Statistical methods for environmental risk assessment

Compositional data module

Written by: Adam Butler, Stijn Bierman, and Glenn Marion

Biomathematics and Statistics Scotland, The University of Edinburgh, James Clerk
Maxwell Building, The King's Buildings, Edinburgh EH9 3JZ.
(htpp://www.bioss.ac.uk/staff.html).

July 2005

This work has been part-funded by the **ALARM** project.

The ALARM project is an international and cross-disciplinary research project, funded under the 6th Framework Programme of the European Union. BioSS is a scientific partner within ALARM (http://www.bioss.ac.uk/alarm/); we contribute to the project both through input to specific research modules, and through a broader "cross-cutting" role. Our work within ALARM involves methodological research, the provision of statistical support, and the development of online training materials.

This pdf file is a text version of an online course, which can be found at:

http://www.bioss.ac.uk/~adam/alarm_training/com0.shtml

Aims of this module

- Describe the common features of **compositional data**, and outline why such datasets cannot be analysed using standard techniques
- Introduce the **log-ratio approach**, explaining the advantages (and limitations) of this approach
- Discuss issues of **statistical modelling & inference**, and outline currently available software for implementing these methods
- Introduce a **conditionally autoregressive model** for the analysis of compositional data on regular grids
- Present **case studies** relevant to the analysis of ecological datasets using compositional methods

1 What are compositional data?

'*Composition deals with the bits and pieces that make up things*' (Wikipedia definition).
'*Composition: the constitution of something made up from different elements*' (Oxford dictionary definition)

Compositional data refer to proportions, or fractions, of a whole. Compositional data arise in many different disciplines, and under many different guises, but share some important common characteristics. Most importantly, compositional data contain information on the relative frequencies with which the different components occur, but can tell us nothing about the actual (absolute) frequencies associated with the different components.

Geologists, for example, are often interested in describing and understanding the mineral composition of rock samples. It may well be scientifically interesting to know that 70% of a particular rock sample consists of quartz (compositional data), but it would be of no conceivable interest to know that a sample of unknown size contained 5g of quartz (data on absolute frequencies). Within biology, dietary data are often compositional - for example, seabird diet studies typically record the relative proportions of two or more different prey species found within the regurgitated stomach contents of individual birds. Again, scientists are really only interested in relative frequencies, because the absolute amount of regurgitated material depends upon a whole lot of factors which are completely unrelated to the diet of the seabird - e.g. the size of the bird, whether the bird regurgitates all of the stomach contents, whether the experimenter collects all of this material etc. etc. etc...

1.1 Examples

Here, we describe a number of example datasets that we will use within this module. The raw data and further details of these data sets have been made available at: http://www.bioss.ac.uk/~adam/alarm_training/datasets.shtml .

As already mentioned, compositional data are ubiquitous in disciplines such as geology, petrology and sedimentology. The 'skye.dat' dataset records the mineral composition of 23 rock samples from the Isle of Skye, whilst the 'arctic.dat' data records the sand/silt/clay composition of sediments in an Arctic lake.

Compositional data frequently arise in the areas of demography and population genetics. The classic 'hair.dat' data record the relative frequencies of different hair colours within the old Scottish counties, whilst the 'gaelic.dat' data record the proportion of Gaelic language speakers on the Scottish islands.

Compositional data on patterns of activity, spending or behaviour often arise in disciplines such as psychology and marketing. The 'leave.dat' data record activity patterns of a statistician over a six month period, whilst the 'nurses.dat' data record levels of glove use by nurses in the emergency ward of a US hopsital.

We use simulated data to demonstrate some of the more subtle aspects of compositional data analysis. The dataset 'bimodal.dat' illustrates the importance of testing the assumption of log-ratio normality, whilst the 'flower.dat' data are used to demonstrate a conditional autoregressive model for spatially referenced compositional data.

1.2 Problems with standard methods

Compositional data are subject to a unit sum constraint - i.e. because they are proportions they must sum to one. This sum constraint in turn imposes some unpleasant constraints upon the variance-covariance matrix of $\boldsymbol{X}$, and so, at a stroke, invalidates most standard statistical approaches - including techniques based on regression and multivariate analysis - which rely on an assumption of multivariate normality. It is important to appreciate that correlations between proportions are difficult (and, for large D, effectively impossible) to interpret in any meaningful way, not least because uncorrelated proportions are not necessarily independent.

The approach which we will outline in this course involves breaking the sum constraint - i.e. using a transformation of the data to remove the constraint, and then applying standard statistical techniques to the transformed data. The approach is analogous to the modelling of binary data using a generalised linear model with logistic link function - i.e. logistic regression (McCullagh & Nelder, 1989) - and will actually turn out to be quite closely related to that approach.

1.3 Some basic principles

Aitchison proposes some fundamental principles of compositional data analysis, and suggests that any reasonable statistical procedure for analysing compositional data should adhere to these principles.

Scale invariance: statistical inferences about compositional data should not depend upon the scale used. For example, we should obtain exactly the same results if we analyse percentages (of 100) as we would if we analysed proportions (of 1).

Subcompositional coherence: statistical inferences about a particular subset of components should depend only upon data about components within that subset. As an example, consider the hair colour data, and suppose that we are interested in comparing the relative frequency of fair and red headed boys. Subcompositional coherence requires that the results of the comparison are the same (a) if we have data on all five hair colours, but concentrate only upon a comparison of these two or (b) in we had only had data on fair and red headed boys. The relative proportions of fair and red headed boys constitute a subcomposition of the full dataset on all five hair colours, hence the name "subcompositional coherence".

Permutation invariance: statistical inferences should not depend upon the ordering (or labelling) of the components. It should not matter, for example, which component we choose to be the "first" and which component we choose to be the "last".

1.4 The sample space

When building a statistical model, one needs to define a sample space: a convenient reference space within which experimental outcomes can be unambiguously recorded. It is generally - although not quite universally - agreed that the appropriate sample space for compositional data is the **standard simplex** (also called the "unit simplex").

The standard *d*-simplex is defined to be the set

$$\left\{ (X^1, \ldots, X^D) \in \mathbb{R}^D : \sum_{j=1}^{D} X^j = 1, X^j > 0 \text{ for } j = 1, \ldots, D \right\}$$

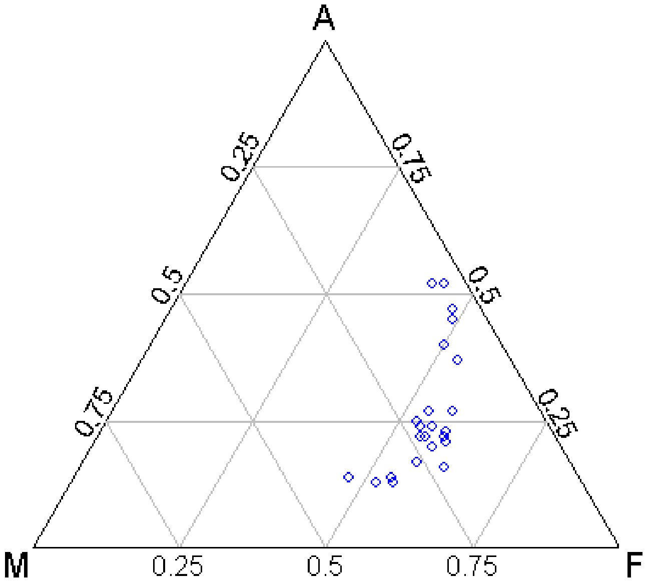
where *D* = d + 1.

Note from this definition that a composition of *D* parts actually lies within a (*D* - 1)-dimensional set. This make sense - if we know the proportions of *D* - 1 components, then (since the proportions must sum to one) we will also know the proportion associated with the final component. In particular, in the case of just two components (*D* = 2) it is obvious that if we know the proportion associated with one component then we also know the proportion associated with the other component, so we really only have information on one dimension (i.e. *d* = 1).

1.5 Graphical representations

Ternary diagrams

We can represent a 3-part composition (*D*=3) using a 2-dimensional plot known as a **ternary diagram**. Ternary diagrams are routinely used in geology & petrology, and are widely used elsewhere.
The ternary diagram is an equilatoral triangle, whose vertices represent the three elements of the composition. Data points which lie close to a vertex have high proportions of the element represented by that vertex, and data points which actually lie at the vertex have zero proportions of the other two elements. Data points lying in the centre of the triangle have equal proportions of all three elements.
We construct a ternary diagram for the skye lavas dataset:

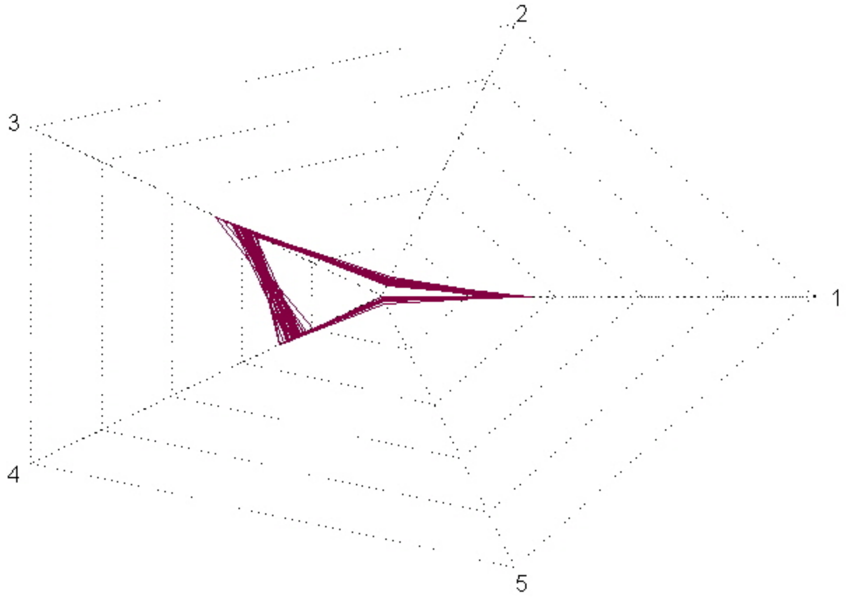


We see that proportions of **F** are generally higher than those of **A** or **M**. The compositions of the rock samples show a moderate level of variability (as indicated by the scatter of points within the triangle), with the proportion of **F** varying from around 40% to around 65%.

We provide an interactive environment for creating ternary plots of compositional data, using the BioSS analytic server at:
http://www.bioss.ac.uk/~adam/alarm_training/com15.shtml

Spiderplots

Higher-part compositions (*D*>3) cannot be represented so easily on a graph. **Spiderplots** provide one kind of graphical representation, but are not always easy to interpret. Spiderplots consist of *D* spokes, representing the *D* elements of the composition. A polygon is plotted for each observation (replicate) of the composition; the length along each spoke is determined by the proportion of the corresponding element in that observation, and the polygon is then formed by "joining" up the spokes.
We construct a spiderplot for the hair colour data:



The key to the spokes is: 1=Fair, 2=Red, 3=Medium, 4=Dark, 5=Jet black.

We see that fair, medium & dark hair are fairly common in all of the counties, whilst jet black and red are quite rare. The proportions show a very low level of variability between counties (indicated by the fact that the red polygons are all quite close to each other), suggesting that hair colour does not strongly depend upon location within Scotland.

We provide an interactive environment for creating spiderplots of compositional data, again using the BioSS analytic server at:
http://www.bioss.ac.uk/~adam/alarm_training/com15.shtml

1.6 Limits to interpretability

Compositional data only contain partial information - they provide information only about the relative values of components - so there are fundamental limitations to what we can learn through the analysis of a compositional dataset.

As a simple example, suppose that John keeps bananas, oranges & apples in his fruit basket. John sometimes eats some of the fruit in the baset, and sometimes he buys some more. All we know, however, is that at the start of a particular week he has 1/3 apples, 1/3 bananas and 1/3 oranges in the basket, but that by the end of the week has has 1/6 apples, 1/6 bananas and 2/3 oranges.

What can we conclude from this information ? Well, actually very little -

- we cannot conclude that John has bought any fruit, or that John has eaten any fruit;
- we cannot conclude that the overall amount of fruit in the basket has gone up, gone down, or stayed the same;
- and we cannot conclude that the number of oranges has gone up, gone down, or stayed the same.

We can only conclude that the proportion of oranges is higher at the end of the week than at the start, and that the proportions of apples and bananas are correspondingly lower.

2 Log-ratios of proportions

We have seen that compositional data are most definitely *non*-normally distributed - since they must always lie between zero and one, and they must sum to one - and that analyses of compositional data based on an assumption of normality can often lead to wildly missleading inferences. Fortunately, however, we can usually transform compositional data onto a scale where they *are* normal - and, in addition, we can do this using a extraordinarily simple transformation.

At this stage, we introduce some notation:

We consider the relative proportions of D components,

$$\mathbf{X} = (X^1, \ldots, X^D),$$

which must, by definition, sum to one:

$$\sum_{j=1}^{D} X^j = 1.$$

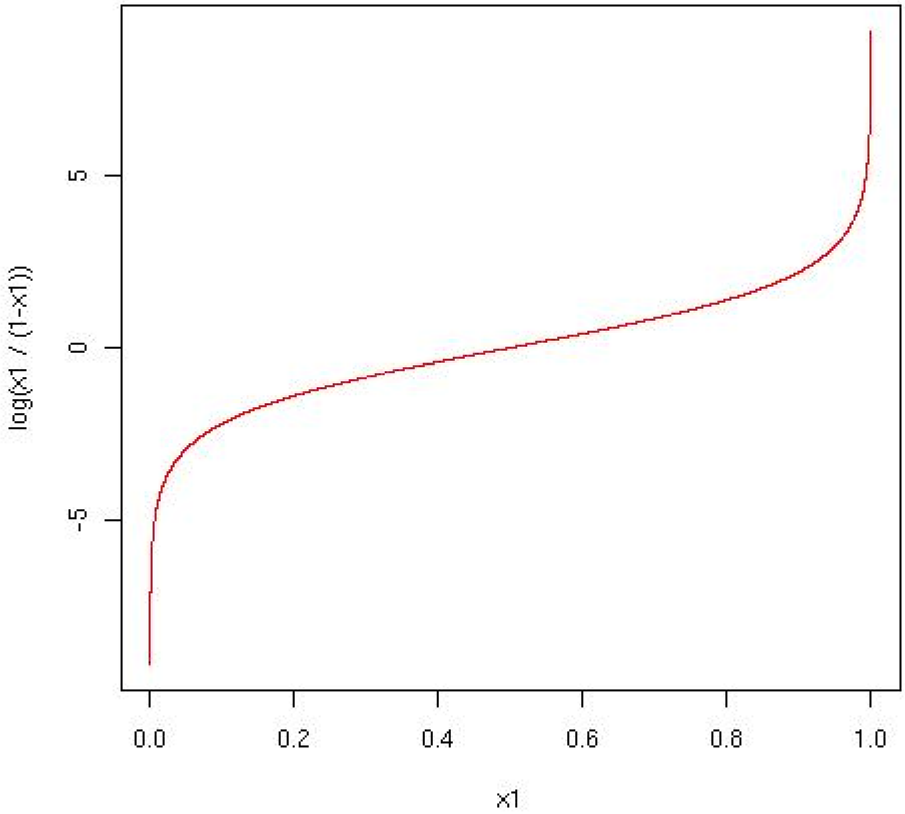
We will assume henceforth that the data contain no zero proportions, so that $X^j > 0$ for all $j = 1,\ldots,D$. We will to the thorny issue of zero proportions at a much later stage.

2.1 A simple case: two components

Begin by considering the simple situation where we have just two components, so that $D = 2$. In his seminal book, Aitchison (1986) suggests that the appropriate model for X^1 and X^2 will generally be to assume that the log-ratio, $\log(X^2 / X^1)$, has a normal distribution.

Why should we use log-ratios at all ?

Within statistics, transformations are usually used to remove constraints. For example, the log transformation is used to convert a variable which must be positive into a variable which can take any real (positive or negative) value. A simple graph:



shows that, even though X^1 must lie between zero and one, the log-ratio

$$\log\left(\frac{X^2}{X^1}\right) = \log\left(\frac{1 - X^1}{X^1}\right)$$

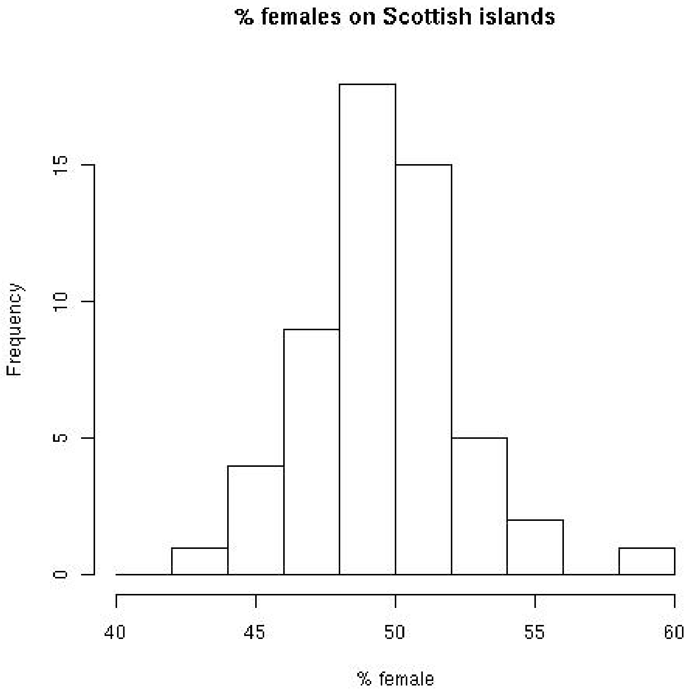
can take any real value. We therefore see that the log-ratio transformation is effective in removing the constraints on $\boldsymbol{X}$.

Why should the log-ratio be normally distributed ?

We have shown that the log-ratio log(X^2 / X^1) can take any real value. This is an essential prerequisite for normality, but it certainly does not demonstrate that the log-ratio *is* normally distributed. Aitchison shows that the standard mathematical arguments for assuming normality in statistics - such as the central limit theorem - also apply to the log-ratios of compositional data. We must still, however, check the validity of the normality assumption for the dataset at hand.

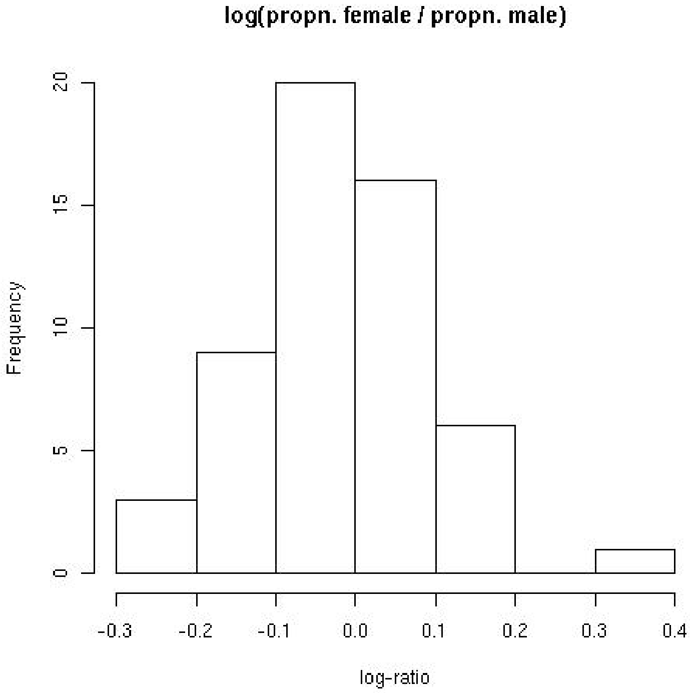
A worked example (using R): gender on remote islands

```
## (Just paste the following code into the R command line...)
## We analyse the gender data,
## which record the percentage of the population who are female for each of 55 Scottish islands.
## We load the data,
gender <- read.table("http://www.bioss.ac.uk/~adam/alarm_training/data/gender.dat",header=TRUE)
## and plot a histogram:
hist(gender$PcFemale,breaks=seq(40,60,2),xlab="% female",main="% females on Scottish islands"):
```



```
## The percentages range from 43% (Eriskay) to 59% (Iona).
## Now we convert from the percentage to the proportion scale...
proportion <- gender$PcFemale / 100
## ...calculate the log-ratios...
```

```
logratio <- log(proportion / (1 - proportion))
## ...and plot another histogram:
hist(logratio,xlab="log-ratio",main="log(propn. female / propn. male)")
```



```
## The log-ratios look fairly close to being normally distributed,
## except for one possible outlier (Iona) which has a particularily high proportion of females.
## Note that Iona is the site of an all-female religious community...
```

An aside: why not use the other log-ratio instead ?

There are two possible log-ratios which we could have modelled, so what would have happened if we had used the "other" log-ratio, log(X^1 / X^2), instead ? Well, using elementary properties of logs we can see that

$$\log\left(\frac{X^1}{X^2}\right) = \log\left(X^1\right) - \log\left(X^2\right) = -\left[\log\left(X^2\right) - \log\left(X^1\right)\right] = -\log\left(\frac{X^2}{X^1}\right)$$

So the two log-ratios will always have identical values, but they will have different sign. We therefore only need to model *one* of the two log-ratios, and it does not matter which one...

2.2 A connection with logistic regression

It is worth noting that log-ratios crop up elsewhere in statistics, particularily in the field of logistic regression.

In logistic regression we have a binary (0/1) outcome which we assume to have arisen from a Bernoulli distribution, and we wish to model the probability *p* of a positive outcome (1) as a function of explanatory variables. We transform the probability onto the real line using the logit link function, log[*p* / (1 - *p*)], which is simply the log-ratio of *p* to (1 - *p*).

We can see that logistic regression has a lot in common with our model for compositional data when *D* = 2. The key difference is that for compositional data we actually observe the proportion *p*, whereas in logistic regression this proportion is an unobserved quantity which represents the probability associated with an observed binary event.

2.3 More than two components

We can adopt a similar approach when the number of components *D* is greater than two.

Multivariate normal model

Aitchison demonstrates that it will frequently be reasonable to assume that the *d = D - 1* log-ratios

$$\log\left(\frac{X^2}{X^1}\right), \ldots, \log\left(\frac{X^D}{X^1}\right)$$

follow a multivariate normal distribution with mean vector &mu and covariance matrix &Sigma .
Note that this model contains a total of *d(d+1)* parameters: there are *d* means, *d* variances, and *d(d-1)* correlations.

Permutation invariance

It can shown that the same statistical inferences will result, regardless of which of the variables is taken as the denominator in the log-ratios - i.e. that an approach based on analysing log-ratios is "permutation invariant" (Aitchison, 1986; Section 5.5). The choice of denominator is, consequently, somewhat arbitrary: we have taken the denominator to be X_1 in our definiton, but it could equally well have been X_D, or indeed any of the other variables. Permutation invariance is an attractive property of the log-ratio approach.

Alternative transformations

The transformation which we introduced above is usually known as the **alr** transformation, and is defined as

$$\phi_j(\mathbf{X}) = \left(\log\left(\frac{X^1}{X^j}\right), \ldots, \log\left(\frac{X^{j-1}}{X^j}\right), \log\left(\frac{X^{j+1}}{X^j}\right), \ldots, \log\left(\frac{X^D}{X^j}\right)\right)$$

Aitchison also proposes some alternative (but closely related) transformations, which are more useful for dealing with certain non-standard forms of compositional data.

Backtransformation

The inverse of the alr transformation is given by

$$\phi_j^{-1}(\mathbf{U}) = \left(\frac{\exp(U^1), \ldots, \exp(U^{j-1}), 1, \exp(U^{j+1}), \ldots, \exp(U^D)}{1 + \sum_{k \neq j} \exp(U^k)}\right)$$

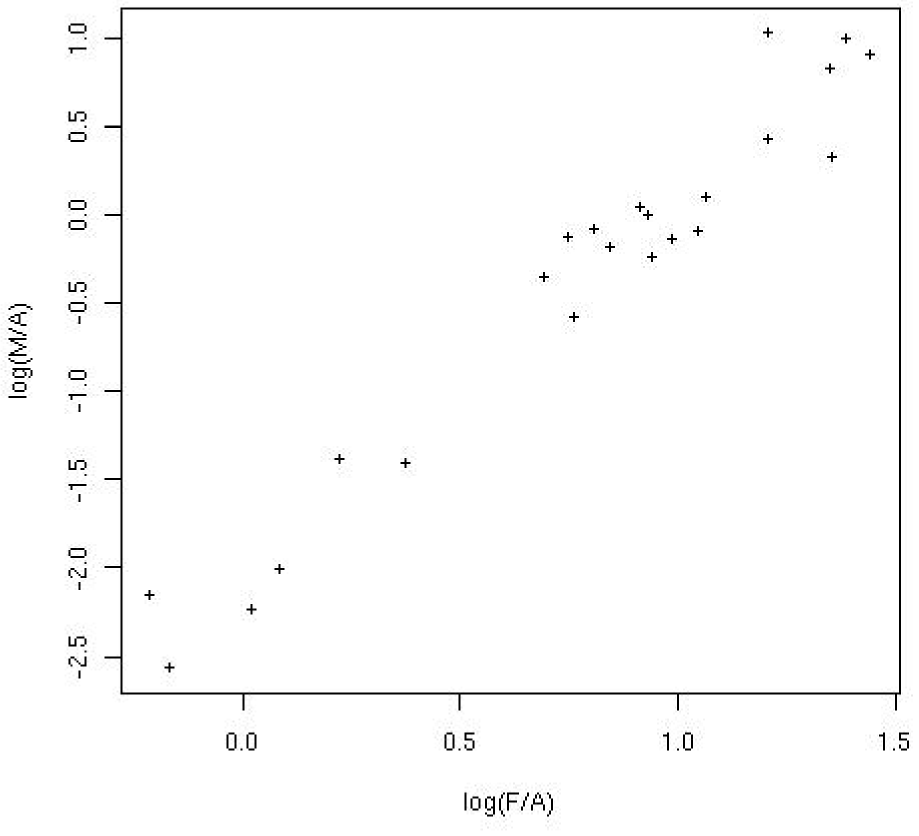
An alternative perspective on the multivariate normal model

Making the assumption that the log-ratios log(X^2 / X^1),...,log(X^D / X^1) have a multivariate normal distribution amounts to an assumption that the original variables X^1,...,X^D have a particular distribution - this distribution is usually called the multivariate log-ratio normal distribution. Fitting a multivariate log-ratio normal distribution to X^1,...,X^D is equivalent to fitting a multivariate normal distribution to the log-ratios log(X^2 / X^1),...,log(X^D / X^1).

A worked example (using R): composition of skye lavas

```
## (Just paste the following code into the R command line...)
## We analyse the skye lavas data,
## which give the mineral composition of 23 rock samples.
skye <-
read.table("http://www.bioss.ac.uk/~adam/alarm_training/data/skye.dat",header=TRUE)
## We take log-ratios...

## We take log-ratios...
logratio <- cbind(log(skye$F / skye$A),log(skye$M / skye$A))
## ...and draw a scatterplot:
plot(logratio[,1],logratio[,2],pch="+",xlab="log(F/A)",ylab="log(M/A)")
```

## We note that the log-ratios of M to A and F to A are highly positively correlated:



2.4 Statistical inference

Once we have transformed onto the log-ratio scale, then we can estimate the mean and variance of the (multivariate) normal distribution using the full range of standard methods available for statistical inference - such as **maximum likelihood**, **least squares**, or **Bayesian methods**. We do not discuss the relative practical & philosophical merits of the various approaches here, but do note that Bayesian inference via **Markov Chain Monte Carlo** is probably the most feasible option for dealing with high dimensional problems (when *D* is large).

2.5 Independence

What does the concept of statistical independence mean within the context of compositional data ? Standard ideas of independence are clearly irrelevant, because compositional data are subject to the constraint that they must sum to one. It turns out that a number of different concepts of "compositional independence" are possible, with different concepts being useful in different contexts. We restrict our attention here only to the most widely used concept of independence for compositional data, that of complete subcompositional independence.

Aitchison (1986, Chapter 10) defines a compositional vector X to have complete subcompositional independence if "the subcompositions formed from any partition of the composition form an independent set". If the log-ratios of X have a multivariate normal distribution, then this definition can be shown to be *equivalent* to a requirement that the covariance matrix Σ has elements of the form:

$$\Sigma_{ij} = \begin{cases} \alpha + \beta_i & \text{if } i = j \\ \alpha & \text{if } i \neq j \end{cases}$$

Complete subcompositional independence *always* holds when D=2 or D=3, so it is only a meaningful & useful concept when $D > 3$.

3 Compositional regression

We will often be interested in searching for relationships between a response variable and one or more explanatory variables (covariates), where either the response or explanatory variables are compositional. These kinds of problem can be dealt with through constructing an appropriate regression model on the log-ratio scale.

Some notation:

Assume that we have data on the relative proportions of D components,

$$\mathbf{X}_i = \left(X_i^1, \ldots, X_i^D\right)$$

for each of i=1,...,N individuals or replicates. Assume that we also know the values of M covariates z_{1i} ..., z_{Mi} for each individual/replicate.

3.1 A simple case

We begin with the simple case where the response variable has just two components (i.e. $D = 2$). The statistical strategy here involves:

- taking the log-ratio, $\log(X_2 / X_1)$;

- fitting a standard regression model to this log-ratio transformed data; and finally

- backtransforming the fitted regression model onto the original scale.

We will skip over the issues involved in selecting an appropriate regression model - since these are generic and widely discussed elsewhere - and focus on the mechanics of transformation and backtransformation.

<u>A linear model</u>

The most common approach will be to assume that the log-ratios are normally distributed,

$$\log\left(\frac{X_i^2}{X_i^1}\right) \equiv \left(\frac{1 - X_i^1}{X_i^1}\right) \sim \mathrm{N}\left(\mu_i, \sigma^2\right) \text{ for } i \in \{1, \ldots, N\},$$

where the mean μ_i is related to the covariates via a linear model of the form

$$\mu_i = a_0 + \sum_{k=1}^{M} a_k z_{ki}.$$

The regression coefficients a_1..., a_M are unknown parameters of the model, and, together with the variance σ^2 and intercept a_0, need to be estimated.

Through a backtransformation onto the original scale, we can use the regression model to obtain fitted values for X_i^1,

$$\hat{X}_i^1 = \frac{\exp\left(\hat{a}_0 + \sum_{k=1}^{M} \hat{a}_k z_{ki}\right)}{1 + \exp\left(\hat{a}_0 + \sum_{k=1}^{M} \hat{a}_k z_{ki}\right)}$$

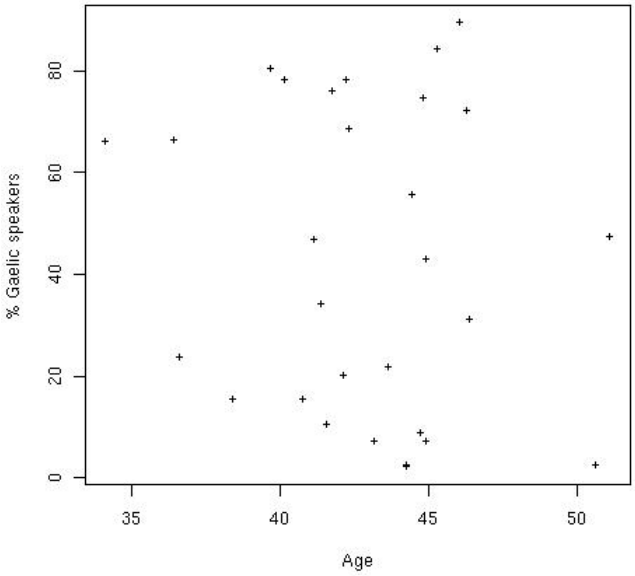
where the "hatted" regression coefficients denote estimated values of these coefficients.

A worked example (using R): backtransformation

```
## (Just paste the following code into the R command line...)
```

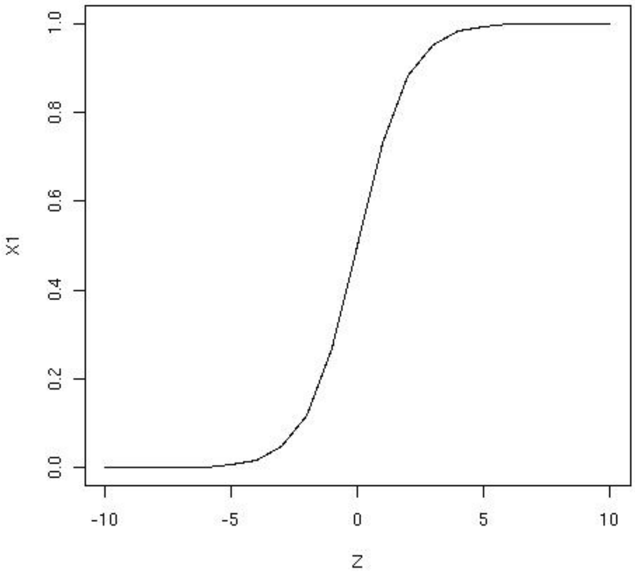
## We illustrate the idea of backtransformation using a simple example.
## Assume that *a* = 0 and *b* = 1 in the above model, so that a plot of *Z* against the log-ratio is linear:

plot(-10:10,-10:10,type='l',xlab="Z",ylab="log(X2 / X1)") :



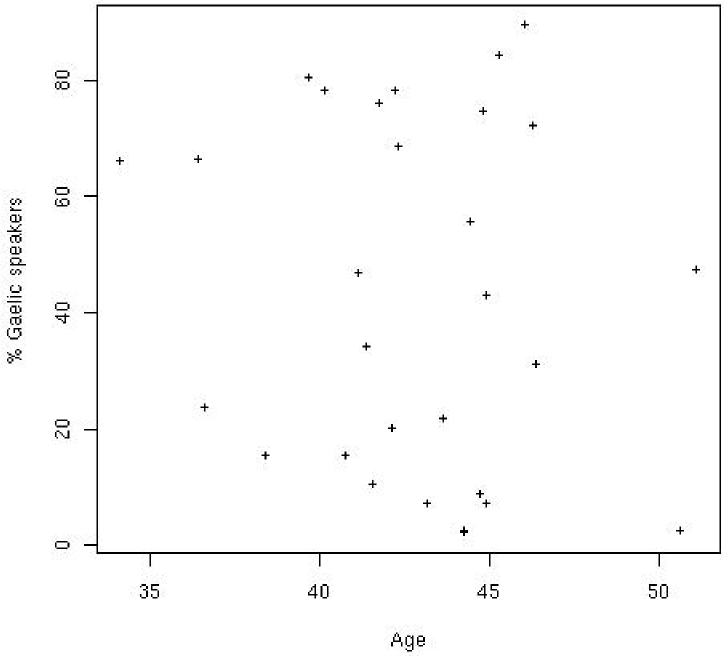
#Then a plot of *Z* against X_1 gives a logit curve:

plot(-10:10,exp(-10:10) / (1 + exp(-10:10)),type='l',xlab="Z",ylab="X1") :



A worked example (using R): the Gaelic language

```
## (Just paste the following code into the R command line...)

## The Gaelic data, which
## record the proportion of Gaelic speakers on 29 Scottish islands.
## We are interested in investigated whether the proportion of Gaelic speakers is related to the average age of the population.

## We load the data...
gaelic <- read.table("http://www.bioss.ac.uk/~adam/alarm_training/data/gaelic.dat",header=TRUE)
## ...and plot average age against the percentage of Gaelic speakers:
plot(gaelic$Age,gaelic$Gaelic,pch="+",xlab="Age",ylab="% Gaelic speakers"):
```



```
## There is no apparent relationship between the two variables,
## but we want to investigate this more formally using a statistical model.

## We transform to the log-ratio scale
logratio <- log(gaelic$Gaelic / (100 - gaelic$Gaelic))
## and fit a linear model:
summary(lm(logratio ~ gaelic$Age)):
```

```
Call:
lm(formula = logratio ~ gaelic$Age)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1075 -1.5176  0.2992  1.6118  3.0036 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.26323    3.77582   0.864    0.395
gaelic$Age  -0.08969    0.08771  -1.023    0.316

Residual standard error: 1.748 on 27 degrees of freedom
Multiple R-Squared: 0.03729,    Adjusted R-squared: 0.001635 
F-statistic: 1.046 on 1 and 27 DF,  p-value: 0.3155
```

## The *p*-value associated with the slope parameter is far from being statistically significant,
## so we certainly *cannot reject the null hypothesis* of no relationship between age and proportion.

## We could try introducing a quadratic term...
gaelic$Age2 <- gaelic$Age^2 ; summary(lm(logratio ~ gaelic$Age + gaelic$Age2)) :

```
Call:
lm(formula = logratio ~ gaelic$Age + gaelic$Age2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1092 -1.5170  0.2975  1.6108  3.0033 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.9998432 28.7625482   0.104    0.918
gaelic$Age  -0.0772554  1.3490311  -0.057    0.955
gaelic$Age2 -0.0001457  0.0157716  -0.009    0.993

Residual standard error: 1.781 on 26 degrees of freedom
Multiple R-Squared: 0.03729,    Adjusted R-squared: -0.03676 
F-statistic: 0.5036 on 2 and 26 DF,  p-value: 0.6101
```

## ...but we find that neither the linear or quadratic term is statistically significant.

3.2 Regression with more than two components

The linear model of the last slide can be extended to allow for more than two components (i.e. $D > 2$). Specifically, we assume that the log-ratios have a multivariate normal distribution,

$$\phi_1\left(\mathbf{X}_i\right) \equiv \log\left(\frac{X_i^2}{X_i^1}\right), \ldots, \log\left(\frac{X_i^D}{X_i^1}\right) \sim \mathrm{MVN}\left(\mu_i, \Sigma\right),$$

whose mean vector &mu$_i$ is related to the covariates through the linear model

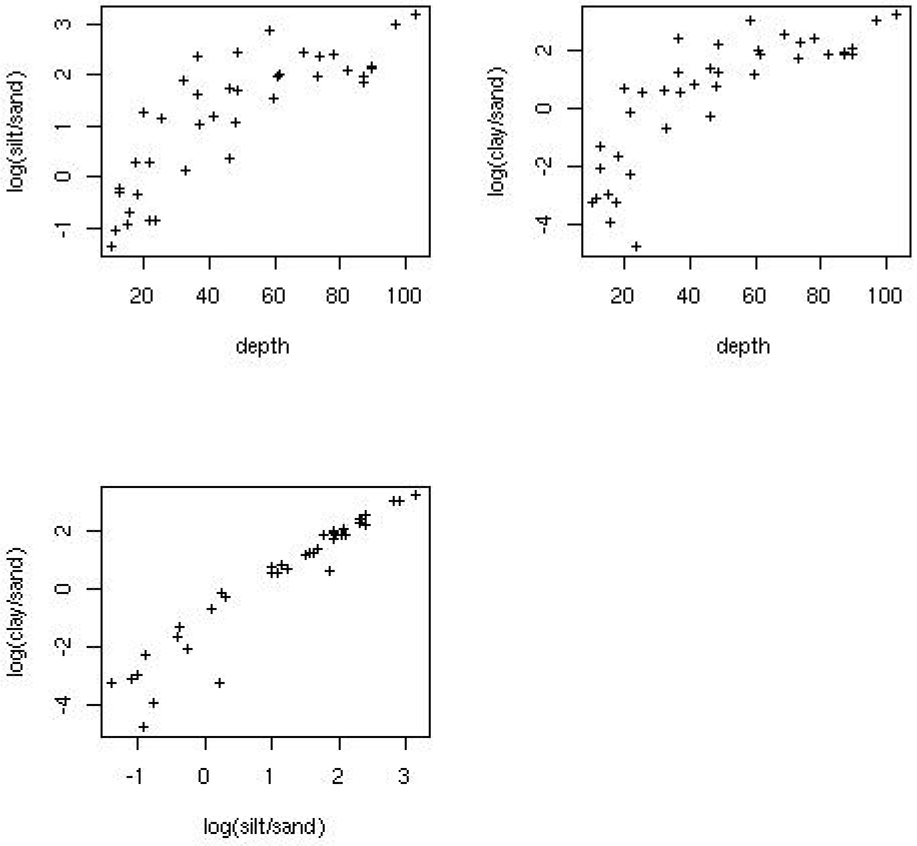
$$\mu_i = \mathbf{a}_0 + \sum_{k=1}^{M} \mathbf{a}_k z_{ki} \text{ for } i \in \{1, \ldots, N\}.$$

Note that the $\mathbf{a}_\mathrm{k}$ now refer to vectors of regression coefficients; the model can also be re-expressed without the use of vector notation, through the formula:

$$\mu_i^j = a_0^j + \sum_{k=1}^{M} a_k^j z_{ki} \text{ for } j \in \{1, \ldots, d\} \text{ and } i \in \{1, \ldots, N\}.$$

A worked example (using R): arctic sediments

```
## (Just paste the following code into the R command line...)

## The arctic sediments data
## record the proportion of sand, silt and clay found in 39 samples taken from different depths of an arctic lake.
## We are interested in studying how sediment composition changes with depth.

## We load the data
arctic <- read.table("http://www.bioss.ac.uk/~adam/alarm_training/data/arctic.dat",header=TRUE)
## and take log-ratios:
logratio <- data.frame(SiltSand = log(arctic$Silt / arctic$Sand), ClaySand = log(arctic$Clay / arctic$Sand))

## We can plot the log-ratios against depth, and against each other:
par(mfrow=c(2,2))
plot(arctic$Depth, logratio$SiltSand, pch="+", xlab="depth",ylab="log(silt/sand)")
plot(arctic$Depth, logratio$ClaySand, pch="+", xlab="depth",ylab="log(clay/sand)")
plot(logratio$SiltSand, logratio$ClaySand, pch="+", xlab="log(silt/sand)",ylab="log(clay/sand)")
```



```
## The scatterplots suggest that the ratios of silt to sand and clay to sand both increase with water depth;
## the plots also indicate a strong positive correlation between the two log-ratios.
```

3.3 Interpretation

Having estimated the parameters of the multivariate normal distribution on the log-ratio scale, we can then backtransform onto the original (compositional) scale. The backtransformed parameter estimates do not necessarily have any obvious interpretation, however.

One important exception is the **intercept parameter**, $\boldsymbol{a}_0$; Billheimer et al. (2001) note that we can backtransform an estimate of the intercept parameter using

$$\phi_1^{-1}(\hat{\mathbf{a}}_0)$$

to obtain an estimate for the multivariate median of the composition.

We can also backtransform fitted values from the model onto the original scale, using

$$\phi_1^{-1}\left(\hat{\mathbf{a}}_0 + \sum_{k=1}^{M} \hat{\mathbf{a}}_k z_{ki}\right)$$

3.4 Multivariate analysis

If *D* is large, then we may be interested in using multivariate methods to summarise the principal source of variation within *X*: i.e. to achieve dimension reduction. Standard methods of multivariate analysis - such as principal components analysis - should *not* be applied in an unmodified way to compositional data (Aitchison, 1986), because these methods typically rely on an underlying assumption of multivariate normality.

Suitably *adapted* versions of multivariate methods, have, however, often been developed for use in the compositional context: for example, principal logcontrast analysis (Aitchison, 1983) provides a (statistically valid) compositional analogue of principal components analysis.

4 Spatial analysis of species composition

An additional statistical challenge when analysing spatially referenced compositional data arises from the possible presence of spatial autocorrelation in the data or model residuals. Spatial autocorrelation is very common in many spatially referenced data sets, because things which are near to each other tend to be more similar than things that are far apart. In the presence of spatial autocorrelation, errors *cannot* be assumed to be independently distributed, violating the basic assumption of usual linear modelling techniques (Haining, 1990). Spatial autocorrelation tends to lead to overestimation of the number of degrees of freedom if it is ignored, and to serious inflation of the Type I error rate (Legendre 1993). The effects of covariates may be also be overestimated (Cressie 1993 ; Anselin & Bera, 1998).

In this section we outline a modern statistical approach for dealing with spatially autocorrelated compositional data, and illustrate this approach using gridded data on the spatial distribution of a specific biological trait (flower colour). Grid-based data frequently arise in the context of species atlases, so the methods which we present have broad applicability within ecology.

Some notation:

Assume that we have data on the relative proportions of *D* components,

$$\mathbf{X}_i = \left(X_i^1, \ldots, X_i^D\right)$$

for each of *i*=1,...,*N* cells on a regular spatial grid. Assume that we also know the values of *M* covariates z_{1i} ..., z_{Mi} for each grid cell *i*.

4.1 A spatial model

Spatial dependence can be incorporated into the analysis of compositional data by using a conditional autoregressive model (CAR model: Besag, 1974; Billheimer et al. 1997). Here, we use an intrinsic version of the multivariate CAR model, proposed by Besag, York and Mollie (1991).

We assume that the log-ratios have a multivariate normal distribution

$$\phi_1\left(\mathbf{X}_i\right) \equiv \log\left(\frac{X_i^2}{X_i^1}\right), \ldots, \log\left(\frac{X_i^D}{X_i^1}\right) \sim \mathrm{MVN}\left(\mu_i, \Sigma\right)$$

and use a linear regression model:

$$\mu_i = \mathbf{a}_0 + \sum_{k=1}^{M} \mathbf{a}_k z_{ki} + \mathbf{S}_i \text{ for } i \in \{1, \ldots, N\}$$

to describe the effect of the spatial covariates z_{1i} ..., z_{Mi} upon the mean vector &mu$_i$.

Spatial random effects

Comparing against our earlier non-spatial model, we can see that the new elements within the CAR model are the multivariate random effects,

$$\mathbf{S}_i = \left(S_i^1, \ldots, S_i^d\right) \text{ for } i \in \{1, \ldots, N\},$$

which are assumed to be spatial correlated and Gaussian (multivariate normal). Define the neighborhood &delta(*i*) of grid cell *i* to be the set of grid cells which are directly adjacent or diagonal to grid cell *i*, and let *n*(*i*) denote the number of grid cells in neighbourhood &delta(*i*). The spatial random effects are assumed to have conditional distributions of the form

$$\mathbf{S}_i | (\mathbf{S}_1, \ldots, \mathbf{S}_{i-1}, \mathbf{S}_{i+1}, \ldots, \mathbf{S}_N) \sim \mathrm{MVN}\left(\sum_{I \in \delta(i)} \mathbf{S}_I / n(i), \Omega / n(i)\right) \text{ for } i \in \{1, \ldots, N\},$$

where Ω is a *d* x *d* dimensional variance-covariance matrix. Note that the conditional means are simply the mean random effects within the neighbourhood &delta(*i*) of grid cell *i*, and that the conditional covariance matrix is inversely proportional to the number of grid cells *n*(*i*) contained in this neighbourhood. The random effects for each component are constrained to sum to zero over space (Besag & Kooperberg, 1995), so that

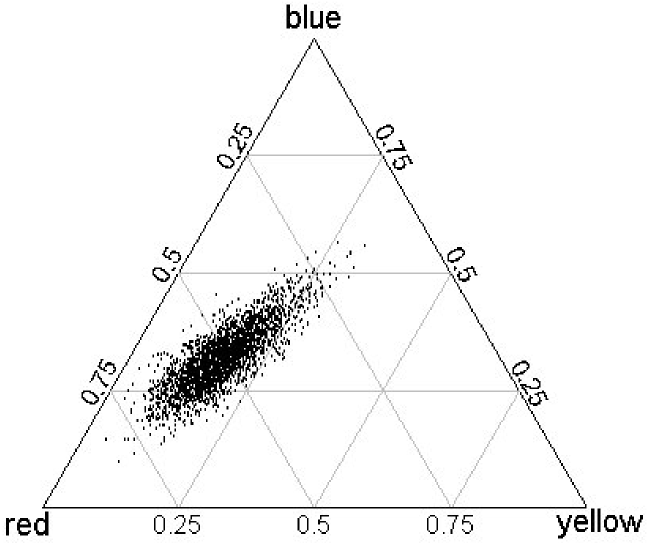
$$\sum_{i=1}^{N} S_i^k = 0 \text{ for } k \in \{1, \ldots, d\}.$$

4.2 Example: flower colours in the UK (simulated data)

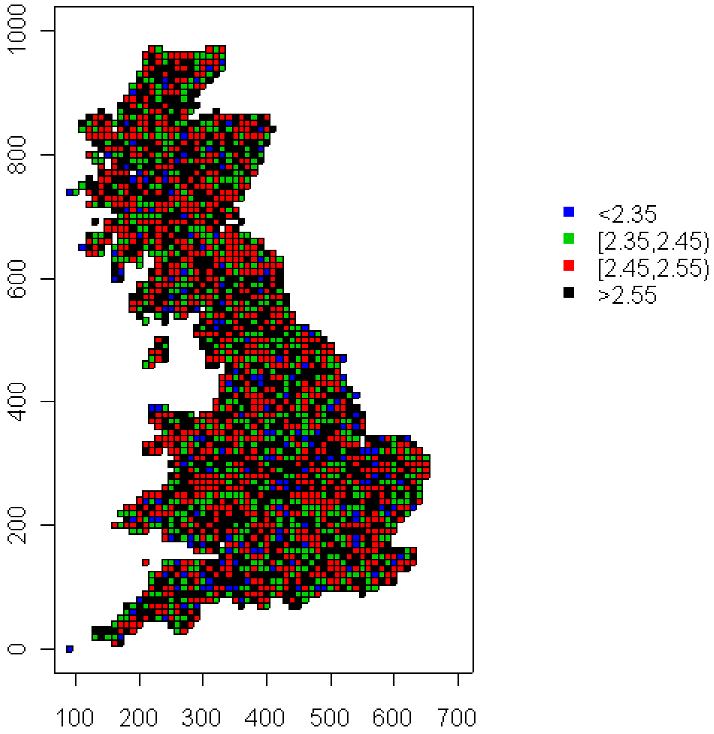
We illustrate the methodology through a case study concerning the (hypothetical) spatial distribution of a biological trait.

We use a simulated dataset on the proportion of flowers having each of three different colours (red, blue & yellow) for cells on the 10km-by-10km Ordnance Survey grid of the United Kingdom. Proportions of flower colours are simulated in such a way that spatial variations in the proportions are related to two (also hypothetical) variables, which we call the biodiversity index and the limestone index. The biodiversity indices are spatially independent, but the limestone indices exhibit strong spatial autocorrelation - i.e. limestone indices for grid cells which are close together will tend to be similar. The raw data, and the methods used to gerenrate the data are available as an appendix.

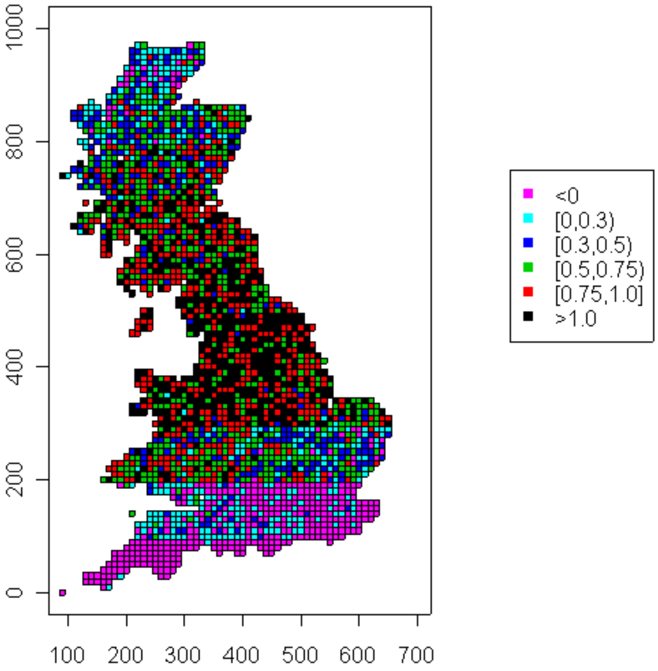
This ternary diagram summarises the composition of simulated flower colours:



This map illustrates the spatial distribution of the biodiversity index,



whilst this one show the spatial distribution of the limestone index:



4.3 Model fitting

We let

$$\begin{aligned} X_i^1 &= \text{proportion of red flowers in grid cell } i \\ X_i^2 &= \text{proportion of blue flowers in grid cell } i \\ X_i^3 &= \text{proportion of yellow flowers in grid cell } i \end{aligned}$$

and model the proportions of flower colours using the CAR model introduced above. Note that

$$X_i^1 + X_i^2 + X_i^3 = 1 \text{ for } i \in \{1, \ldots, N\},$$

so that the data are genuinely compositional.

We will assume (pretend) that we have no knowledge of the limestone index, and will try to relate the proportions of different flower colours solely to the biodiversity index. We fit a linear regression model of the form:

$$\mu_i^1 = a_0^1 + a_1^1 z_i, \text{ and } \mu_i^2 = a_0^2 + a_1^2 z_i$$

where the log-ratios are assumed to have a bivariate normal distribution

$$\left(\log\left(\frac{X_i^2}{X_i^1}\right), \log\left(\frac{X_i^3}{X_i^1}\right)\right) \sim \text{BVN}\left((\mu_i^1, \mu_i^2), \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}\right)$$

Note that $D = 3$, $d = 2$, $M = 1$, and $N = 2523$.

4.4 Parameter estimation using WinBugs

We estimate the parameters of the model within a Bayesian framework (see e.g. Gelman et al., 1995). Within the Bayesian context parameters are treated as random variables, and all inferences about the parameters are based on the posterior distribution. The posterior distribution is obtained by using Bayes' theorem to combine prior knowledge about the parameters together with knowledge about the parameters gained from the data.

Or equivalently, but put into more formal language...

If $f(\theta)$ denotes the prior distribution for the parameters &theta of a statistical model, and if $f(X|\theta)$ denotes the likelihood of data X within this model, then Bayes' theorem states that

$$f(\theta|\mathbf{x}) \propto f(\theta)f(\mathbf{x}|\theta)$$

where $f(\theta|\boldsymbol{X})$ is the posterior distribution of the parameters &theta .

We used the GeoBUGS module (Thomas et al. 2004) of the WinBUGS package (Spiegelhalter et al. 1999; Link et al., 2002) to fit the multivariate CAR model using Markov chain Monte Carlo (McMC) techniques. WinBUGS/GeoBUGS provides an efficient and user friendly environment for fitting a wide range of sophisticated statistical models, many of which would be prohibitively difficult and/or expensive (in time and labour) to fit in any other way. The WinBUGS code which we have used is available here.

We must specify prior distributions for the parameters of the spatial model. We choose to use uninformative priors (also called vague priors) for all of the parameters, since we do not wish to make strong prior assumptions about the values of the parameters.

We take the prior for the intercept term to be a location invariant uniform distribution, so that

$$a_0 \sim U(-\infty, \infty).$$

Note that this is an improper prior.

We take priors for the regression coefficients to be normally distributed with zero mean and a large variance, so that

$$a_k^j \sim N(0, 100) \text{ for } j \in \{1, \ldots, d\} \text{ and } k \in \{1, \ldots, M\}.$$

Finally, we take priors for the precision matrices (i.e. inverted variance-covariance matrices) within our model to be Wishart distributions,

$$\Sigma^{-1} \sim \text{Wishart}(A, h);\ \ \Omega^{-1} \sim \text{Wishart}(B, h)$$

with $h = 2$ degrees of freedom and variance matrices of the form

$$A = B = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$$

The parameter *a* determines the precision of the spatial random effect within the CAR model, and it is important to test the sensitivity of the final results to the choice of prior value for this parameter. We take *a* = 0.1 in our analysis, but obtained similar results with the alternative choices *a* = 0.01 and *a* = 0.005.

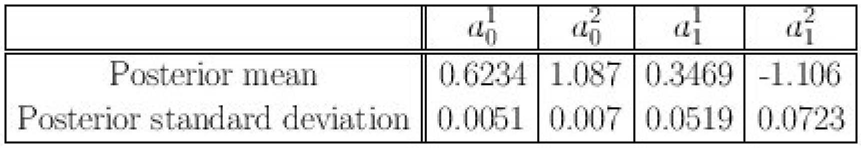
Multiple Markov chains were run starting from different points, and the chains were assumed to have converged when the 2.5%, 50% and 97.5% quantiles of the posterior distributions obtained using the combined chains were sufficiently similar to those obtained from the individual chains. We further checked convergence of the Markov chains using visual inspection of the series of parameter estimates, autocorrelations of these series, and the Brooks-Gelman-Rubin statistic (Brooks & Gelman, 1998).

We ran our Markov chains for 200000 iterations, discarded the first 20000 draws as burn-in, and used only every 1 in every 10 iterations in order to overcome the effects of autocorrelation. We based our inferences on the remaining 18000 draws from the the posterior distribution.

4.5 Results from the flower colour example

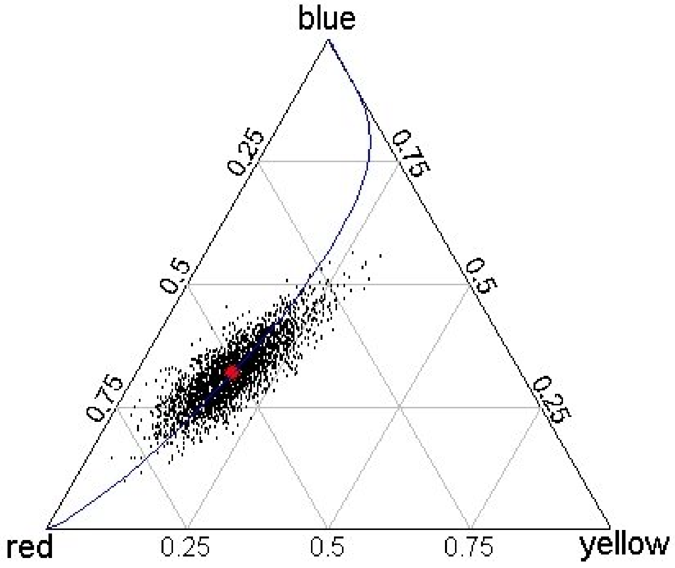
We fit the Bayesian CAR model to the data using GeoBugs. The code for the WinBUGS model is given below.

The posterior means and posterior standard deviations for the regression parameters are:

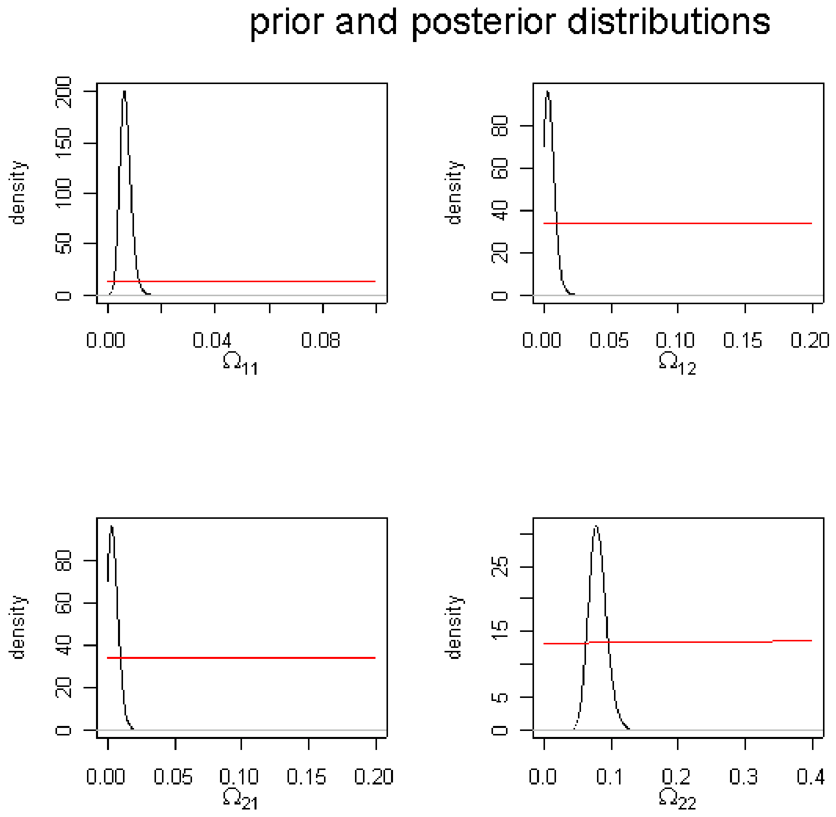


	a_0^1	a_0^2	a_1^1	a_1^2
Posterior mean	0.6234	1.087	0.3469	-1.106
Posterior standard deviation	0.0051	0.007	0.0519	0.0723

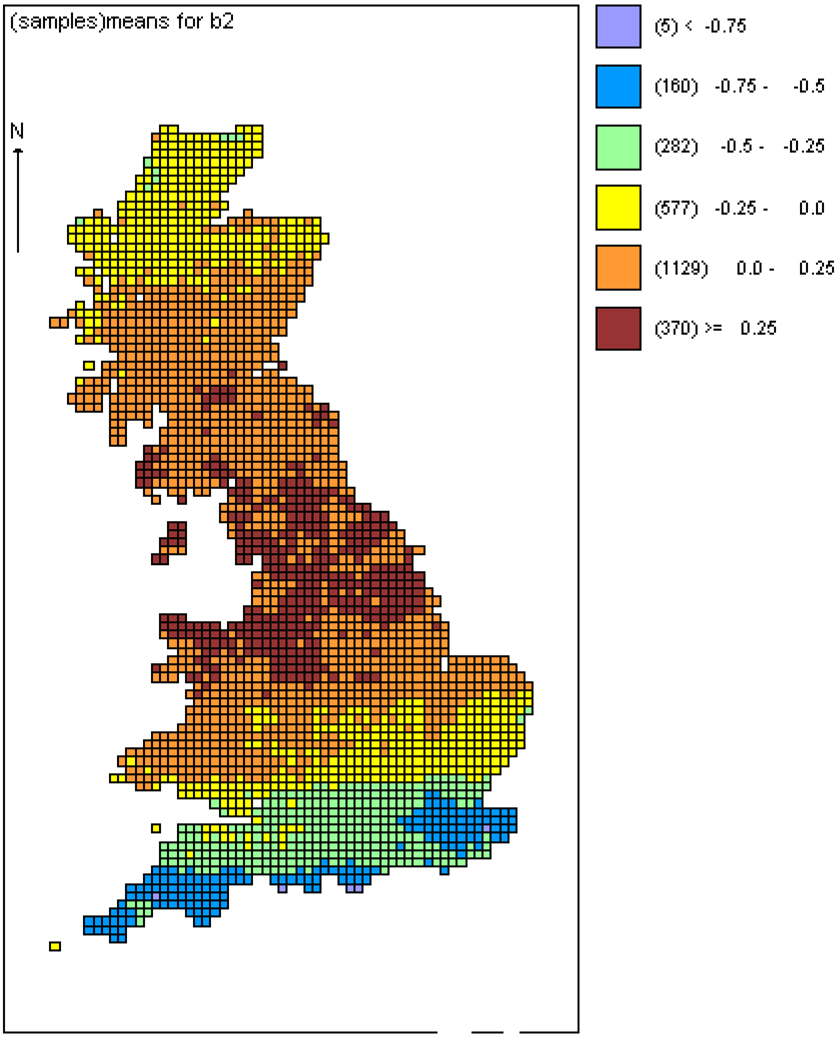
The following ternary plot depicts the observed log-ratios (points), the fitted median composition (obtained by back-transforming the posterior means of a_0^1 and a_0^2), and the fitted regression line of the composition, as predicted by the observed gradient in species-richness :



The prior distributions (red) and posterior distributions (black) for the elements of the variance-covariance matrix of spatial random effects, Ω, are shown in the next set of graphs. The posterior means for the variances Ω_{11} and Ω_{22} are small, indicating substantial levels of spatial autocorrelation in the model residuals. This residual spatial autocorrelation results from the fact that the limestone index is not included as a covariate within our model.



Spatial patterns in the posterior means of the spatial random effects $\hat{\mathbf{S}}_i$ closely resemble the spatial pattern of the omitted "limestone index" covariate z_{2i}. The following map depicts the spatial distribution of $\hat{S}_i^1$, with a similar pattern for $\hat{S}_i^2$.



The inclusion of the spatial random effects greatly decreased the residual deviance of the model, indicating a substantial improvement in model performance over the non-spatial version of the model.

5 Problematic compositional data

Statistical methods for dealing with compositional data using multivariate normal models for log-ratios are well-established, and are motivated by sound theoretical principles. This does not mean, however, that the techniques are suitable for analysing absolutely *any* compositional dataset which we may encounter. In this section we introduce some particularily "messy" datasets, explain why log-ratio normal models are inappropriate for analysing these data, and - where appropriate - suggest some alternative analyses.

5.1 Too many components

Some compositional datasets contain data on a very large number of components, *D*. For example, household expenditure studies often record expenditure for an extensive range of different commodities - bread, jam, toothbrushes, soap, toasters, etc. etc.. We cannot usually apply LR-normal models to such datasets, because:

the number of parameters in the model, *d(d+3)*, will often be almost as large as - or potentially even larger than - the number of replicates *n*, leading to overfitting; and

some of the components are likely to contain zero proportions, preventing us from calculating log-ratios (see below).

The best solution will often - although not always - be to aggregate the data into a smaller number of compositions, and then to fit an LR-normal model to this "reduced" dataset. For example, we may group commodities together into a small number of broad classes (food, toiletries, electrical goods etc.). Note that we are inevitably going to lose a lot of information by doing this, so this approach will only be appropriate if we can aggregate components together in some sensible and uncontroversial way.

A worked example (using R): time off

```
## (Just paste the following code into the R command line...)

## We use a dataset on the proportion of time that I spend in the office and elsewhere: These data describe my daily activities during my first 6 months at BioSS. For each of six blocks of four working weeks (20 working days), the data show the number of days in which I was: at work in my office, away at a conference/meeting, away due to a public holiday (e.g. Christmas), away on annual leave, or away on sick leave.

## We begin by loading the data...
leave <-
read.table("http://www.bioss.ac.uk/~adam/alarm_training/data/leave.dat",header=TRUE)
## ...and converting them into proportions:
leave <- leave / 20
```

```
## The data record days spent in five different ways by a statistician (me) during six blocks of four weeks:
## at the office, at conferences/meetings, on public holidays, on annual leave, or away sick.
## They contain a number of zero proportions (e.g. because some four week blocks do not contain any public holidays).

## We could aggregate the data into just two constituents ("at the office" and "away"):
leave_aggregated <- data.frame(Office = leave$Office,
Away = leave$Conferences + leave$Public_holiday + leave$Annual_holiday + leave$Sick):
```

```
  Office Away

1   0.70 0.30

2   0.80 0.20

3   0.80 0.20

4   0.65 0.35

5   0.75 0.25

6   0.85 0.15
```

```
## This certainly gets rid of the zero proportions, but in simplifying things we also get rid of possibly interesting information...
```

5.2 Zero proportions

Many compositional datasets contain some zero proportions - i.e. for some of the observations, one or more of the components is entirely absent. We cannot take the logarithm of zero, so compositional data which contain zero proportions cannot be dealt with using log-ratios. Statistical methods for dealing with zero proportions are still under development, and no truly generic techniques for dealing with this problem have yet been put forward.

In some circumstances, zero proportions in the data may correspond to non-zero proportions which are so small as to be below the limits of detection (so-called "trace zeroes"), in which case it is generally agreed that the imputation methods of Fry et al. (2000) and Martin-Fernandez et al. (2003) should be used.

If - as in the example below, and as in many ecological applications - the zero proportions in the data genuinely correspond to zero proportions (so-called "essential zeroes") then things are less straightforward. The hierarchical modelling approach proposed by Aitchison & Kay (2003) appears promising, but to date there have been no substantial applications of this methodology.

A worked example (using R): do nurses wear gloves ?

We consider the nurses data:

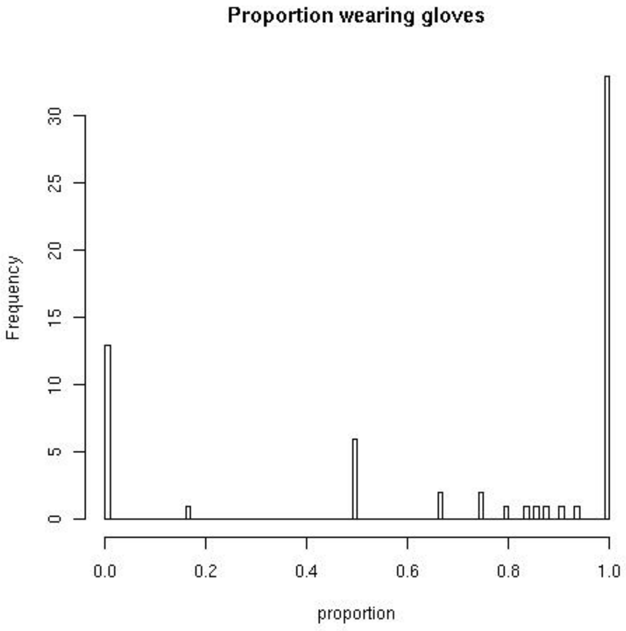
These data quantify the extent to which nurses in a particular paediatic hospital used gloves when performing routine medical activities. There are a total of 63 records, corresponding to data for 23 individual nurses. The data record the number of procedures undertaken by nurses during the period of observation (the "Obs" column), and the number of these procedures for which gloves were worn (the "Gloves" column). Data were collected prior to an short educational programme designed to encourage glove use, and were subsequently collected one, two and five months after the completion of this programme (the "Period" column denote the time point: 1 = before programme, 2 = one month after, 3 = two months after, 4 = five months after). Note, however, that not all of the nurses were monitored at all four time points. The number of years experience for each nyse is also recorded (the "Experience" column). The data are due to Friedland et al. (1992), and are available through the DASL website (http://www.dasl.datadesk.com/parse.acgi?datafile=Nurses).

The data record whether or not nurses in an emergency department wear gloves when undertaking

```
## (Just paste the following code into the R command line...)
```

medical procedures.

```
## We begin by reading in the data...
nurses <- read.table(file="http://www.bioss.ac.uk/~adam/alarm_training/data/nurses.dat",header=TRUE)

## ...and then working out the proportion of nurses that do and do not wear gloves:
proportion <- data.frame(gloves = nurses$Gloves / nurses$Obs, nogloves = (nurses$Obs - nurses$Gloves) / nurses$Obs)

## We can plot a histogram of the proportions:
hist(proportion$gloves,breaks=seq(0,1,0.01),xlab="proportion",main="Proportion wearing gloves")
```



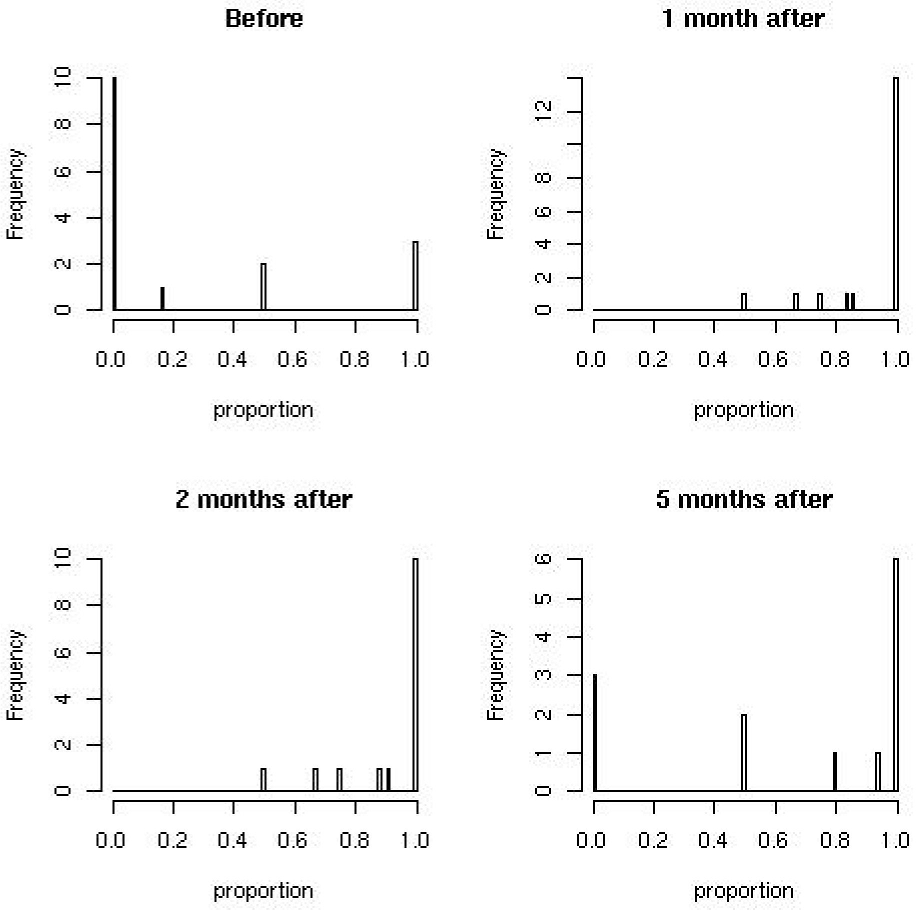
## We see that the majority of proportions are either zero or one.
## The zero values in this example are *essential zeroes*: some nurses *never* wear gloves, whilst some nurses *always* wear gloves.

### **The effect of training**

## Data were collected before, and at three timepoints after, the nurses attended an educational programme.
## The variable 'Period' identifies the timepoint (1 = before training, 2 = one month after, 3 = two months after, 4 = five months after).
## There is interest is in understanding whether the training course has increased the proportion of nurses that wear gloves.

## We can plot seperate histograms for data at each of the four timepoints:

```
par(mfrow=c(2,2))
hist(proportion$gloves[nurses$Period ==
1],breaks=seq(0,1,0.01),xlab="proportion",main="Before")
hist(proportion$gloves[nurses$Period ==
2],breaks=seq(0,1,0.01),xlab="proportion",main="1 month after")
hist(proportion$gloves[nurses$Period ==
3],breaks=seq(0,1,0.01),xlab="proportion",main="2 months after")
hist(proportion$gloves[nurses$Period ==
4],breaks=seq(0,1,0.01),xlab="proportion",main="5 months after")
```



## We see that prior to the training course, a large proportion of the nurses never wore gloves.
## One to two months after the training course, most of the nurses wear gloves for every procedure.
## Five months after the training course levels of glove use have fallen substantially -
## but are still higher than they were before the training course.

## This exploratory analysis gives us some interesting insights about the possible effects of the educational programme.
## In order to gain solid *evidence* for any effect, however, we would need to analyse these data using a formal statistical model.

## Any model would have to take account of:
## - the fact that many of the proportions are zero or one;
## - the fact that observations for the same nurse in different periods are almost certain to be dependent; and
## - the fact that the dataset is actually quite small.
## Finding an appropriate model is likely to be difficult !

5.3 Non-normality

The log-ratios of X will not always be well described by a multivariate normal model, so it is important to verify that the assumption of normality is indeed a reasonable one. The good news is that, once we have transformed our compositional data onto the log-ratio scale, then we can use standard statistical tools to investigate whether the data are normal.

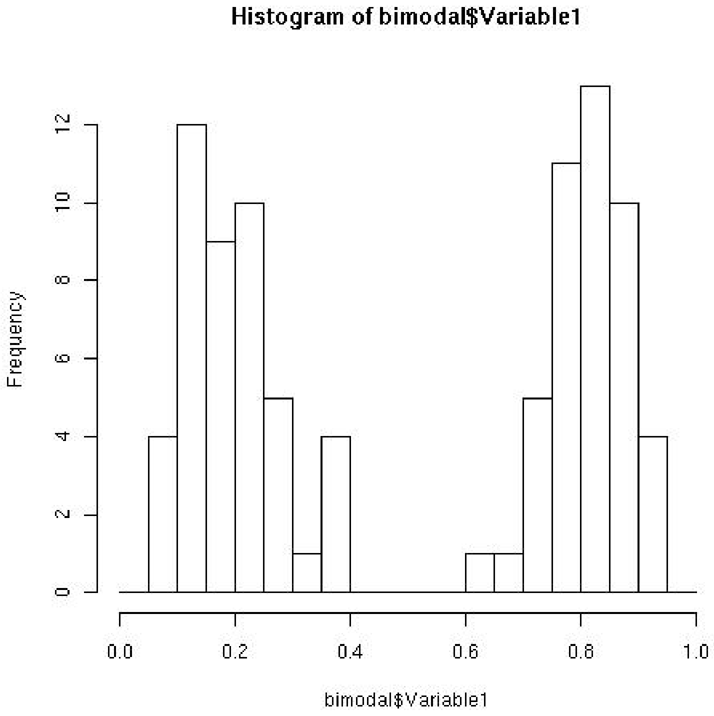
Quantile-quantile and probability-probability plots can be used to search for deviations from normality, and a wide range of formal tests for normality are also available. Some tests for *multivariate* normality are also available. Many tests for normality have low statistical power, however, and it is important to note that a failure to reject the null hypothesis of normality *does not necessarily imply that the data are actually normal*.

If the diagnostic tests do find evidence of non-normality, then this suggests that the normal model is not reasonable, so we must choose an alternative model for our data. Choosing an appropriate model for compositional data is unlikely to be straightforward, and at this point it may be sensible to seek the advice of a statistician. Aitchison (1986, Ch.13) proposes some powerful generalisations of the LR-normal distribution, and suggested that these are likely to be useful for dealing with non-normal data.

What if our diagnostic tests do not find any evidence for non-normality ? In this case we still cannot conclude that the data are normally distributed, but we can now conclude that the data at least appear to be *consistent* with an assumption of normality. We would then - usually - go ahead and fit an LR-normal model to the data, always bearing in mind that the outcomes from our analyses will be dependent upon the validity of the normality assumption.

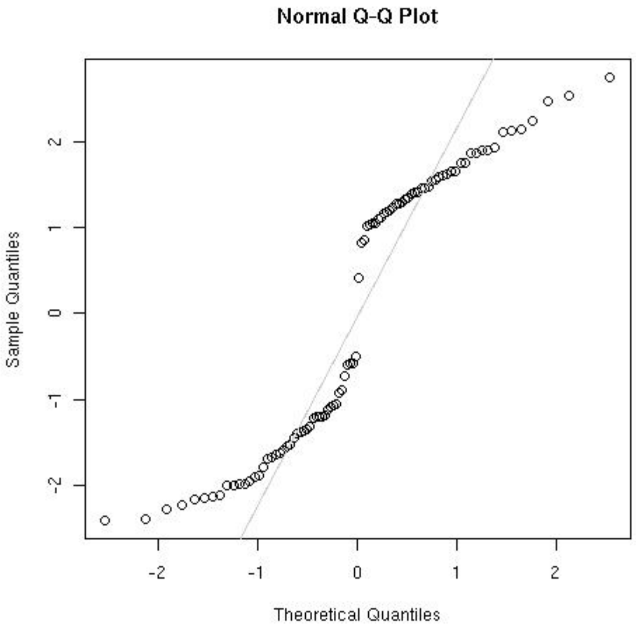
A worked example (using R): the curse of bimodality

```
## (Just paste the following code into the R command line...)

## We use a small simulated dataset (the "bimodal" dataset, which is described here)
## to demonstrate one way in which the normality assumption may fail.

## We begin by loading the data:
bimodal <-
read.table(file="http://www.bioss.ac.uk/~adam/alarm_training/data/bimodal.dat",header=TRUE)

## Now we plot a histogram...
hist(bimodal$Variable1,breaks=seq(0,1,0.05))
```



## ...which shows (pretty clear) evidence of bimodality

```
## We transform the data onto the log-ratio scale:
logratio <- log(bimodal$Variable1 / bimodal$Variable2)

## Now we can assess the validity of the normality assumption...
## ...graphically, using a Q-Q plot:
qqnorm(logratio) ; qqline(logratio,col=gray(0.8))
```



```
## ...and formally, using a Shapiro test:
shapiro.test(logratio)

        Shapiro-Wilk normality test

data:  logratio

W = 0.8793, p-value = 5.37e-07

## The QQ plot shows serious discrepancies between the data and the fitted (normal)
model,
## whilst the Shapiro test conclusively rejects the null hypothesis of normality (with a
p-value of less than 0.001)
```

5.4 Log-ratio sceptics

Most experts accept that analyses of compositional data should usually be based on fitting a normal model to log-ratios, unless there are specific problems in adopting such an approach (e.g. there are zero proportions, or the log-ratios exhibit non-normality). A group of skeptical mathematical geologists have, however, argued that the whole approach of fitting a statistical model to log-ratios of the proportions is fundamentally misguided (e.g. Philip & Watson, 1988; Philip & Watson, 1989). They advocate an alternative approach, based upon modelling the angles between compositions using the geometry of the sphere rather than that of the simplex.

6 Additional resources

6.1 Software

Statistical methods for compositional data are not usually included within standard statistical software packages, possibly explaining why these techniques are not more widely used. Most of the methods are comparatively straightforward to code up from scratch, however, and a number of add-ons to existing packages are also available.

R / Splus

The **R** (http://cran.r-project.org/) and **S** languages provide a powerful interactive environment for statistical programming. **Splus** (http://www.mathsoft.com/) is a widely-used commerical package which uses the S language, whilst R is a free, open-source variant. A number of existing **R** functions can prove very useful when analysing compositional data:

- the **ternaryplot** function in the **vcd** library, for visualising data with three components;
- the **qqnorm**, **qqline** and **shapiro.test** functions in the **stats** library, for testing normality;
- the **rmvnorm** function in the **mvtnorm** library, for simulating from a multivariate normal;
- the **mvnormtest** library, for testing multivariate normality.

The full range of contributed R libraries can be downloaded - easily, and free of charge - from the CRAN repository. Matevz Bren and Vladimir Batagelj are writing an R library called **mixture** (http://ima.udg.es/Activitats/CoDaWork03/paper_Ben_Batagelj.PDF) on compositional data analysis, but this doesn't yet seem to be available on CRAN. Joel Reynolds has written a set of R/Splus functions (http://www.biostat.wustl.edu/archives/html/s-news/2003-12/msg00134.html) to implement various methods in compositional data analysis, based partly on previous code written by Dean Billheimer. The **R** code which we have used in this course is available online at http://www.bioss.ac.uk/~adam/alarm_training/rcode.html .

General regression models for compositional data can be fitted by Bayesian methods using **WinBugs** (http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml), or the open-source version **OpenBugs**.

The spatial models (chapter 4 of this module) can be fitted using **GeoBugs** (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/geobugs.shtml), a spatial modelling variant of WinBugs.
The GeoBugs code which we have used within this course is available as an appendix.

John Aitchison has written a suite of BASIC programs, called **CODA**, to implement a wide range of techniques for compositional data analysis. If you wish to buy CODA, then you should contact Chapman and Hall directly. We have no experience in using CODA ourselves, but, according to the INQUA Working Group on Data-Handling Methods, 1990 (http://www.kv.geo.uu.se/inqua/news3/nl3hj3b3.htm) "...the CODA programs are easy-to-use and powerful, but output is rather limited and graphics are poor...".

Smith Ecology ltd. produce and sell an Excel add-in called **Compos Analysis** (http://www.smithecology.com/sotware.htm) for implementing many of the techniques given in Aitchison (1986).

6.2 Key Publications

The literature on compositional data analysis is quite extensive, and ranges from the highly mathematical to the very applied. Most research papers are published in statistics, geology or economics journals, but papers are also becoming increasingly common within ecology/biology journals.

The book by Aitchison (1986) remains the seminal work in this area, and a new edition has recently been published (Aitchison, 2003a). Aitchison outlines the problems in applying standard statistical methods to compositional data, puts forward an alternative approach based upon the analysis of log-ratios, and discusses issues surrounding statistical modelling and inference. In a recent review paper Aitchison (2003b) gives an overview of the history of compositional data analysis, and of methodological developments in the subject since 1986.

Ecological applications

Aebischer et al. (1993) use a log-ratio approach to analyse animal radio-tracking data. Billheimer et al. (2001) show how log-ratio regression models can be used to analyse ecological data relating to species composition. Elston et al. (1996) show how data on the diet compositions of animals that are offered a range of options can be analysed using log-ratio analysis.

Workshops & conferences

A workshop on compositional data analysis was held in 2003 - the CODAWORK'03 workshop (http://ima.udg.es/Activitats/CoDaWork03/) - and the resulting papers were posted online. A second workshop, CODAWORK'05 (https://www.fundacioudg.org/fudgif/cat/prog-codaworka.asp?codi=2833), will be held during 2005.

6.3 Bibliography

Adolph, C. (2005)
Succession in the Temple: Central Banker Careers and the Politics of Appointment. Working paper, http://faculty.washington.edu/cadolph/homepage/app.pdf.

Aitchison, J. (1982).
The statistical analysis of compositional data (with discussion).
Journal of the Royal Statistical Society B, **44**, 139-177.

Aitchison, J. (1983)
Principal component analysis of compositional data.
Biometrika, **70**, 57-65.

Aitchison, J. (1986)
The Statistical Analysis of Compositional Data (1st edn.), Chapman and Hall (London).

Aitchison, J. (2003a)
Compositional Data Analysis: where are we and where should we be heading ?
Conference paper for the *CoDaWORK workshop on Compositional data analysis, Girona, September 2003*,
http://ima.udg.es/Activitats/CoDaWork03/Girona_invited_lecture_Aitchsion.pdf.

Aitchison, J. (2003b)
The Statistical Analysis of Compositional Data (2nd edn.), Chapman and Hall (London).

Aebischer, N.J., Robertson, P.A. & Kenward, R.E. (1993)
Compositional analysis of habitat use from animal radio-tracking data. *Ecology*, **74**, 1313-1325.

Aitchison, J. and Kay, J.W. (2003)
Possible solutions of some essential zero problems in compositional data analysis.
Conference paper for the *CoDaWORK workshop on Compositional data analysis, Girona, September 2003*,
http://ima.udg.es/Activitats/CoDaWork03/paper_Aitchison_and_Kay.pdf.

Anselin, L. and Bera, A. (1998)
Spatial dependence in linear regression models with an application to spatial econometrics.
In A. Ullah and D. Giles (eds.), *Handbook of Applied Economic Statistics*, pp.237-289, Marcel Dekker (New York).

Besag, J., York, J and Mollie, A. (1991).
Bayesian image restoration, with two applications in spatial statistics (with discussion).
Annals of the institute of statistical mathematics, **43**, 1-59.

Besag, J. and Kooperberg, C.L. (1995).
On conditional and intrinsic autoregressions.
Biometrika, **82**, 733-746.

Besag, J. (1974).
Spatial interaction and the statistical analysis of lattice systems.
Journal of the Royal Statistical Society B, **36**(2), 192-236.

Billheimer, D., Guttorp, P. and Fagan, W.F. (2001)
Statistical Interpretation of Species Composition.
J. Amer. Statist. Assoc., **96**, 1205-1214.

Billheimer, D., Guutorp, P. and Fagan, F. (2001)
Statistical interpretation of species composition.
Journal of the American Statistical Association, 96(456): 1205-1214.

Billheimer, D., Cardoso, T., Freeman, E. Guttorp, P., Ko, H. and Silkey, M. (1997)
Natural variability of benthic species in the Delaware Bay.
Journal of environmental and ecological statistics, 4:95-115.

Billheimer, D. and Guttorp, P. (1995)
Spatial statistical models for discrete compositional data.
Technical report, University of Washington, Seattle, Department of Statistics.

Brooks, S.P. & Gelman, A. (1998)
General methods for monitoring convergence of iterative simulations.
Journal of Computational and Graphical Statistics, **7**, 434-455

Coakley, J.P. and Rust, B.R. (1968)
Sedimentation in an Arctic lake.
J. Sed. Petrology, **38**, 1290-1300.

Cressie, N.A.C. (1993)
Statistics for Spatial Data (revised edition), John Wiley & Sons (New York).

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

Friedland, L., Joffe, M., Moore, D., et al. (1992)
Effect of Educational Program on Compliance With Glove Use in a Pediatric Emergency Department.
American Journal of Diseases of Childhood, **146**, 1355-1358.

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

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995)
Bayesian data analysis, Chapman & Hall/CRC.

Haining, R.P. (1990)
Spatial Data Analysis in the Social and Environmental Sciences, Cambridge University Press.

Legendre, P. (1993)
Spatial autocorrelation: Trouble or new paradigm ?
Ecology, **74**, 1659-1673.

Link, W.A., Cam, E., Nichols, J.D. & Cooch, E.G. (2002)
Of Bugs and Birds: Markov Cain Monte Carlo for hierarchical modeling in wildlife

research.
Journal of Wildlife Management, **66**, 277-291.

McCullagh, P. and Nelder, J.A. (1989)
Generalized Linear Models, Chapman & Hall (London).

Philip, G.M. and Watson, D.F. (1988)
Angles measure compositional differences.
Geology, **16**, 976-979.

Smith, P.G. (2004)
Automated log-ratio analysis of compositional data: software suited to analysis of habitat preference from radio tracking data.
Bat Research News, **45**(1), 16.

Spiegelhalter, D.J., Thomas, A, Best, N. and W.R. Gilks (1999)
WinBUGS Version 1.2 User Manual. MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK

Thomas, A., Best, N., Lunn, D., Arnold, R. and Spiegelhalter, R. (2004)
GeoBUGS User manual, http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/geobugs.shtml

Thompson, R.N., Esson, J. and Duncan, A.C. (1972)
Major element chemical variation in the Eocene lavas of the Isle of Skye, Scotland.
J. Petrology , **13**, 219-253.

Tocher, J.F. (1908)
Pigmentation survey of school children in Scotland.
Biometrika, **6**, 129-235.

Watson, D.F. and Philip, G.M. (1989)
Measures of variability for geological data.
Math. Geol., **21**(2), 233-254.

APPENDIX A: DATA SETS

The following data sets are available online at:
http://www.bioss.ac.uk/~adam/alarm_training/datasets.shtml

Data file	*Description*	*Course chapters*
arctic.dat	sand/silt/clay composition of arctic sediments	1, 3
bimodal.dat	simulated data exhibiting bimodality	5.3
flowers.dat	simulated spatial data on flower colours	4
gaelic.dat	% of Gaelic speakers on Scottish islands	1, 2, 3
hair.dat	hair colour of boys in Scotland	1, 2
leave.dat	activity pattern of a statistician	5.1
nurses.dat	glove use among nurses	5.2
skye.dat	mineral composition of Skye lavas	1

APPENDIX B: DETAILS OF SIMUALTION OF UK FLOWER COLOURS

A simulated dataset, demonstrating the use of a conditionally autoregressive model to analyse spatial ecological data. For each of 2523 10-by-10km grid cells, the following data are given: x and y coordinates in km from a basepoint ("x" and "y"), value of the limestone index ("limestone"), value of the biodiversity index ("biodiversity"), proportions of flowers having the colours red, blue and yellow ("red", "blue" and "yellow").

Algorithm

Let L_i denote the latitude of grid cell i.

Simulate the biodiversity index z_{1i} from a normal distribution, $N(2.5,0.1^2)$, where $i = 1,...,2523$.

Simulate the limestone index z_{2i} from a normal distribution, $N(\sin(L_i / 250), 0.25^2)$, where $i = 1,...,2523$.

Simulate the proportion of red flowers X_i^1 from a normal distribution, N(z_{1i} / 4, 0.035^2), where i = 1,...,2523.

Adjust the proportion of red flowers : if less than 0 then set to 0.01, if greater than 1 then set to 0.99

Adjust the proportion of red flowers again : set $X_i^1 = X_i^1 - z_{2i} / 10$

Simulate the proportion of blue flowers X_i^2 from a normal distribution, N(z_{1i} / 10, 0.035^2), where i = 1,...,2523.

Adjust the proportion of blue flowers : set $X_i^2 = X_i^2 - z_{2i} / 20$

Adjust the proportion of blue flowers again : if less than 0 then set to 0.01, if greater than 1 then set to 0.99

Calculate the proportion of yellow flowers X_i^3, by deducting the proportions of red and blue flowers from one.

Further details of the R-code used to generate the data can be found at (http://www.bioss.ac.uk/~adam/alarm_training/R/flowers_simulate.R)

APPENDIX C: WINBUGS MODEL USED TO FIT THE BAYESIAN CONDITIONAL AUTOREGRESSIVE MODEL

For more details, such as initial values, and the data sets see:
http://www.bioss.ac.uk/~adam/alarm_training/com45.shtml

```
model

      {

      for (g in 1:N) {

            LR1[g] <- LR[g,1];

            LR2[g] <- LR[g,2];

      }

      for (i in 1 : N) {

            LR[i,1:2] ~ dmnorm(mu[i,], T[,]);

            for ( j in 1:2 ) {

            mu[i,j] <- a0[j] + a1[j] * expl[i] + b[j,i];
            }
      }

      b[1:2,1:N] ~ mv.car(adj[], weights[], num[], tau[,]);

      for ( q in 1:N ) {
            b1[q] <- b[1,q];
            b2[q] <- b[2,q];
      }
      for ( k in 1:sumNumNeigh) { weights[k] <- 1; }

      tau[1:2,1:2] ~ dwish(O[,],2);
      O[1, 1] <- 0.1;
      O[1, 2] <- 0;
      O[2, 1] <- 0;
      O[2, 2] <- 0.1;

      for ( l in 1:2 ) {
            a0[l] ~ dflat();
            a1[l] ~ dnorm(0,0.0001);
        }

      T[1 : 2 , 1 : 2]  ~ dwish(R[ , ], 2);
      R[1, 1] <- 0.1;
      R[1, 2] <- 0;
      R[2, 1] <- 0;
      R[2, 2] <- 0.1;
}
```