1717 (.0268)	-.0063 (.0021)
Ethnic	.553 (.037)	.273 (.027)	.1640 (.0228)	.0862 (.0123)	-.0014 (.0013)
Beverages	.064 (.015)	-.010 (.011)	.0445 (.0052)	.0166 (.0023)	.0048 (.0023)

Maximum likelihood estimates are given in Table 1, and the probability contours of the fitted model are presented in Figure 2 for four of the food groups; for each group we also show a simulated dataset, generated by setting the parameter values equal to the maximum likelihood estimates. Note that the MLEs for μ lie outside the unit simplex for seven of the eight food groups, reflecting the fact that the data for each group contain a high proportion of zero values in one or two of the three components but a low proportion of zero values in the remaining components. For example, the four meat groups - pork, poultry, lamb and fish - tend to contain a high proportion of fat and protein but a low or zero proportion of carbohydrates.

Using Table 1 we compare the distribution of the maximum likelihood estimates for the five groups which relate to meat or fish products. We can see by eye that there are differences between the parameter estimates of the different groups. In particular, estimates for the "fish" group differ markedly from those associated with the other four groups. Model selection procedures based on likelihood ratio testing and the Akaike Information Criterion reach the same conclusion - that the best model retains separate parameters for each of the five groups, rather than attempting to pool parameters across two or more of the groups.

In Table 2 we assess goodness of fit by comparing the observed frequencies with which particular patterns of zero values occur against the corresponding values that we would have obtained from the fitted model. A comparison of these frequencies provides an important indicator of the extent to which the model provides a reasonable description of the data. We see that the "observed" and "fitted" frequencies are typically very similar for the groups that relate to meat and fish products: lamb, fish, beef, pork and poultry. For these groups, the observed data indicate that foods either contain only protein and fats, or else contain all three nutrient types, and the model performs well in distinguishing the frequency with which the data belong to these two groups. The only discrepancy is that the fitted model suggests that there should be a small number of food products containing "only protein", and, for pork and fish, an even smaller number containing "protein and carbohydrate", whereas there are no such products in the real data.

There are discrepancies between observed and modelled frequencies for the remaining three groups - "fats", "beverages" and "ethnic" - indicating lack of fit: in particular, the model tends to under-estimate the proportion of points that lie on the interior of the simplex, and to over-estimate the proportion that lie on the vertexes. The poor performance of the model reflects the fact that these are heterogeneous groupings of food products, and that the data for these groups show evidence of multi-modality. Our model could be adapted to deal with multi-modal data by regarding **Z** as a mixture of Gaussian distributions, but we

Fig. 2. Latent Gaussian model fitted to four food groups in the USDA dataset by maximum likelihood. Contours of the fitted latent distribution (i.e. $\mathbf{z}$) are shown (shaded), together with simulated data from the fitted model (crosses) - i.e. $\mathbf{y} = g(\mathbf{z})$. Simulated datasets are of the same size as the original dataset for that group.
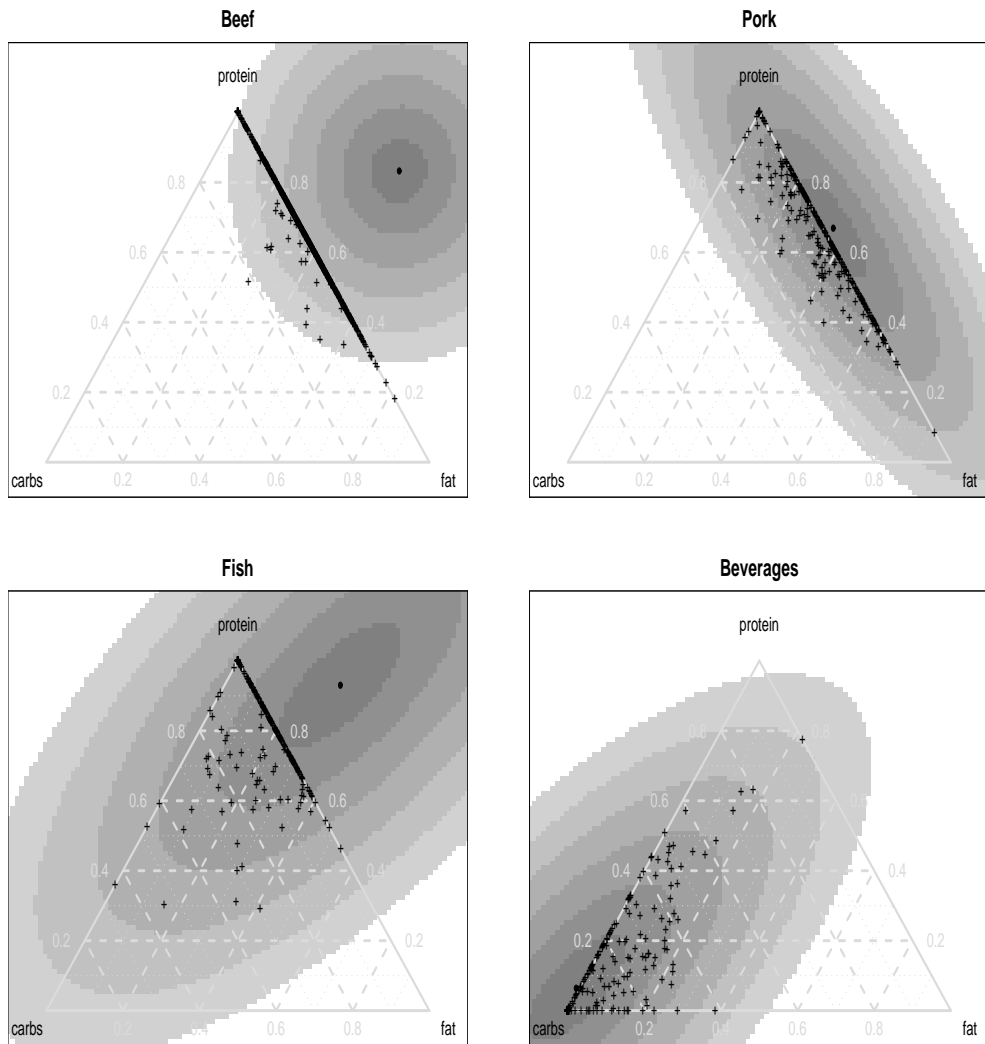
Table 2. Observed and expected patterns of zero values in the USDA data for eight individual food groups. We tabulate the number of products that contain only protein (p), only fat (f), only carbohydrate (c), only protein and fat (pf), only protein and carbohydrate (pc), only fat and carbohydrates (fc), and all three nutrients (pfc). Expected values are calculated by numerical integration, based on the maximum likelihood estimates. The corresponding standard errors (SE) are calculated as $\sqrt{np(1-p)}$, where n is the number of products in the food group and p is the expected proportion of values having that pattern of zero values.

Food group					Number of individual foods			
		p	f	c	pf	pc	fc	pfc
Beef	Observed	0.0	0.0	0.0	767.0	0.0	0.0	22.0
	Expected	11.3	0.0	0.0	755.8	0.3	0.0	21.6
	(SE)	3.3	0.1	0.0	5.6	0.5	0.0	4.6
Lamb	Observed	0.0	0.0	0.0	327.0	0.0	0.0	16.0
	Expected	12.2	0.0	0.0	315.2	2.2	0.0	13.4
	(SE)	3.4	0.1	0.0	5.1	1.5	0.0	3.6
Pork	Observed	0.0	0.0	0.0	184.0	0.0	0.0	105.0
	Expected	6.0	0.0	0.0	185.1	4.0	0.0	93.9
	(SE)	2.4	0.2	-	8.2	2.0	0.1	8.0
Poultry	Observed	0.0	0.0	0.0	254.0	0.0	0.0	84.0
	Expected	14.0	0.0	0.0	242.6	1.4	0.0	79.9
	(SE)	3.7	0.0	0.0	8.3	1.2	0.2	7.8
Fish	Observed	0.0	0.0	0.0	193.0	0.0	0.0	56.0
	Expected	19.4	0.0	0.0	170.8	6.2	0.0	52.5
	(SE)	4.2	0.0	0.1	7.3	2.5	0.1	6.4
Fats	Observed	0.0	107.0	0.0	8.0	0.0	4.0	85.0
	Expected	0.0	114.7	0.4	7.9	0.4	19.8	60.8
	(SE)	0.0	7.1	0.7	2.7	0.7	4.2	6.5
Ethnic	Observed	0.0	4.0	1.0	41.0	1.0	1.0	84.0
	Expected	6.6	0.1	3.9	41.9	18.0	8.3	53.2
	(SE)	2.5	0.3	1.9	5.3	3.9	2.8	5.6
Beverages	Observed	4.0	0.0	49.0	3.0	63.0	10.0	101.0
	Expected	0.0	0.0	79.3	0.0	57.5	20.4	72.8
	(SE)	0.0	0.0	7.2	0.2	6.6	4.3	7.1

have not attempted such an extension here.

Our analysis should be regarded as exploratory rather than definitive, since we have assumed here that the nutritional data for products within a particular food group are independent and identically distributed. In reality, methods of sampling and levels of intra-product variability will differ markedly between individual food products, and the data are probably subject to residual dependence at the level of products within groups. In addition, the products are not of equal importance in terms of human consumption, and would ideally be weighted according to data on intake. It should, at least in principle, be possible to deal with each of these issues by adapting or extending the latent Gaussian model.

4.   Inference in higher dimensions

The integrals involved in calculating the log-likelihood function (3) are straightforward when $D = 2$ or 3, for the choice of g that we have adopted, but become difficult once D is moderately large (e.g. $D \geq 4$) because the sets $\{\mathbf{z} \in H_D : g(\mathbf{z}) = \mathbf{y}\}$ become increasingly complex and difficult to define. Numerical maximum likelihood estimation does not, therefore, provide a viable approach to statistical inference for general D. An alternative, Bayesian, approach is to assign a prior distribution $\pi(\theta)$ to the parameters $\theta = (\mu, \Sigma)$, and to draw inferences about θ by simulating values from the posterior distribution of $\theta|\mathbf{y}$.

A standard method for simulating the posterior distribution of $\theta|\mathbf{y}$ is to generate values from a Markov chain whose equilibrium distribution is equal to $\theta|\mathbf{y}$ (Markov chain Monte Carlo; Gilks et al., 1996), and the simplest way to do this is via Gibbs sampling (Casella and George, 1992). For our model, we suggest the use of a hybrid Gibbs-rejection sampling algorithm that alternates between two steps

1 simulation of $\mathbf{z}|(\mathbf{y}, \theta)$ using the rejection algorithm presented in Appendix B; and

2 simulation of parameter values from $\theta|(\mathbf{z}, \mathbf{y}) = \theta|\mathbf{z}$;

following initialisation of the parameters θ. The algorithm generates a series of dependent realisations from the posterior distribution of $\theta|\mathbf{y}$, once convergence to the equilibrium distribution of the chain has been achieved. Standard Bayesian methods allow us to simulate from the parameters of the multivariate normal distribution, $\theta|\mathbf{z}$. The parameters are subject to sum constraints ($\mu^T \mathbf{1} = 1$ and $\Sigma \mathbf{1} = \mathbf{0}$), so, to avoid singularities, we work only with the first $D - 1$ components of μ and Σ.

We do not wish to make any strong prior assumptions about the values of the parameters $\theta = (\mu, \Sigma)$. One possibility is to adopt the Jeffrey's prior (Jeffreys, 1946)

$$\pi(\mu, \Sigma) \propto |\Sigma|^{-(D+1)/2}, \tag{3}$$

which is uninformative (diffuse) and invariant to reparameterisation. Simulation of $\theta|\mathbf{z}$ then depends on a Gibbs algorithm that alternates between

2a simulation of $(\mu|\Sigma, \mathbf{z}) = (\mu|\Sigma, \mathbf{z})$ from a multivariate normal distribution with mean vector $\bar{\mathbf{z}} = \sum_{i=1}^{n} \mathbf{z}_i/n$ and variance matrix Σ/n; and

2b simulation of $(\Sigma|\mu, \mathbf{z}) = (\Sigma|\mu, \mathbf{z})$ from an inverse Wishart distribution with $D + n - 1$ degrees of freedom and variance matrix $V(\mathbf{z}) = \sum_{i=1}^{n} (\mathbf{z}_i - \mu)(\mathbf{z}_i - \mu)^T$.

The prior in Equation 3 is improper, however, and we cannot demonstrate that it will necessarily lead to a proper posterior distribution for $(\theta, \mathbf{z})|\mathbf{y}$ (Robert and Casella, 1999, pp. 328-329).

An alternative approach would involve assigning proper but diffuse priors to μ and Σ. The complicating factor is that these prior distributions would need to fulfil the constraints $\mu^T \mathbf{1} = 1$ and $\Sigma \mathbf{1} = \mathbf{0}$, and that we would also expect that they ought, in the absence of any genuine prior information, to be symmetric in the components of $\mathbf{Y}$. Priors for the variance matrix are conventionally formulated in terms of a distribution for the precision matrix, Σ^{-1}, but this approach is not possible here because Σ is singular. The construction of appropriate proper prior distributions therefore appears to be non-trivial.

4.1. Simulation study

We use a small simulation study to compare the posterior means and credible intervals generated by the rejection-Gibbs scheme against the corresponding estimates and confidence intervals generated by numerical maximum likelihood. We restrict attention to the case $D = 3$ (qualitatively similar results are obtained using $D = 2$, not shown), and apply the methods to four datasets, each of size 200. The values of μ are taken to be either (A) $\mu_1 = \mu_2 = \mu_3 = 1/3$ or (B) $\mu_1 = 4/3, \mu_2 = \mu_3 = -1/6$ and the covariance matrix is either (a) $\Sigma = \Omega/9$ or (b) $\Sigma = \Omega$, where $\Omega = (3\mathbf{I} - \mathbf{1})/2$. In dataset Aa a high proportion of points lie on the interior (104/200) with few on the vertexes (9/200), whilst in datasets Ba and Bb the majority of points lie on the vertexes (148 and 134, respectively) and few lie on the interior (4 and 10 respectively). Dataset Ab is intermediate, with 23 points on the interior and 85 on the vertexes.

Maximum likelihood estimates are computed via numerical optimisation, with confidence intervals constructed either using the inverse of the hessian or via profile likelihood - the latter is more computationally intensive, but allows for asymmetry in the distributions of the parameter estimators. The MCMC results for datasets Aa and Ab are based on 50000 iterations, following a burn-in of 10000 iterations. The results for datasets Ba and Bb are based on four parallel chains, each of length 60000, which were started from widely dispersed initial values - these chains are then combined, so that the final results are based on 200000 iterations. The Heidelberger-Welch tests, Geweke Z-scores, and (for datasets Ba and Bb) the Gelman-Rubin convergence diagnostics detected no significant evidence of non-stationary. The chain lengths and burn-in periods were selected based on the Raftery-Lewis convergence diagnostic, and on visual inspection of the trace-plots.

For datasets Aa and Ab, in which the majority of data lie on the interior or edges of the simplex, we see (Figure 3) that the results obtained using MCMC are very similar to those obtained via numerical maximum likelihood. For datasets Ba and Bb most of the data lie at the vertexes, so that, in the context of our model, these data contain a high proportion of fully or partially censored values. The high degree of censoring leads to a shallow gradient in the likelihood function, and so presents challenging problems for statistical inference. We see that there are reasonably substantial differences between the results obtained using maximum likelihood and MCMC, with the credible intervals obtained via MCMC tending to be wider than the corresponding confidence intervals based on the profile likelihood.

5. Discussion

We have outlined a novel methodology for analysing compositional data that contain zero values, and have illustrated the application of the model to simulated and real data. The latent Gaussian model that we have proposed is conceptually simple, and similar to models that are used in many other contexts to describe the properties of data which are subject to constraints. The model is relatively parsimonious, especially when D is large, and the parameters should be estimable even from data that contain a high proportion of zero values. Simulation from the model is also straightforward, because of the assumption that the latent variable is multivariate normal.

We acknowledge, however, that there are theoretical difficulties with the model that we have proposed. In particular, the model treats compositional data which lie on the interior of the unit simplex as having arisen from a multivariate normal distribution. In doing so, it violates two of the principles - scale invariance and subcompositional coherence - that

Fig. 3. Maximum likelihood estimates with 95% confidence intervals (lower) and posterior means with 95% credible intervals (upper) for each parameter in each of the four simulated datasets (Aa, Ab, Ba and Bb). Confidence intervals are calculated using both profile likelihood (black) and the inverse of the hessian evaluated at the MLE (grey), except for dataset Ba (where the profile likelihood failed to converge).
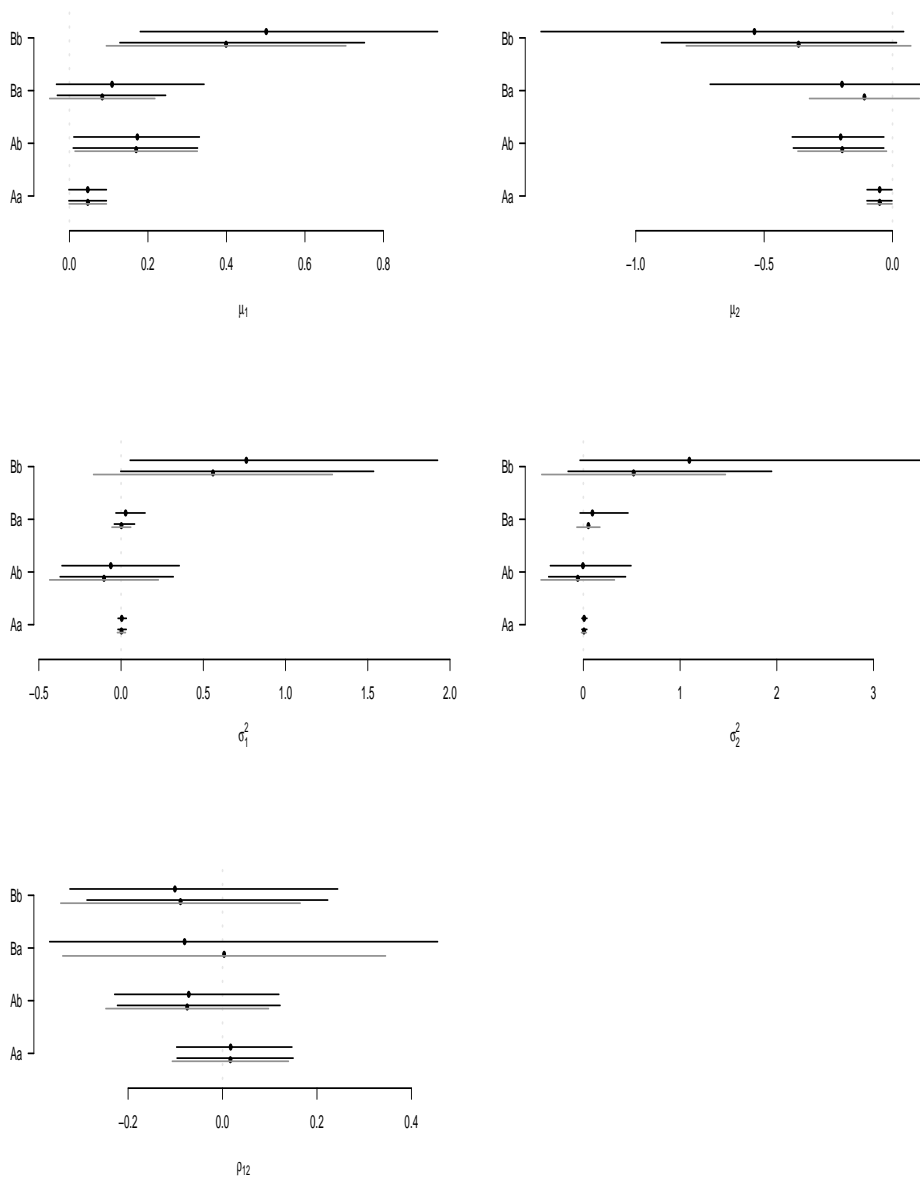
models for data on the interior of the simplex should adhere to (Aitchison, 1986; note that it does adhere to the third property, that of permutation invariance). Scale invariance requires that all inferences should be expressed solely in terms of the ratios of proportions, whilst subcompositional coherence requires that inferences about a particular subset of components should be unaffected if components outside that subset are combined together. Any approach that attempts to describe zero and non-zero proportions using a common model will inevitably break these principles, because ratios of proportions are infinite along the boundaries of the simplex. A related issue concerns the interpretation of the model parameters, μ and Σ, which do not - for the reasons outlined by Aitchison (1986) - have any natural or convenient interpretation in terms of the statistical properties of Y. In particular, the unit sum constraint implies that the elements of the covariance matrix, Σ, cannot be used as a basis for making statements about the (in)dependence of components of Y. As a result, our model should not be used as a basis for principal components analysis, or for other methods that use the form of Σ to explore the dependence structure of Y.

Although our model in principle provides a basis for the analysis of any compositional dataset containing zero values, these theoretical limitations imply that, in practice, it should only be used (a) in situations where existing methods are unavailable or inappropriate, (b) as a diagnostic tool to assess the performance of other methods, or (c) for exploratory data analysis. Zero values that arise as a result of rounding error can be dealt with using zero replacement strategies, so situation (a) will generally only apply to data that contain structural zeroes. In that context, a hierarchical approach (Aitchison and Kay, 2003) can be used if the ratio of data to components is relatively large, but may not be viable if the data are sparse or contain a high proportion of zero values. The potential diagnostic uses of our model merit further research. The model provides a basis for drawing inferences about the patterns of zero values, and, in particular, for calculating the probability that data will lie on a particular vertex or edge of the simplex. These probabilities could subsequently be used, for example, to assist in the formulation of an appropriate hierarchical model, or for the detection of outliers in the context of a zero replacement model.

Our model makes implicit assumptions about the relative frequencies with which data lie on the vertexes, edges and interior of the unit simplex. These assumptions depend on the form of the function, g, that has been selected to describe the relationship between the latent variable $\mathbf{Z}$ and the data $\mathbf{Y}$. We have chosen g such that it performs a Euclidean projection from the unit hyperplane to the unit simplex, since the Euclidean transformation is widely used and has attractive general theoretical properties. One drawback of this choice, however, is that inferences about the distribution of particular component, Y_1, will differ depending on whether or not we choose to aggregate some of the other components within $\mathbf{Y}$. It would therefore be interesting to explore the theoretical and empirical properties of alternative transformation functions. We would, ideally, like to infer the form of g based on the empirical properties of Y, by, for example, estimating the unknown parameters associated with a flexible parametric family of possible transformation functions, but we expect that real compositional datasets will rarely contain sufficient information to enable us to distinguish between alternative choices for g.

Finally, we have applied the model to data with three components, in order to illustrate the methodology, but we acknowledge that most practical applications involve four or more components. The hybrid Gibbs-rejection sampling approach offers a promising approach to inference in higher dimensions, but requires further refinement in order to provide an efficient and robust algorithm for use in that context. Approximate Bayesian Computation provides a possible alternative approach to inference in situations where, as here, the like-

lihood function is difficult to compute, but initial simulation studies have suggested that these methods have relatively poor performance in the context of our model (Butler et al., 2007). Another important issue, which becomes critical in higher dimensions, concerns the assessment of model fit in latent Gaussian models, and in future research we hope to develop diagnostic tools that are appropriate for this purpose.

Acknowledgements

Appendix A: form of the log-likelihood function when $D = 3$

Let $D = 3$ and consider the log-likelihood contribution from a single point $\mathbf{y} = (y_1, y_2, y_3)$.

Interior

If $\mathbf{y}$ lies in the interior of the unit simplex, so that $y_1 > 0, y_2 > 0, y_3 > 0$, then

$$\log \mathbb{P}(\mathbf{Y} = \mathbf{y}) = -\frac{3}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\mathbf{y} - \mu)^T \Sigma^{-1}(\mathbf{y} - \mu).$$

Vertexes

If $\mathbf{y}$ lies at a vertex of the unit simplex, then we can assume without loss of generality that $\mathbf{y} = (0, 0, 1)$. If $\Delta = \{\mathbf{z} \in H_3 : g(\mathbf{z}) = (0, 0, 1)\}$ then

$$\log \mathbb{P}(\mathbf{Y} = \mathbf{y}) = \log \int_\Delta \frac{\exp\left(-\frac{1}{2}(\mathbf{z} - \mu)^T \Sigma^{-1}(\mathbf{z} - \mu)\right)}{\sqrt{(2\pi)^2 |\Sigma|}} d\mathbf{z}.$$

We know that $\mathbf{z} \in \Delta$ if and only if $z_1 + z_2/2 < 0$ and $z_2 + z_1/2 < 0$. Let $\mathbf{u} = A^T \mathbf{z}$, where

$$A = \frac{1}{3} \begin{bmatrix} -1 & 1 \\ 1 & -1 \\ 3 & 3 \end{bmatrix},$$

so that

$$\mathbf{u} = \left(\frac{z_2 - z_1}{3} + z_3, \frac{z_1 - z_2}{3} + z_3\right) = \left(1 - \frac{4}{3}\left[z_1 + \frac{z_2}{2}\right], 1 - \frac{4}{3}\left[z_2 + \frac{z_1}{2}\right]\right).$$

It follows that $\mathbf{z} \in \Delta$ if and only if $u_1 > 1$ and $u_2 > 1$, so that

$$\log \mathbb{P}(\mathbf{Y} = (0, 0, 1)) = \log \int_{\mathbf{u}=(1,1)}^{(\infty,\infty)} \frac{\exp\left(-\frac{1}{2}(\mathbf{u} - A^T\mu)^T (A^T\Sigma A)^{-1}(\mathbf{u} - A^T\mu)\right)}{\sqrt{(2\pi)^2 |A^T\Sigma A|}} d\mathbf{u}.$$

Equivalent formulae can be derived for other vertexes simply by switching labels. The bivariate integral can be evaluated using standard algorithms.

Edges

We can assume, without loss of generality, that $\mathbf{y} = (0, y_2, y_3)$ for $y_2 > 0, y_3 > 0$. Hence

$$\log \mathbb{P}(\mathbf{Y} = \mathbf{y}) = \log \mathbb{P}(Y_2 = y_2, Y_3 = y_3) + \log \mathbb{P}(Y_1 \leq 0 | Y_2 = y_2, Y_3 = y_3).$$

We perform a change of variable. Let $\mathbf{U} = B^T \mathbf{Y}$, where

$$B = \frac{1}{\sqrt{3}} \begin{bmatrix} 0 & \sqrt{3} \\ 1 & 0 \\ -1 & 0 \end{bmatrix},$$

so that $U_1 = Y_1$ and $U_2 = (Y_2 - Y_3)/\sqrt{3}$. It follows that $\mathbf{U} \sim N(m, \Psi)$, where $m = B^T \mu$ and $\Psi = B^T \Sigma B$. If $\mathbf{u} = B^T \mathbf{y}$ then

$$
\begin{aligned}
\log \mathbb{P}(\mathbf{U} = \mathbf{u}) &= \log \mathbb{P}(U_2 = u_2) + \log \mathbb{P}(U_1 \leq 0 | U_2 = u_2) \\
&= -\frac{1}{2}(\log 2\pi + \log \Psi_{22}) - \frac{(u_2 - m_2)^2}{2\Psi_{22}} + \log \Phi \left(-\frac{m_1 + \frac{\Psi_{12}(u_2 - m_2)}{\Psi_{22}}}{\sqrt{\Psi_{11} - \frac{\Psi_{12}^2}{\Psi_{22}}}} \right),
\end{aligned}
$$

where Φ denotes the cumulative distribution function for a standard Gaussian distribution.

Appendix B: algorithm for simulating $\mathbf{z}|(\mathbf{y}, \mu, \Sigma)$

We present an algorithm for simulating values of the latent variable $\mathbf{z} \in H_D$ conditionally upon data $\mathbf{y} \in S_D$ and upon values of the parameters μ and Σ; this is used in the hybrid rejection-Gibbs sampling algorithm of Section 4. Let

$$J = \{1 \leq k \leq D : y_k = 0\}$$

denote the set of the indices of the components of $\mathbf{y}$ which are zero. Then the algorithm proceeds as follows:

(a) If $|J| = 0$ then set $\mathbf{z} = \mathbf{y}$ and exit;
(b) Repeatedly simulate $\mathbf{z}_J$ from a $|J|$-dimensional multivariate normal distribution with mean μ_J and variance matrix Σ_{JJ} until we generate a value such that

$$\left[z_J^{(j)} + \sum_{k=1}^{j-1} \frac{z_J^{(k)}}{D - j + 1} \right] < 0 \text{ for all } j \leq |J|;$$

(c) Calculate the remaining components of $\mathbf{z}$ using

$$\mathbf{z}_{-J} = \mathbf{y}_{-J} - (1/(D - |J|))\mathbf{z}_J^T \mathbf{1}.$$

The justification for the algorithm relies on recognising that it is just a variant of rejection sampling. The most direct way to generate values of $\mathbf{z}|(\mathbf{y}, \mu, \Sigma)$ would be to simulate $\mathbf{z}$ from a multivariate normal distribution with mean μ and variance matrix Σ, and to accept $\mathbf{z}$ if and only if $g(\mathbf{z}) = \mathbf{y}$. However, we already know, from Section 2.2, that values of $\mathbf{z}$ must fulfil the constraint that

$$\mathbf{z}_{-J} = \mathbf{y}_{-J} - (1/(D - |J|))\mathbf{z}_J^T \mathbf{1}.$$

It follows that it is only necessary to simulate the subset of values $\mathbf{z}_J$ relating to the set of components for which $\mathbf{y}$ has zero values, since the remaining values $\mathbf{z}_{-J}$ will depend upon these in a deterministic way. Section 2.3 also provides us with the order restrictions that the elements of $\mathbf{z}_J$ must satisfy to ensure that $g(\mathbf{z}) = \mathbf{y}$, and step (b) of the algorithm involves checking that these restrictions do indeed hold for the value of $\mathbf{z}_J$ which we have simulated.

References

Adolph, C. (2004). Succession in the temple: Central banker careers and the politics of appointment. Working paper, University of Washington.

Agricultural Research Service, US Department of Agriculture (2006). *USDA Nutrient Database for Standard Reference, Release 19*. http://www.ars.usda.gov/nutrientdata.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Aitchison, J. and J. J. Egozcue (2005). Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology 37*(7), 829–850.

Aitchison, J. and J. W. Kay (2003). Possible solutions of some essential zero problems in compositional data analysis. In S. Thió-Henestrosa and J. A. Martín-Fernández (Eds.), *Proceedings of CODAWORK'03, The First Compositional Data Analysis Workshop*, University of Girona, Spain. October 15-17.

Allcroft, D. J. and C. A. Glasbey (2003a). Analysis of crop lodging using a latent variable model. *Journal of Agricultural Science 140*, 383–393.

Allcroft, D. J. and C. A. Glasbey (2003b). A latent Gaussian Markov random field model for spatio-temporal rainfall disaggregation. *Applied Statistics 52*, 487–498.

Bacon-Shone, J. (2003). Modelling structural zeros in compositional data. In S. Thió-Henestrosa and J. A. Martín-Fernández (Eds.), *Proceedings of CODAWORK'03, The First Compositional Data Analysis Workshop*, University of Girona, Spain. October 15-17.

Bull, K., S. Wanless, D. A. Elston, F. Daunt, S. Lewis, and M. P. Harris (2004). Local-scale variability in the diet of Black-legged Kittiwakes *rissa tridactyla*. *Ardea 92*(1), 43–52.

Butler, A., C. A. Glasbey, and S. Wanless (2007). Approximate Bayesian inference for a latent Gaussian model of compositional data. Technical report, Biomathematics and Statistics Scotland, Edinburgh.

Casella, G. and E. I. George (1992). Explaining the Gibbs sampler. *The American Statistician 46*, 167–174.

Egozcue, J. J., J. L. Diaz Barrero, and V. Pawlowsky-Glahn (2006). Hilbert spaces of probability density functions based on Aitchison geometry. *Acta Mathematica Sinici - English Series 22*(4), 1175–1182.

Elston, D. A., A. W. Illius, and I. J. Gordon (1996). Assessment of preference among a range of options using log ratio analysis. *Ecology 77*(8), 2538–2548.

Fletcher, R. (1981). *Practical Methods of Optimization*. Chichester (England): Wiley.

Fry, J., T. R. L. Fry, and K. R. McLaren (2000). Compositional data analysis and zeros in micro data. *Applied Economics 32*(8), 953–959.

Fry, T. R. L. and D. Chong (2005). A tale of two logits, compositional data analysis and zero observations. In G. Mateu-Figueras and C. Barceló-Vidal (Eds.), *Proceedings of CODAWORK'05, The Second Compositional Data Analysis Workshop*, Univeristy of Girona, Spain. October 19-21.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. London. A 186*, 453–461.

Martín-Fernández, J. A., C. B. Barceló-Vidal, and V. Pawlowsky-Glahn (2000). Zero replacement in compositional data. In H. A. L. Kiers, J. P. Rasson, P. J. F. Groenen, and M. Schader (Eds.), *Advances in Data Science and Classification. Proceedings of the 7th Conference of the International Federation of Classification Societies, University of Namur (Belgium)*, pp. 155–160. Springer-Verlag (Berlin).

Martín-Fernández, J. A., C. B. Barceló-Vidal, and V. Pawlowsky-Glahn (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology 35*(3), 253–278.

Martín-Fernández, J. A., J. Palarea-Albaladejo, and J. Gómez-García (2003). Markov chain monte carlo method applied to rounding zeros of compositional data: First approach. In S. Thió-Henestrosa and J. A. Martín-Fernández (Eds.), *Proceedings of CODAWORK'03, The First Compositional Data Analysis Workshop*, University of Girona, Spain. October 15-17.

Palarea-Albaladejo, J., J. A. Martín-Fernández, and J. A. Gómez-García (2007a). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computer and Geosciences*. In revision.

Palarea-Albaladejo, J., J. A. Martín-Fernández, and J. A. Gómez-García (2007b). Parametric approach for dealing with compositional rounded zeros. *Mathematical Geology 39*(8).

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Robert, C. P. and G. Casella (1999). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

Salter-Townshend, M. and J. Haslett (2006). Zero-inflation of compositional data. In *Proceedings of the 21st International Workshop on Statistical Modelling*, Galway, pp. 448–456. July 3-7.

Tobin, J. (1958). Estimation for relationships with limited dependent variables. *Econometrica 26*(1), 24–36.

von Eynatten, H., V. Pawlowsky-Glahn, and J. J. Egozcue (2002). Understanding perturbation on the simplex: a simple method to better visualise and interpret compositional data in ternary diagrams. *Mathematical Geology 34*(3), 249–257.