

# A latent Gaussian model for multivariate consumption data

D.J. Allcroft, C.A. Glasbey\*  
Biomathematics and Statistics Scotland  
King's Buildings, Edinburgh, EH9 3JZ, Scotland

and  
M.J. Paulo  
Biometris  
P.O. Box 16, 6700 AA Wageningen, The Netherlands

March 1, 2006

## Abstract

We propose a multivariate statistical model for individual consumption of multiple food types, to provide a more objective basis for exposure assessment from chronic consumption. Intake of each type of food is modelled by a latent Gaussian variable, where intake is zero if the latent variable is below a threshold, and otherwise is a monotonically increasing function of the latent variable. Further, we use a Factor Analysis model to describe the association in intakes between different foods. This reduces the number of parameters to be estimated and aids interpretation. The method is illustrated using data from the *Dietary and Nutritional Survey of British Adults, 1986-1987*.

**Key words:** Chronic intake, Factor Analysis, Latent Gaussian model, Multivariate consumption.

---

\*Corresponding author: Professor C.A. Glasbey, email: [chris@bioss.ac.uk](mailto:chris@bioss.ac.uk), FAX +(44) 131 650 4901

# 1 Introduction

The perceived risk posed to health from eating too much and eating the wrong type of food is a topic rarely out of the news. It could be that a particular food is considered to be unhealthy, or maybe that the food is good for you in small amounts but becomes detrimental to health when more is consumed. The process of risk assessment can be divided into four steps (Codex Alimentarius Commission, 2004), the first being hazard identification, where known or potential adverse health effects of a given substance or nutrient are identified. In the second step, hazard characterization, a qualitative and quantitative evaluation takes place of the adverse effects associated with the substance. Following that, there is an exposure assessment, where the likely intake of the substance via the food is quantitatively evaluated. The final step is risk characterisation, that integrates information from the previous steps to characterize the risk, or probability of an adverse effect. In this study we present a model that can be used in step three, exposure assessment.

In risk assessment a distinction is made between acute risks, concerning health effects which arise from a single or short-term intake, and chronic risks, related to long-term intake. Safety levels are usually determined from diet trials performed on laboratory animals and observing the maximum level of a substance that can be consumed by the animal with no observable adverse effect on health. Two measures are derived from such studies. The ADI (Acceptable Daily Intake) is the amount of the substance that can be consumed daily over a lifetime without any known adverse effect on health. It is expressed in relation to bodyweight. The ARfD (Acute Reference Dose) is an estimate of the amount of the substance that can be consumed over a short period of time, usually one meal or one day, without any known harm to health.

In general, intakes of a chemical present in food cannot be measured directly, so in order to assess the exposure to the chemical it is essential to combine consumption data with residue concentration values. Consumption data are usually available from national surveys such as the *Dietary and Nutritional Survey of British Adults, 1986-1987* (UK Data Archive, 1991). Concentration data come from monitoring programmes run at several stages of the agro-food chain, and consist of measurements of residue concentrations present in samples of raw agricultural products. A typical feature of monitoring datasets is their sparsity: toxic substances are rarely found in food samples, either because they are not present or because the concentrations are below the limit of reporting (LOR). Other factors, such as food washing or processing, are likely to reduce or eliminate the residue in food. In a realistic risk assessment, a sensitivity analysis should be conducted to investigate the effects of processing factors, and of setting non-detected concentrations at either zero or at the LOR. Residue intakes are obtained from consumption data and concentration values by multiplying them and summing over the different food commodities (see, for example, Kroes et al., 2002).

A potential difficulty in exposure assessment arises from the fact that exposure to a substance may be caused from eating simultaneously several products that contain that substance. Therefore the total intake resulting from all relevant food products should

be modelled. Statistical modelling provides a framework for more objective exposure assessment. Until recently most attempts to model multivariate food consumption data have been through the use of non-parametric models, for example Ferrier et al. (2002). Parametric models for chronic exposure to substances are usually univariate (see, for example, Cullen and Frey, 1999), and models for intakes adopt a variance components approach (Slob, 1993; Nusser et al., 1996; Myles et al., 2003). For acute exposure there are also univariate models using extreme value theory (Tressou et al., 2004). Paulo et al. (2005) used parametric multivariate methods for modelling acute risks, but as far as we know parametric multivariate models have not yet been used in chronic exposure assessment.

There are many objectives for which a parametric multivariate model for consumption would be useful. In probabilistic exposure assessment we need a model to generate consumption patterns to be combined with concentration information. Exposure assessments typically relate to extreme situations, for which there will be few, if any, data. This is even more true in multivariate settings: it can easily occur that no person in the dataset consumed breakfast cereals, liver and chocolate in the same day, however there is no reason to believe this would never occur. Clearly, a parametric model can still generate such consumption patterns. It is also likely that for practical use we may want to look at subpopulations (pregnant women, babies, people living in a specific region, fish eaters, etc.) and that the number of consumers in the database would become very small for empirical sampling. Or we may want to model consumption as a function of e.g. age. In all these cases it is useful to have a parametric model.

In this study we propose a multivariate statistical model for individual consumption of multiple food types, to provide a more objective basis for exposure assessment from chronic consumption. We use data from the *Dietary and Nutritional Survey of British Adults, 1986-1987* to illustrate our model.

## 2 Model

Given a typical dataset of consumption with a large number of food types, it is highly unlikely that all individuals will have eaten foods from all the categories, hence the data matrix will typically contain many zeroes. This poses an immediate problem for the modelling. One approach would be to have a binary variable for each food type, taking the value 0/1 according to whether a particular individual has consumed that food type or not. Then, for non-zero consumptions, some distribution could describe the amount consumed. However it is not then straightforward to model the pairwise correlations between food types. A more elegant approach is to use an underlying multivariate Gaussian distribution. The idea is that for each food type, we create a thresholded Gaussian variable such that the part of the distribution below the threshold corresponds to zero consumption of that food type and non-zero consumptions are transformed to fit the part of the distribution above the threshold. So for each food type we have a single variable to account for both absence/presence and the amount consumed. Considering all food types together,

we have a latent multivariate Gaussian process. Therefore as well as considering the marginal distribution for each food type it is easy to incorporate the correlations between all pairs of food types.

In §2.1 we discuss the transformation to latent Gaussian variables, in §2.2 the estimation of the correlation matrix and in §2.3 a model to parameterise the full correlation matrix. Then, in §2.4 we consider model validation, and finally, in §2.5 we discuss potential applications.

## 2.1 Transformation to normality

Let  $y_{ij}$  denote the intake by individual  $i$  of food type  $j$ , for  $i = 1, \dots, n$  and  $j = 1, \dots, p$ . Consider the distribution of intakes of food type  $j$ . We assume that there exist realisations  $z_{ij}$  of a latent Gaussian variable, with zero mean, unit variance, such that

$$y_{ij} = \begin{cases} 0 & \text{if } z_{ij} < T_j \\ f_j(z_{ij}) & \text{otherwise.} \end{cases}$$

Here,  $f_j$  is a monotonic function which transforms the latent variable  $z_{ij}$  to  $y_{ij}$ , and  $T_j$  is a threshold such that intake is zero if the latent variable is below this threshold, and otherwise is a transformed value of the latent variable.

We use a quadratic power transformation for the inverse of  $f_j$ , a function of the form:

$$f_j^{-1}(y_{ij}) = \alpha_0 + \frac{1}{\alpha_1} \left( y_{ij}^\gamma + \alpha_2 y_{ij}^{2\gamma} \right).$$

This function has the advantage of being analytically invertible, so we can use it to get both a consumption value from the latent variable and vice versa.

We estimate  $f_j$  separately for each  $j = 1, \dots, p$ , by minimising the sum of squares of differences between observed and expected normal scores over all non-zero  $y_{ij}$ :

$$\sum_{i=n_{0j}+1}^n \left\{ \Phi^{-1} \left( \frac{i - 0.375}{n + 0.25} \right) - f_j^{-1}(y_{(i)j}) \right\}^2,$$

where  $n_{0j}$  denotes the number of zero values for the  $j$ th food type,  $\Phi^{-1}$  is the inverse standard normal cumulative distribution function, and  $y_{(i)j}$  denotes the  $i$ th smallest observation in the set  $\{y_{ij}, i = 1, \dots, n\}$ . To estimate  $T_j$  we apply  $\Phi^{-1}$  to the proportion of zero values for that food type, i.e.

$$\hat{T}_j = \Phi^{-1} \left( \frac{n_{0j}}{n} \right).$$

For illustration, we consider a dataset from the UK Data Archive (1991) *Dietary and Nutritional Survey of British Adults, 1986-1987* containing consumption data for 2197 adults. The data are food consumption over a 7 day trial period, for 51 main food types and 85 sub-types (see Appendix). We modelled the 51 main types, though we could equally well have worked with the 85 sub-groups. The four parameters in  $f_j$  estimated for each food type are displayed in Table 1, together with the proportion of zero consumptions.

Food type	Propn zeroes	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\gamma}$	$\hat{\alpha}_2$
1	0.30	-0.7716	40.59	0.7165	-0.001573
2	0.13	-1.3395	36.70	0.6712	-0.000882
3	0.48	-0.1344	144.36	0.8407	-0.000495
4	0.55	-0.0693	35.20	0.6930	-0.001932
5	0.60	-0.0147	44.87	0.7269	-0.001891
6	0.70	0.4435	124.91	0.9408	-0.000607
7	0.23	-1.1340	11.37	0.6342	-0.005066
8	0.23	-1.0189	42.73	0.7546	-0.001309
9	0.34	-0.6725	53.42	0.7479	-0.001117
10	0.12	-1.1079	136.01	0.7452	-0.000257
11	0.77	0.7478	224.51	0.6022	0.013344
12	0.82	0.9122	197.39	0.5987	0.011913
13	0.57	-0.5207	3.26	0.3934	-0.022145
14	0.19	-1.2296	19.19	0.7218	-0.002859
15	0.73	0.4653	182.92	0.8868	-0.000408
16	0.20	-1.2163	33.91	0.7857	-0.001676
17	0.40	-0.4456	9.56	0.5022	0.030412
18	0.67	0.4676	123.23	1.0137	-0.000687
19	0.81	0.8298	79.08	0.8973	-0.000849
20	0.79	0.2434	3.21	0.3174	0.054108
21	0.62	0.2308	25.46	0.7163	-0.001277
22	0.23	-1.1371	18.71	0.7156	-0.003040
23	0.24	-0.9514	61.05	0.7882	-0.000918
24	0.68	-0.0214	23.43	0.6709	-0.003650
25	0.58	-0.3363	23.98	0.7130	-0.002973
26	0.91	1.0073	150.93	0.9830	-0.000635
27	0.34	-0.7857	40.48	0.7625	-0.001647
28	0.76	0.5739	163.75	1.0247	-0.000570
29	0.72	0.1622	37.36	0.7511	-0.002221
30	0.47	-0.4548	36.46	0.8029	-0.001942
31	0.47	-0.4270	65.22	0.8159	-0.001049
32	0.55	-0.1102	29.60	0.6759	-0.002309
33	0.56	-0.2315	167.80	1.0327	-0.000464
34	0.66	0.2940	166.54	0.9750	-0.000491
35	0.65	0.1728	61.74	0.8522	-0.001324
36	0.20	-0.9996	36.81	0.7237	-0.001626
37	0.01	-2.5500	42.30	0.7463	-0.001058
38	0.14	-1.4622	37.81	0.7125	-0.001349
39	0.10	-1.6272	72.85	0.8079	-0.000729
40	0.20	-0.9638	71.59	0.7368	-0.000844
41	0.18	-1.2056	6.89	0.4363	0.015804
42	0.49	-0.2569	23.63	0.7716	-0.002907
43	0.79	0.6912	13.54	0.6316	-0.006034
44	0.46	-0.3340	22.50	0.7134	-0.003208
45	0.58	0.0279	182.35	0.8135	-0.000345
46	0.34	-0.8207	24.94	0.5350	-0.002411
47	0.75	0.2720	11.14	0.5421	-0.007485
48	0.62	0.1389	101.97	0.7447	-0.000749
49	0.53	0.0207	232.58	0.6630	-0.000292
50	0.05	-1.9130	12.31	0.5727	-0.004040
51	0.01	-3.0809	101.46	0.6744	-0.000377

Table 1: Estimated parameters for the transformation of the 51 food types to latent Gaussian variables, for data from *Dietary and Nutritional Survey of British Adults, 1986-1987* (UK Data Archive, 1991).

## 2.2 Estimation of correlation matrix

We model the joint distribution of latent variables for all food types by a multivariate normal distribution. So, for individual  $i$  the vector of latent values for the  $p$  food types, denoted  $z_i$ , is a realisation of

$$z_i \sim \text{MVN}(0, S).$$

Here  $S$  is a  $p \times p$  symmetric matrix with unit terms on the diagonal, so it is both the variance and correlation matrix for  $z_i$ . Each element  $S_{jl}$  may be estimated for  $j = 1, \dots, p-1, l = j+1, \dots, p$ , taking the approach of Allcroft and Glasbey (2003). We numerically maximise the pairwise marginal log-likelihood

$$\sum_{i=1}^n \log p(z_{ij}, z_{il}) \quad (1)$$

with respect to  $S_{jl}$ , where the pairwise probabilities,  $p$ , take one of four forms depending on whether neither, one or both of the consumptions were zero for a given individual:

$$p(z_{ij}, z_{il}) = \begin{cases} \Phi_2(\hat{T}_j, \hat{T}_l; S_{jl}) & \text{if } z_{ij} \leq \hat{T}_j \text{ and } z_{il} \leq \hat{T}_l \\ \phi(z_{ij}) \Phi\left(\frac{\hat{T}_l - S_{jl} z_{ij}}{\sqrt{1 - S_{jl}^2}}\right) & \text{if } z_{ij} > \hat{T}_j \text{ and } z_{il} \leq \hat{T}_l \\ \phi(z_{il}) \Phi\left(\frac{\hat{T}_j - S_{jl} z_{il}}{\sqrt{1 - S_{jl}^2}}\right) & \text{if } z_{ij} \leq \hat{T}_j \text{ and } z_{il} > \hat{T}_l \\ \phi_2(z_{ij}, z_{il}; S_{jl}) & \text{otherwise.} \end{cases}$$

Here  $\phi$  denotes the standard normal probability density function (pdf),  $\phi_2$  the standard two-dimensional normal pdf,  $\Phi$  the standard normal cumulative distribution function (cdf) and  $\Phi_2$  the standard two-dimensional normal cdf. In situations where there are no individuals for whom  $z_{ij} > \hat{T}_j$  and  $z_{il} > \hat{T}_l$ , (1) may be maximised by  $S_{jl} \rightarrow -1$ . This situation did not occur in our dataset, so estimation was not a problem. When it does occur it is necessary to estimate the parameters in the factor analysis model, described below in §2.3, directly from the data, rather than using the two-stage approach presented here.

Figure 1 shows the estimated correlation matrix, firstly with food types ordered as they appear in the list given in the Appendix and secondly in order of terms in the first eigenvector of  $S$ . This is convenient for display, as the first food types are positively correlated with each other, and the last types have the biggest negative correlation with the first ones. At the top of the order are foods such as skimmed milk, polyunsaturated margarine, brown bread, fruit and vegetables, which are all positively correlated with each other. At the opposite end are foods such as margarine, eggs and butter, which are negatively correlated with those at the top. Correlation coefficients ranged from  $-0.53$  (between white and wholemeal bread) to  $+0.41$  (between liqueurs and wine). Of the 1275 correlations, 39% were negative and only 3% lay outside the range  $\pm 0.25$ .

## 2.3 Factor Analysis model

The estimated correlation matrix contains  $p(p-1)/2$  terms (here 1275 for  $p = 51$ ). Rather than having such an unconstrained correlation matrix with so many free parameters we would like to impose some structure on it. One way to do this is to consider the following model for  $S$ :

$$S = \sum_{k=1}^K \beta_k \beta_k^T + \text{diag}(\sigma_1^2, \dots, \sigma_p^2) = BB^T + \Sigma. \quad (2)$$

Here,  $\beta_1, \dots, \beta_K$  are vectors of length  $p$ , with the number of them,  $K$ , is to be determined, and  $\sigma_j^2$ , termed ‘unique variances’, are such that  $S_{jj} = 1, j = 1, \dots, p$ . We denote by  $B$  the  $p \times K$  matrix whose columns are  $\beta_1, \dots, \beta_K$ , and by  $\Sigma$  the  $p \times p$  diagonal matrix with diagonal elements  $\sigma_1^2, \dots, \sigma_p^2$ .

The parametrisation of  $S$  given in (2) is equivalent to assuming that, for the  $i$ th individual, we have  $K$  latent factor,  $e_{i1}, \dots, e_{iK}$ , which are independent standard normal variates, and that  $z_i$  is specified by

$$z_i = \beta_1 e_{i1} + \dots + \beta_K e_{iK} + \epsilon_i,$$

where  $\epsilon_i$  is a  $p$ -vector, also of independent normal variates, but with  $\epsilon_{ij}$  having variance  $\sigma_j^2$ . The  $\beta$ 's can be interpreted as either a general appetite, or preferences for particular groups of food types, whereas the  $\sigma_j^2$  are specific variances for each food type. The model is the same as that used Factor Analysis (FA), and methods for estimating  $B$  and  $\Sigma$  have already been extensively researched (see, for example Krzanowski, 1988).

For a maximum likelihood approach we need to maximise the log likelihood, for fixed  $K$ :

$$\log L \propto -\log |BB^T + \Sigma| - \text{trace}(BB^T + \Sigma)^{-1}S.$$

However, note that in our censored case  $S$  is not the sample covariance matrix, so  $L$  is only an approximate likelihood.  $L$  is best maximised by an iterative procedure (Jöreskog, 1967). From an initial estimate of  $\Sigma$ , such as

$$\hat{\sigma}_j^2 = 1 - \max_{l \neq j} |\text{Corr}(y_{ij}, y_{il})|,$$

we alternate between the following two procedures, one being analytic and the other a numerical optimisation, until convergence is achieved:

- Set  $\Sigma = \hat{\Sigma}$ . Perform an eigenanalysis of  $\Sigma^{-1/2}S\Sigma^{-1/2}$  and let  $\Omega$  be the  $p \times q$  matrix containing the  $q$  largest eigenvectors and  $\Theta$  be the  $q \times q$  diagonal matrix with the  $q$  largest eigenvalues on the diagonal. Then

$$\hat{B} = \Sigma^{1/2}\Omega(\Theta - I)^{1/2}.$$

- Set  $B = \hat{B}$  and minimise

$$\log |BB^T + \Sigma| + \text{trace}(BB^T + \Sigma)^{-1}S$$

with respect to  $\Sigma$  to get a revised estimate of  $\hat{\Sigma}$ .

In order to decide how many factors are needed, a common approach is to start from  $K = 1$ , i.e. assuming that a single latent factor is sufficient, and progressively increment  $K$  until a satisfactory fit is achieved. With full data, a likelihood ratio test can be used to determine whether the model is adequate. Here, with only partial data, the degrees of freedom will not be correct and so this cannot be used without modification. Also, with large datasets, simple models are almost always rejected by statistical significance tests, although more complex models may be of no greater scientific usefulness. Therefore, we have limited consideration to small values of  $K$  and not developed formal statistical tests.

Figure 2 shows the fitted correlation matrix from models with  $K = 1$  and  $K = 2$ . Using a single factor, the correlation between the first few food types and the last few, and the negative correlation between the two groups has been captured. As successive factors are added, more patterns in the matrix are captured. A single factor explains 83% of the variability in the correlation matrix and two factors explain 88%.

We can also plot the scores of each food type on the first two factors and display them on a two-dimensional plot. This is shown in Figure 3 and illustrates that similar foods appear close together on the plot. This sort of plot can also be useful for investigating subpopulations, as scores can be calculated for each factor for different subpopulations and then plotted to see if certain subpopulations eat, on average, more or less of certain food groups.

## 2.4 Model validation

Figure 4 shows histograms of untransformed and transformed non-zero consumption values for white bread (food type 2) and low-fat spread (19), which had 13% and 81% zero values, respectively. The linearity of the normal probability plots shows that we have examples of the transformation being a good fit for both food types that are eaten a lot and for those eaten by relatively few individuals.

We also need to check the assumption of multivariate normality. We have the problem that zero consumption corresponds to a censored value on the Gaussian scale, and so any approaches to model validation need to take this into account. We can take a simulation approach and for censored values, simulate a value in the part of the distribution below the threshold, and thereby create a realisation of the latent variable.

If the assumption of full multivariate normality holds, then any linear combination of the latent variables should be univariate normal. Although examination of the normality of such linear combinations does not confirm the multivariate normality hypothesis, a good fit would show consistency with the hypothesis and would be unlikely if the hypothesis did not hold.

As there are such linear combinations of the variables within the factor analysis, it seems natural here to examine the univariate normality of the factors. Figure 5 shows normal probability plots for the first two factors. These were formed as described above and so

when a particular individual has not consumed a particular food, a value for the latent variable was simulated from the relevant part of the distribution. Strictly, we should also take account of correlations between latent variables when simulating censored values. However, this is far more difficult to achieve so, for simplicity, here we have simulated each latent variable independently. The two plots can be seen to be roughly linear for the main part of the plots; some lack of fit can be seen at the ends, however this involves only a very small proportion of the sample values, e.g. less than 1%. Therefore these plots are largely consistent with the assumption of normality of the latent factors.

## 2.5 Potential applications

The most obvious types of question that this model equips us to answer are what proportions of the population consume above or below given amounts of certain food types or nutrients. For instance it might be thought to be bad to have an average sugar intake of more than Xg per day. But then the risk of a given disease might be considered much worse if this is combined with an average daily intake of less than Yg of fibre. Consumption of any pair of food types or nutrients is not independent and therefore any estimate of the proportion of people who eat both more than Xg of sugar and less than Yg of fibre should take into account the correlation between the two. Working with the multivariate Gaussian model enables these correlations to be easily included and the idea can be extended to more than two food groups. In a Gaussian framework, any marginal, joint or conditional distribution can be derived very easily and are univariate or multivariate Gaussian themselves. Hence a set of relatively straightforward questions that can be answered are of the form:

- What proportion of people consume above a given amount of a particular food type or nutrient?
- What proportion of people consume above given amounts of two or more food types?
- What proportion of people consume above a given amount of a particular food type, given they consume above a given amount of another food type?

The first of these is just about the marginal distribution of a single food type. The second and third involve joint and conditional distributions respectively, and for these it is necessary to take into account the correlation between the two (or more) variables and so it is for this sort of question that a model such as the one developed is useful. The model also enables us to put standard errors or confidence limits on the estimates – this can be done via simulation. Some of these quantities could be calculated empirically, however as we have already pointed out, if there happen to be no people in the dataset who have eaten particular combinations of foods, then this would be incorrectly estimated at zero, whereas a parametric model would provide a non-zero estimate. Also, for concentrating on certain subpopulations where even less data were available this would be even more of a problem.

### 3 Discussion

Perceived risks from food are of increasing concern. The purpose of the present paper is to provide a statistical framework for objective exposure assessment. We highlight the potential of the proposed method and describe some of the questions that it may be used to address. We have not attempted to thoroughly analyse the dataset described, though this is possible, as socio-economic and physiological information on the individuals was also collected.

In terms of methodology, we have shown that multivariate consumption data may be successfully modelled by a latent multivariate Gaussian model. Quadratic power functions were used to marginally transform consumption for each individual food type to thresholded Gaussian distributions. The full correlation matrix was estimated for the underlying Gaussian process. Then, in order to achieve a more parsimonious model, a further set of latent variables was introduced, allowing the correlation matrix to be modelled by an approach equivalent to factor analysis. Confirmation that the factors were approximately univariate normal allows us to conclude that the assumption of multivariate normality was reasonable. The main benefits of the whole latent Gaussian approach are firstly that binary consumption/non-consumption and the amount consumed can be modelled simultaneously using a single variable, and secondly that the correlation between foodstuffs is also modelled.

Correlated consumptions were observed and modelled in a published study of acute exposures (Paulo et al., 2005). That study found that consumptions of some commodities were negatively correlated and others positively correlated. In general, dependencies between consumption of different products are rather complex, as they are usually the consequence of there being several possible ways to combine food commodities. Correlations are an imperfect way to summarise these dependencies, but are still far superior to univariate methods.

We have modelled food types directly, using the 51-level categorisation of main types as shown in the Appendix. An alternative would be to model nutrients, which are also available in the dataset. In either case the methodology is the same, but the types of questions to be answered will determine which approach is more appropriate. In the current dataset, the nutrient data are obtained deterministically from the food type data, by assuming that a given food type is composed of particular proportions of nutrients. Therefore the nutrients are simply a linear transformation of the food types. However because the transformation between observations and the latent Gaussian variables is non-linear, identical results will not be obtained working directly with nutrients as opposed to with food types.

The data we have used for illustration are weekly averages for each individual and therefore we have the problem that to estimate typical daily intake, if we work with the 7-day total or average, this is only an approximation to typical daily intake, as it is contaminated by within-person between-day variation. This problem has been addressed in several ways (e.g. see Slob, 1993; Nusser et al., 1996; Myles et al., 2003), all involving hierarchical

modelling. It would be of interest to incorporate this into our proposed method.

We have used a factor analysis (FA) approach to simplify the correlation matrix using fewer parameters. This had the advantage of a plausible interpretation in terms of latent factors. Another approach would have been principal components analysis (PCA). Given a dataset, PCA and FA often give similar results, leading to the understandable general confusion as to what the differences between these two approaches are. PCA is generally recommended when the objective is simply to empirically reduce the information in many variables into a smaller set of variables, whereas Factor Analysis identifies latent factor that contribute to the common variance in the set of variables. So the first important difference is that Factor Analysis assumes an underlying model, PCA does not. A further difference is that Factor Analysis concentrates on the variance that is common among the variables, it excludes the ‘unique variance’ terms  $\sigma_j^2$  from the analysis; PCA does not differentiate between common and unique variance – in effect it sets the unique variances to zero. Therefore the primary aim of FA is to explain the association between the variables, whereas PCA concentrates on the variances. A further useful distinction to note is that the results of FA are invariant under changes of scale, whereas PCA gives different results depending on whether it is based on the covariance or correlation matrix. See Krzanowski (1988) for more discussion. In many cases, if the unique variances as estimated in the FA are small, similar results are obtained from the two methods.

We have displayed the Factor Analysis results from one and two factor models and a visual assessment showed that the main features of the correlation structure are captured even with such few factors. In general though it would be desirable to have a formal method for choosing the number of factors needed. A suggestion would be to perform some sort of cross-validation (leave-one-out, leave-more-out, or just simple a split of data in training and test set), and then optimise the predictive performance of the model. However the results from this sort of method are generally dependent on the size of the dataset and with such a large dataset as that used here you would generally end up using more factors than you would ideally want. However some methodology on this would be useful.

## Acknowledgements

We thank the UK Data Archive for making such data available to researchers. We also thank Hilko van der Voet for helpful comments. BioSS is supported by the Scottish Executive Environment and Rural Affairs Department.

## References

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

- Codex Alimentarius Commission (2004). *Joint FAO/WHO Food Standards Programme: Procedural Manual*. Food and Agriculture Organization of the United Nations and World Health Organization, Rome, thirteenth edition.
- Cullen, A. C. and Frey, H. (1999). *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. Plenum Press, New York.
- Ferrier, H., Nieuwenhuijsen, M., Boobis, A., and Elliott, P. (2002). Current knowledge and recent developments in consumer exposure assessment of pesticides: a UK perspective. *Food Additives and Contaminants*, 19:837–852.
- Jöreskog, K. G. (1967). Some contributions to maximum likelihood Factor Analysis. *Psychometrika*, 32:443–482.
- Kroes, R., Muller, D., Lambe, J., Loewik, M. R. H., van Klaveren, J., Kleiner, J., Massey, R., Mayer, S., Urieta, I., Verger, P., and Visconti, A. (2002). Assessment of intake from the diet. *Food and Chemical Toxicology*, 40:327–385.
- Krzanowski, W. J. (1988). *Principles of Multivariate Analysis: A User's Perspective*. Oxford University Press, Oxford.
- Myles, J., Price, G., Hunter, N., Day, M., and Duffy, S. (2003). A potentially useful distribution model for dietary intake data. *Public Health Nutrition*, 6:513–519.
- Nusser, S., Carriquiry, A., Dodd, K., and Fuller, W. (1996). A semi-parametric transformations approach to estimating usual daily intake distributions. *Journal of the American Statistical Association*, 91:1440–1449.
- Paulo, M., van der Voet, H., Jansen, M. J. W., ter Braak, C. J. F., and van Klaveren, J. D. (2005). Risk assessment of dietary exposure to pesticides using a bayesian method. *Pest Management Science*, 61:759–766.
- Slob, W. (1993). Modelling long-term exposure of the whole population to chemicals in food. *Risk Analysis*, 13:525–530.
- Tressou, J., Crépet, A., Bertail, P., Feinberg, M., and Leblanc, J. (2004). Probabilistic exposure assessment to food chemicals based on extreme value theory. Application to heavy metals from fish and sea products. *Food and Chemical Toxicology*, 42:1349–1358.
- UK Data Archive (1991). Office of Population Censuses and Surveys Social Survey Division, Ministry of Agriculture, Fisheries and Food, Department of Health, Dietary and Nutritional Survey of British Adults, 1986-1987 [computer file]. Colchester, Essex: UK Data Archive [distributor], 18 September 1991. SN: 2836.

## Appendix - List of main food types and sub-types

01A PASTA; 01B RICE; 01R OTHER CEREALS  
02R WHITE BREAD  
03R WHOLEMEAL BREAD  
04R OTHER BREAD  
05R HI FIBRE BKFAST CEREALS  
06R OTHER BKFAST CEREALS  
07R BISCUITS  
08A FRUIT PIES; 08R BUNS, CAKES, PASTRIES  
09A MILK PUDDINGS; 09B ICE CREAM; 09C OTHER PUDDINGS; 09D SPONGE PUDDINGS  
10R WHOLE MILK  
11R SEMI-SKIMMED MILK  
12R SKIMMED MILK  
13R OTHER MILK  
14A COTTAGE CHEESE; 14R OTHER CHEESE  
15R YOGHURT  
16R EGGS  
17R BUTTER  
18R POLYUNSAT MARGARINE  
19R LOW FAT SPREAD  
20R BLOCK MARGARINE  
21A SOFT MARGARINE; 21B YELLOW SPREADS; 21R OTHER SPREADS/MARG  
22R BACON & HAM  
23R BEEF & VEAL  
24R LAMB  
25R PORK  
26R COATED CHICKEN  
27R CHICKEN/TURKEY DISHES  
28R LIVER & PRODUCTS  
29R BURGERS/KEBABS  
30R SAUSAGES  
31R MEAT PIES & PASTRIES  
32R OTHER MEAT PRODUCTS  
33R FRIED WHITE FISH  
34A OTHER WHITE FISH; 34B SHELLFISH  
35R OILY FISH  
36A CARROTS; 36B OTHER SALAD VEG; 36C TOMATOES  
37A PEAS; 37B GREEN BEANS; 37C BAKED BEANS; 37D LEAFY GREEN VEG; 37E CARROTS;  
37F FRESH TOMATOES; 37R OTHER VEG  
38A POTATO CHIPS; 38B FRIED/ROAST POTS.; 38R OTHER FRIED POTATO PRODS  
39R OTHER POTATOES  
40A APPLES & PEARS; 40B ORANGES etc; 40C BANANAS; 40D FRUIT CANNED IN JUICE;  
40E FRUIT CANNED IN SYRUP; 40F UNSALTED NUTS & MIXES; 40R OTHER FRUIT & NUTS  
41A SUGAR; 41B PRESERVES; 41R OTHER SUGAR PRODS  
42R SAVOURY SNACKS  
43R SUGAR CONFECTIONERY  
44R CHOCOLATE CONFECTIONERY  
45R FRUIT JUICE  
46A SOFT DRINKS (NOT DIET); 46B DIET SOFT DRINKS  
47A LIQUEURS; 47B SPIRITS  
48A WINE; 48B FORTIFIED WINE  
49A BEERS; 49R CIDER, PERRY etc  
50R MISCELLANEOUS  
51A COFFEE; 51B TEA; 51R WATER

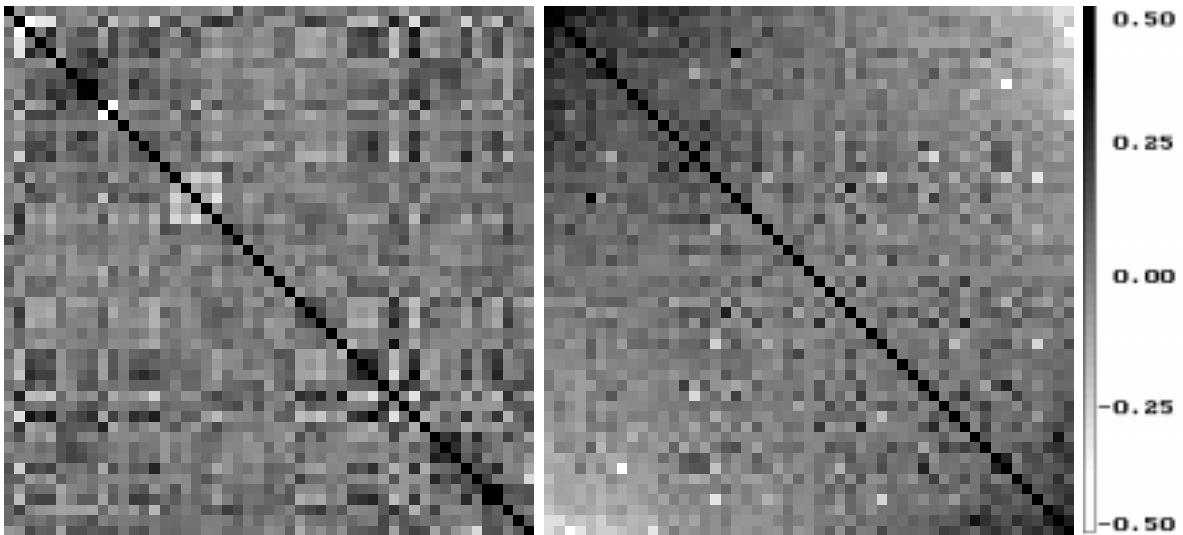


Figure 1: Estimated correlation matrix, with food items: a) ordered as listed in the Appendix; b) ordered by terms in the first eigenvector. Correlation coefficients are shown as shades of grey, with higher values displayed as darker shades in accordance with the key on the right.

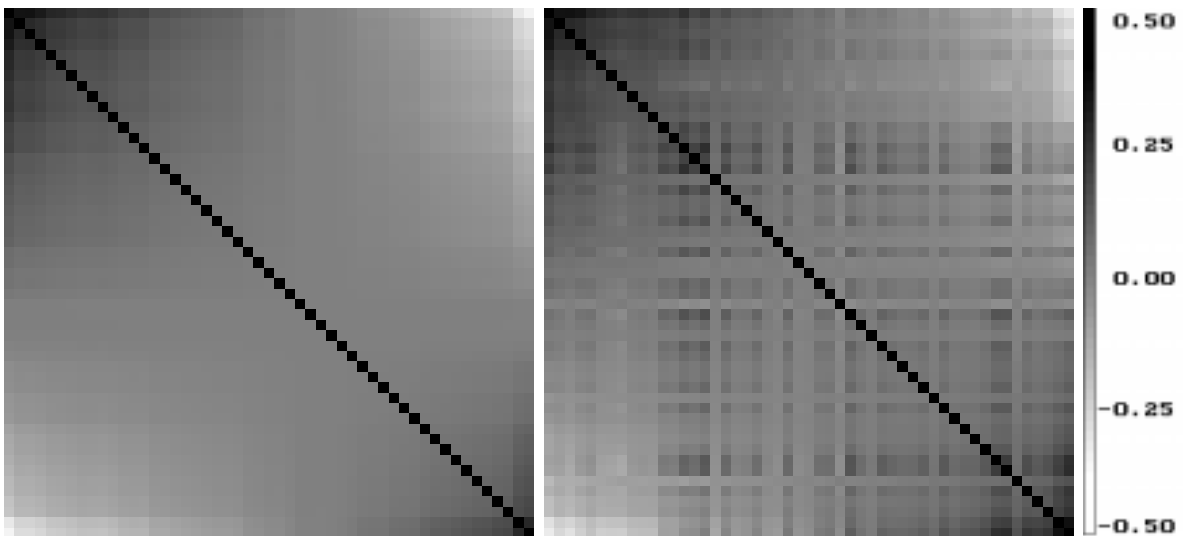


Figure 2: Fitted correlation matrix using Factor Analysis: a) with  $K = 1$ ; b) with  $K = 2$ , with correlation coefficients denoted by shades of grey according to scale on right.

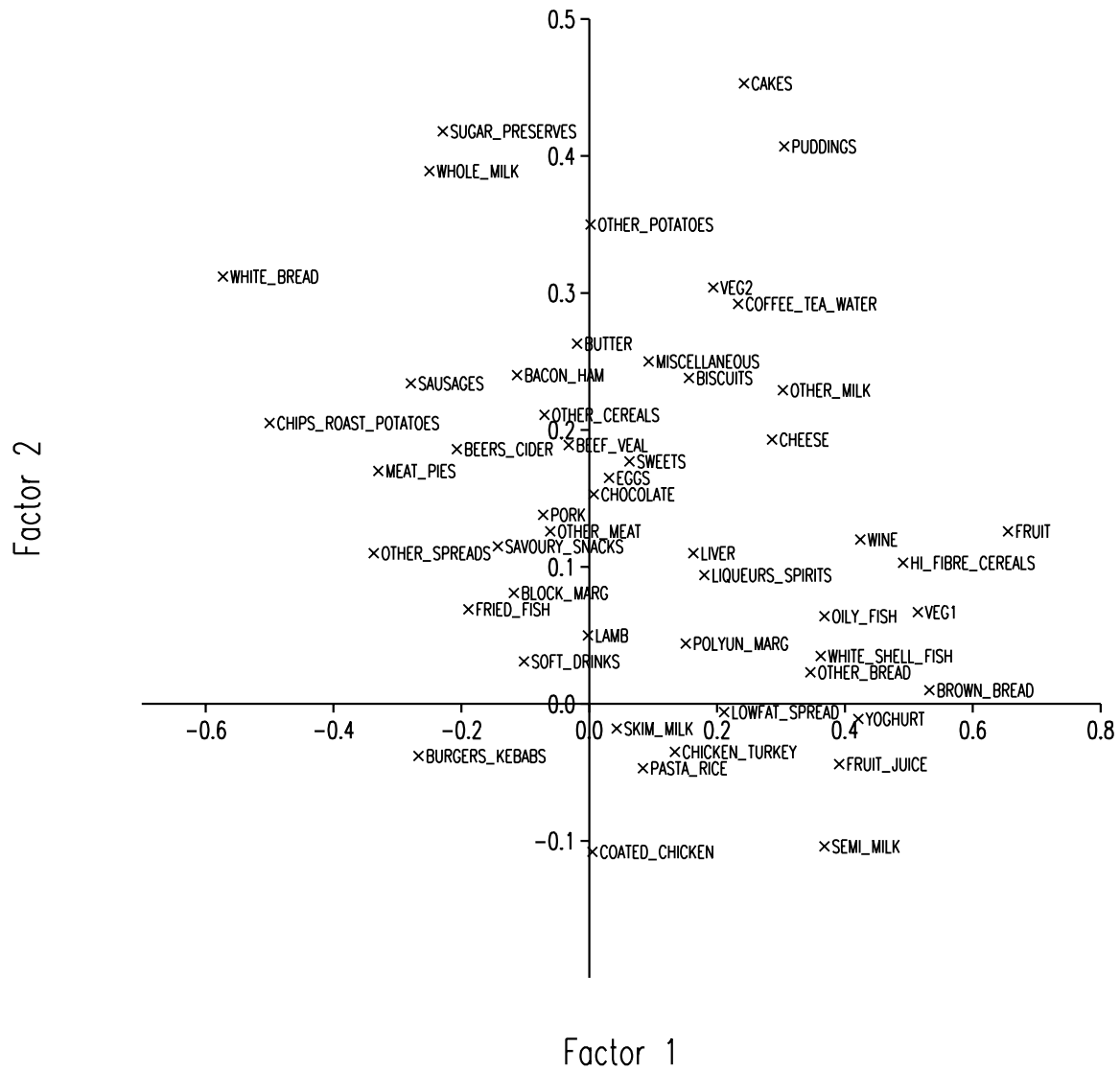


Figure 3: Factor scores from Factor Analysis for the 51 food types on the first and second factors.

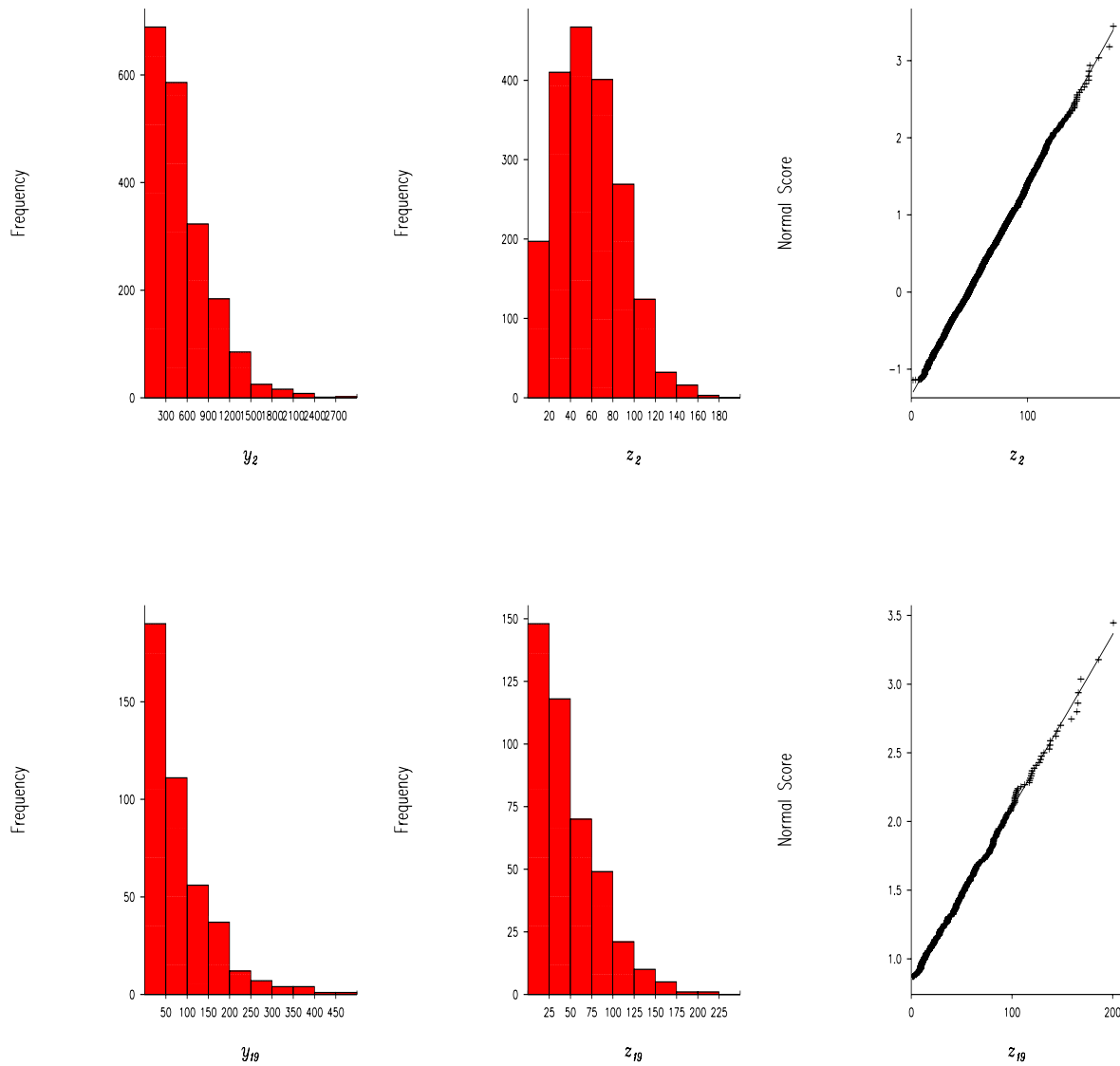


Figure 4: For white bread (top) and low-fat spread (bottom), distribution of non-zero consumption values: 1) histogram of untransformed values; 2) histogram of quadratic power-transformed values; 3) normal probability plot

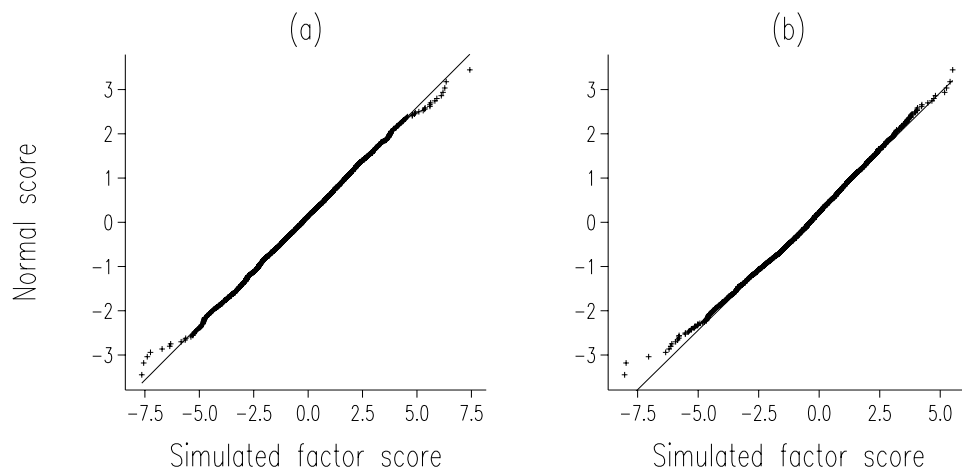


Figure 5: Normal probability plot for (a) the first factor, (b) the second factor, based on 2197 individuals, simulating values of latent variables in cases where none of that food type was consumed by that individual