Bayesian Stable Isotope Mixing Models

Andrew C. Parnell∗1, Donald L. Phillips†2, Stuart Bearhop†3, Brice X. Semmens†4, Eric J. Ward†5, Jonathan W. Moore†6, Andrew L. Jackson†7, and Richard Inger†3

1School of Mathematical Sciences (Statistics), Complex and Adaptive Systems Laboratory, University College Dublin, Ireland
2U.S. Environmental Protection Agency, National Health & Environmental Effects Research Laboratory, Oregon, USA
3Centre for Ecology and Conservation, School of Biosciences, University of Exeter, UK
4Scripps Institution of Oceanography, University of California, San Diego, 9500 Gilman Drive, La Jolla, California, USA
5Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, USA
6Earth2Ocean Research Group, Simon Fraser University, Burnaby, Canada
7School of Natural Sciences, Trinity College Dublin, Ireland

October 1, 2012

Abstract

In this paper we review recent advances in Stable Isotope Mixing Models (SIMMs) and place them into an over-arching Bayesian statistical framework which allows for several useful extensions. SIMMs are used to quantify the proportional contributions of various sources to a mixture. The most widely used application is quantifying the diet of organisms based on the food sources they have been observed to consume. At the centre of the multivariate statistical model we propose is a compositional mixture of the food sources corrected for various metabolic factors. The compositional component of our model is based on the isometric log ratio (ilr) transform of Egozcue et al. (2003). Through this transform we can apply a range of time series and non-parametric smoothing relationships. We illustrate our models with 3 case studies based on real animal dietary behaviour.

1 Introduction

Stable isotope analysis is an increasingly important tool in the study of ecological food webs. The technique utilises the fact that biologically active elements exist in more than one isotopic form. Generally the lighter isotopic form is much more abundant in the environment than the heavier form, although their relative abundance is altered by a range of biological, geochemical and anthropogenic processes. These processes produce isotopic gradients which are reflected in the tissues of plants and animals. Differences in relative abundance of these isotopes within a particular sample can be measured using a mass spectrometer and expressed as the ratio of heavy to light form, which can then be standardised against international reference samples and reported in the delta (δ) notation as parts per thousand or per mil (‰).
As a consumer’s tissues are ultimately derived from the dietary sources they consume, it is possible to use stable isotope mixing models (SIMMs) to derive the assimilated diet of an individual, or a group of individuals, given the isotopic ratios of the consumers’ tissues and food sources (Phillips, 2012). A number of recent papers have proposed models to analyse such data, gathering over 1500 citations since their first introduction. More recently, the models proposed for such data are Bayesian. In this paper we review the different models proposed and bring them into an over-arching framework. We include three case studies ranging from the simple to the complex, together with JAGS (Just Another Gibbs Sampler; Plummer, 2003) code for their implementation.

Generally the isotopic ratios of a sample of a consumer’s tissues (e.g. blood, feathers, whiskers) are measured along with a representative sample of potential items from a consumer’s diet. The consumer isotopic values are represented in the model as the convex combination of the source values where the coefficients in the simplex are the ‘dietary proportions’; strictly speaking they are the proportion of the consumers’ dietary proteins obtained from the sources. Estimation of these dietary proportions is the main focus of our analysis. Most commonly the isotopic observations on the consumers and sources are multivariate with dimension 2. A thorough description of the uses of stable isotopes can be found in Inger and Bearhop (2008).

Once the isotopic data have been collected for both consumer and sources, it is usual to create an iso-space plot which shows the consumer and source values. An example is shown in Figure 1. It is desirable for the consumer values to lie within the fuzzy convex hull of the sources. However, a further phenomenon is often observed here, that of trophic enrichment, whereby light isotopes are lost during the conversion of source proteins into consumer tissues. The isotopic values of the consumer (or equivalently the sources) are thus adjusted by a trophic enrichment factor (TEF) which may vary by food source and consumer. These TEF corrections arise from laboratory studies, and thus contribute another set of (uncertain) data to our analysis.

The inference challenge involved in a SIMM is thus to estimate the dietary proportions whilst taking account of the uncertainty in the source and TEF values. Clearly not all of the consumers will eat exactly the same diet, so it is common to use a hierarchical model. Furthermore, covariates such as time or age may be available which are thought to influence the dietary proportions. The model we present in this paper takes account of these common features in a multivariate hierarchical Bayesian model.

The paper is organised as follows. In Section 2 we introduce our general SIMM and show it may be extended to form part of a larger class of hierarchical compositional generalised linear models. In Section 3 we outline previous work in this area and show how each of these fits into our general SIMM. Section 4 outlines the statistical issues concerning the formulation of such models and how they can be fitted. We outline three case studies in Section 5 showing how the model works in different situations depending on the information available. We discuss future directions in Section 6.

### 2 Model formulation

We first provide the notation we use to formulate our model. We suppose that there are \( N \) consumers on \( J \) isotopes, with \( K \) sources. We outline the most important components of our model below:

- \( Y_{ij} \) represents the isotope measurement on consumer \( i \) for isotope \( j \). We write \( Y_i \) as the \( J \)-vector of isotope values for consumer \( i \)

---

1[See mathsci.ucd.ie/~parnell_a/](http://mathsci.ucd.ie/~parnell_a/)
Using the notation above, we write our general model as:

$$Y_i \sim N(p_i^T(s_i + c_i), \Sigma).$$  \hspace{1cm} (1)

We assume a hierarchical formulation so that \( s_{ik} \sim N(\mu^s_k, \Sigma^s_k) \) and \( c_{ik} \sim N(\mu^c_k, \Sigma^c_k) \) where the means and covariances of the sources and TEFs are estimated from the source and TEF experimental data.

Of particular interest is the modelling structure for the dietary proportions \( p_i \). We use an isometric log-ratio (ilr) approach as proposed by Egozcue et al. (2003), though other transformations are available (see next two sections for further discussion). The transformation is written as:

$$\phi_i = \text{ilr}(p_i) = V^T \log \left[ \frac{p_{i1}}{g(p_i)}, \ldots, \frac{p_{iK}}{g(p_i)} \right] \text{ with } g(p_i) = \left( \frac{1}{K} \prod_{i=1}^{K} p_{ik} \right)^{1/K}. \hspace{1cm} (2)$$
with $V$ a $K - 1 \times K$ matrix of orthonormal basis functions on the simplex. The inverse transformation $p_i = \text{ilr}^{-1}(\phi_i)$ simply involves exponentiating and re-normalising the values. There are two consequences of working with the ilr. The first is that we now work in a $K - 1$ dimensional space. The second is that there is no obvious link between the elements of $\phi_{ik}$ and $p_{ik}$, so we lose some degree of interpretability. We further parameterise the transformed proportions so that $\phi_{ik} \sim N(\gamma_{ik}, \kappa_k)$ with $\kappa_k$ quantifying a random effect variance. Were we to use the centred log-ratio (clr) transform (equivalent to setting $V = I$; though see Section 4 as to why we might not do this) then $\kappa_k$ would represent the consumer-level variance in diet for source $k$. Finally we set $\gamma_{ik}$ to be a mean term restricted in some fashion to allow for estimation. A multivariate prior for $\phi_i$ would also be feasible.

In situations where covariates $x_i$ are available, they are usually linked to the model through the dietary proportions. The covariates may take the form of age, sex, time or any other variables upon which diet is expected to depend. In certain cases (such as our third case study) both the diet and the sources are expected to be functions of a covariate. In the simpler case we apply the covariates by making $\gamma_{ik}$ functions of $x_i$. In the more advanced case we additionally apply them to $\mu_k$ and $\Sigma_k$.

We term the TEF data set $D_c$ and the source data set $D_s$, each consisting of collections of $J$-vectors of isotope ratios for each of the $K$ sources. A full Bayesian posterior distribution gives:

$$
\pi(p, \phi, \gamma, \kappa, \Sigma, \Sigma_s, \Sigma_c, \mu_s, \mu_c, s, c | Y, x, D_c, D_s) \propto \left[ \prod_{i=1}^{N} \pi(Y_i | p_i, s_i, c_i, \Sigma) \right] \times \left[ \prod_{i=1}^{N} \pi(\phi_i | \gamma_i, x_i, \kappa) \right] \\
\times \left[ \prod_{i=1}^{N} \prod_{k=1}^{K} \pi(s_{ik} | \mu^s_k, \Sigma^s_k) \right] \times \left[ \prod_{i=1}^{N} \prod_{k=1}^{K} \pi(c_{ik} | \mu^c_k, \Sigma^c_k) \right] \\
\times \left[ \prod_{k=1}^{K} \pi(\mu^s_k, \Sigma^s_k | D_s) \right] \times \left[ \prod_{k=1}^{K} \pi(\mu^c_k, \Sigma^c_k | D_c) \right] \\
\times \left[ \prod_{k=1}^{K} \pi(\kappa_k) \right] \times \pi(\Sigma)
$$

(3)

A Directed Acyclic Graph (DAG) is shown in Figure 2. The sub-models for the sources and TEFs, namely $\pi(\mu^s_k, \Sigma^s_k | D_s)$ and $\pi(\mu^c_k, \Sigma^c_k | D_c)$ can be updated as part of the modelling steps, but we prefer a more efficient empirical Bayes approach (Carlin and Louis 2000) whereby we approximate $\mu^s_k, \mu^c_k, \Sigma^s_k, \Sigma^c_k$ by their sample estimates, thus cutting feedback and removing these from the updates and the posterior. Note that this has minimal effect on the source and TEF random effects $s_{ik}$ and $c_{ik}$ which are still updated as part of the modelling process. The prior distributions we give to $\Sigma$ and $\kappa_k$ are vague Inverse-Wishart and Inverse-Gamma respectively.

### 3 Previous work

The earliest attempts at applying a mixing model framework to stable isotope data did not use probability as the basis for estimation. Initially, SIMMs were restricted to systems involving a single consumer (or the mean of multiple consumers), and where the number of isotopes and sources was arranged such that $J + 1 = K$. Such an arrangement yields a linear system with a single solution. [Phillips and Gregg (2001)] provided propagation of error calculations for such a system in their IsoError model to establish confidence intervals around the estimates based on the variances of the consumer and source isotopic measurements. The method was expanded upon in IsoSource (Phillips and Gregg 2003) to relax the $J + 1 = K$ restriction and allow for multiple sources but without explicit incorporation of source and consumer variability. IsoSource works by simulating values of the dietary proportions $p$ on a grid to produce multiple valid solutions to the
linear system. The valid solutions could be plotted in a histogram-like fashion, though they did not represent probability distributions, rather simply the range of values which might be plausible given the geometry of the system.

These initial attempts were formalised in a Bayesian fashion in the models MixSIR (Moore and Semmens, 2008) and SIAR (Parnell et al., 2008). The MixSIR model can be thought of as a simplification of Equation 1 without explicit random effects across the dietary proportions, and with $\Sigma$ set to zero. The sources and TEFs are treated as independent across isotopes and are given fixed values for their mean and variance. The dietary proportions are given independent Beta-distributed priors or, in a later version, a Dirichlet distribution. Since the dietary proportions are the only parameters in the model, it can be fitted extremely efficiently using Importance Resampling (e.g. Robert and Casella, 2005) on a grid encompassing the range of proportion values. Updated versions of the MixSIR model have included random effects in the dietary proportions through the clr transform, and also have allowed for hierarchical models to be fitted, most elegantly in capturing familial relationships affecting the diet of gray wolves in British Columbia (Semmens et al., 2009). These latter advanced versions of MixSIR are fitted using the JAGS software (Plummer, 2003).

The SIAR model (Parnell et al., 2010) is in many ways similar to the basic MixSIR model (and thus still a simplification of Equation 1) though includes a residual component which is treated as independent between isotopes and given an Inverse Gamma prior. The model also allows for concentration dependencies ($q_{kj}$, vectorised as $q$) which quantify the proportion of the isotope in the given food source (Phillips and Koch, 2002). They can be added in to our model by replacing $p_i$ in Equation 1 with $C(p_i \oplus q)$ where $C$ is the simplex closure operator and $\oplus$ is the simplex perturbation (Egozcue et al., 2003). Usually $q$ are either fixed or given a suitably informative prior distribution. The SIAR model is fitted using standard MCMC with Metropolis-Hastings steps to update the dietary proportions.
More recently, the IsotopeR model has been introduced which extends the SIAR/MixSIR models to a multivariate setting (both sources and TEFs are multivariate normal) and partitions the residual covariance $\Sigma$ into a mass spectrometer calibration error and that of residual error. They further allow for the sources, TEFs and dietary proportions to be random effects with the latter obtained through the clr transform (see next section for further discussion). The model is fitted in JAGS (Plummer 2003) but does not allow for covariate information or for the estimation of the dietary random effect variance.

Aside from explicit SIMM model development, recent focus has also been on the performance of SIMMs in non-ideal conditions, most notably with respect to the characterisation of source and TEF values by Bond and Diamond (2011). Clearly it is absolutely vital that food sources are not excluded from the model as they will yield biased dietary proportions. Similarly, the estimation of the TEF values must be conducted appropriately or there will be some extra uncertainty in the estimated dietary proportions. In related work, Ward et al. (2011) considered the problem of (dis)aggregating sources and its effect on the resulting estimates.

For example, if a consumer only eats part of a food source it may be hard to obtain isotopic values from just that part. Similarly if two sources, though different species, lie in the same location in iso-space it may be impossible for the model to determine the difference in their dietary consumptions. Thus on occasion it may be pertinent to aggregate sources without any loss of information. Alternatively the aggregation can be accomplished with fewer assumptions if it is calculated a posteriori by combining the negatively correlated dietary proportions.

Lastly, it should be noted that there are strong connections between SIMMs and the ‘end member’ analysis used by geologists to determine the composition of e.g. river sediment. In such cases the sources are usually known with minimal error and the challenge is to estimate the proportional contributions of different sediment sources. A number of different methods have been proposed, e.g. Soulsby et al. (2003); Brewer et al. (2011) (and references therein). Ours most closely resembles that of Palmer and Douglas (2008), though they use the alr transformed proportions (see below for definition) which are given a spatial prior distribution.

4 Statistical issues in SIMMs

The models we fit use ideas from Bayesian hierarchical modelling (e.g. Gelman et al. 2003) and compositional data analysis (Aitchison 1986). BHM is now part of the standard toolbox of Bayesian statistics and we do not discuss them further here. Of more interest however is the compositional structure applied to the dietary proportions as this can strongly affect the behaviour of the posterior distribution. A recent review of the state of compositional data analysis can be found in Pawlowsky-Glahn and Buccianti (2011). Our models differ fundamentally from many standard problems as our compositions are not observed directly but are latent parameters to be estimated, constrained by their geometrical position in iso-space and the covariates upon which they may depend.

The starting point for the early Bayesian SIMMs models was that of the Dirichlet distribution, being perhaps the simplest valid distribution on the simplex. The usual prior distributions used were either flat where all Dirichlet parameters are set to 1 or the Jeffreys prior where all are set to $1/K$. Unfortunately, as is well known (e.g. Aitchison 1986), the Dirichlet distribution suffers from a very rigid sub-compositional independence assumption. This is not necessarily a problem when used as a prior distribution with fixed components as the posterior distribution may well show interesting sub-compositional properties. However, if dependence is to be modelled through hyper-parameters of the Dirichlet distribution (e.g. with covariates) this restriction will remain in the posterior.

A number of extensions to the Dirichlet have been proposed (e.g. Wong 1998), but we focus here on the logistic-normal transformations of Aitchison (1986) and Egozcue et al. (2003), through which more flexible sub-compositional dependence can be obtained. The simplest of these is perhaps the additive log-ratio (alr),
where \( g(p_i) \) in Equation 2 is set to be one of the chosen proportions (which is then removed from the composition) and \( V = I \). However, this can perform poorly when there is no obvious choice of denominator and is not permutation-invariant. The centred log-ratio (clr; defined in Equation 2 when \( V = I \)) removes the need for a choice of denominator in the log ratio but produces a covariate matrix of rank \( K - 1 \). It is also not sub-compositionally coherent (see Pawlowsky-Glahn and Buccianti, 2011 for further discussion of these terms). Finally the isometric log-ratio (ilr) uses orthonormal basis functions in the simplex to obtain coordinates that are isometric and satisfy the usual compositional requirements of coherence and permutation/subcomposition invariance. The choice of basis functions is somewhat subjective; we follow the method in Egozcue et al. (2003). From a Bayesian perspective, the latent compositional parameters are more easily identifiable when working with the ilr.

Once a suitable transform has been chosen, it is feasible to include covariates or perform any of the traditional multivariate analysis techniques. Numerous examples can be found, including a geo-statistical framework (Tolosana-Delgado et al., 2011), discrete time series (Barceló-Vidal et al., 2011), or spectral methods (Pardo-Igúzquiza and Heredia, 2011). A popular topic is that of zero compositions or zero-inflation (e.g. Butler and Glasbey, 2008) which is not a severe issue in SIMMs because we know from experimental observation that all dietary sources are consumed.

5 Case studies

We now present three case studies and show how our general model can be used in each of the scenarios. In the first case study we analyse the diet of a small sample of geese from data previously studied by Inger et al. (2006). This first case study uses the model as proposed in Section 2 and can be seen as a small extension to that of Hopkins and Ferguson (2012). The second case study extends the geese model to allow for the inclusion of covariates or basis function models. Our final case study includes a compositional time series component where the sources and consumers are observed at different time points. In this final case study the data are swallows consuming chironomid midges, other freshwater invertebrates and terrestrial invertebrates. In all cases we run the models using the JAGS software and check convergence using the coda package (Plummer et al., 2006) and the Brooks-Gelman-Rubin diagnostic (Gelman and Rubin., 1992; Brooks and Gelman, 1998).

5.1 Case study 1

Our first case study contains 9 (\( \delta^{13}C, \delta^{15}N \)) pairs of isotopic values taken from the blood plasma of Brent geese sampled on 26th October 2003. The food sources are Zostera spp, terrestrial grasses, Ulva lactuca and Enteromorpha spp. Our empirical Bayes estimates for the sources are given in Table 1. The TEFs were taken from values in the literature (see references in Inger et al., 2006 for more details) so that \( \mu_c^k = (1.63, 3.54)^T \) and \( \Sigma_c^k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \) for all \( k \). For a simple random effects formulation we set \( \gamma_{ik} = 0 \) over all \( i \) and \( k \). An iso-space plot for the data (where the sources have been corrected by the TEFs) is shown in Figure 1. The JAGS model was run for 3 chains over 50,000 iterations, removing 10,000 for burn-in and thinning by a factor of 20. A density plot of the mean dietary proportions is shown in the left panel of Figure 3. They can be seen to compare favourably with the simpler SIAR model (see the function siardemo in Parnell et al., 2008), though in this case we have extra information covering individual dietary estimates, as well as improved estimation of the source and TEF random effects. In particular, the flexibility of the hierarchical formulation through the ilr transform allows for some multi-modality in the posterior distributions. The right panel of Figure 3 shows a matrix plot of the joint behaviour of the dietary proportions. These can be useful in determining unavoidable model inadequacy, for example when it is impossible to ascertain which food sources are being
consumed together. A strong negative correlation indicates that the food sources are indistinguishable. For example the strong negative correlation between Zostera and Ulva lactuca indicates that, whilst it is clear they are consuming mainly Zostera, the balance between the two cannot be exactly determined.

<table>
<thead>
<tr>
<th>Source</th>
<th>Enteromorpha</th>
<th>Terr Grasses</th>
<th>Ulva lactuca</th>
<th>Zostera</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_k^s$</td>
<td>$(-14.06, 9.82)^T$</td>
<td>$(-30.88, 4.43)^T$</td>
<td>$(-11.17, 11.2)^T$</td>
<td>$(-11.17, 6.45)^T$</td>
</tr>
<tr>
<td>$\Sigma_k^s$</td>
<td>$\begin{bmatrix} 1.37 &amp; 0 \ 0 &amp; 1.37 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 0.41 &amp; 0 \ 0 &amp; 5.15 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 3.83 &amp; 0.85 \ 0.85 &amp; 1.24 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 1.48 &amp; -0.56 \ -0.56 &amp; 2.16 \end{bmatrix}$</td>
</tr>
</tbody>
</table>

Table 1: Estimates of the source means and covariance matrices for the geese data of case study 1.

There are many other useful statistics we can calculate here quite simply from the posterior distributions. In particular, we can focus on individual level variation by calculating, for example, the probability that an individual consumes more Zostera than another. Similarly with access to the variance parameters $\kappa_k$ we can determine whether there is more variation amongst consumers in some sources rather than others. Lastly, it is often desirable to perform model comparison diagnostics to determine whether certain parts of the model can be removed without a detrimental effect on prediction. However, we do not perform any further analysis on this data set, preferring instead to use these tools when analysing the more sophisticated data below.

5.2 Case study 2

We now extend our Goose model to a larger data set of 248 observations collected over the period of October 2003 to April 2005 of which the previous case study was just a small part. The sources and TEFs are believed
to be stable over the course of the study so the source means and covariances are the same as Table 1. The diet of the geese will however vary during the season due to variations in abundance of food sources along with social and demographic factors. An iso-space plot for the full data is shown in Figure 4. The iso-space plot appears to show that the diet during October to be focussed mainly on Zostera, moving on to Ulva lactuca and/or Enteromorpha during November/December. In January and February the diet appears to be relatively mixed, but focussing almost solely on Terrestrial grasses around April. In addition to the stable isotope information we have covariates which state the Goose’s sex and whether they are juvenile or adult.

Figure 4: Isospace plot of the full goose data set where the consumers are labelled by month. The data in case study 1 correspond to the data shown for month 10. Note that the sources and the TEFs are unchanged from Figure 1.

We consider 6 possible models for the dietary behaviour, accounting for the covariates. In each case we set $\gamma_{ik} = X_i^T \beta_k$ where $X_i$ is an $L$-vector of covariates or basis functions for consumer $i$, and $\beta_k$ is an $L$-vector of regression parameters associated with $\phi_k$. Note that, when using the ilr transform, the parameters $\beta_k$ do not have any association with source $k$. With the extra parameters in the model, convergence of the MCMC algorithm is a problem because the parameters are often highly correlated across sources, though this is somewhat lessened with appropriate choice of $V$ in the ilr. We find it helpful to re-parameterise with Helmert contrasts across sources so that $\gamma_{i1} = X_i^T \beta_1$ and $\gamma_{ik} = X_i^T (\beta_1 + \beta_k)$ for $k \geq 2$. We use the first source (Enteromorpha) as $k = 1$ as this seems to be consumed throughout the season though this is obviously not known in advance so does involve some degree of trial and error.

In all cases the parameters $\beta$ are given vaguely informative $N(0, 10)$ distributions as a value of $|\phi|$ in excess of 10 is likely to yield dietary proportions near 100% (though again this depends on the correlation structure in the data set). We compare the different models using the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) and, for the final chosen model, posterior predictive distributions of the data. We use two versions of the DIC, the standard $p_D$ method of Spiegelhalter et al. (2002), and the $p_	ext{f}$ method of Plummer (2008) which estimates the optimism as half the variance of the deviance, and thus penalises complex models more harshly. Due to the extra complexity in the models, we run them with 3 chains for 200,000 iterations,
removing 20,000 for burn-in and thinning by 90.

The first model we try involves no covariates and is thus the same as that in Section 5.1. The second model includes an intercept term and time as a simple linear covariate. The third model replaces linear time with a single harmonic component so that \( X^T_i = [1, \cos (\frac{2\pi t_i}{365}), \sin (\frac{2\pi t_i}{365})] \) where \( t_i \) is the Julian day. The fourth, fifth and sixth models are expansions of model 3 to include juvenile/adulthood (model 4), sex (model 5) and also their interaction (model 6) as covariates. Table 2 shows the different models and the associated DIC values. Both versions of DIC seem to prefer models with the harmonic covariate, and also the addition of either sex or adulthood.

Figure 5 shows the relationship between Julian day and dietary proportion for the different sources for model 4. The switch from Zostera to Enteromorpha, to terrestrial grasses is very clearly seen in both juveniles and adults, though the uncertainty in juveniles is slightly higher, especially with respect to Enteromorpha. Figure 6 shows the posterior predictive distribution of the data under model 4. The model seems to predict the data well, the three modes corresponding to the main sampling times of October, February and April.

<table>
<thead>
<tr>
<th>Model</th>
<th>Covariate(s)</th>
<th>DIC (using ( p_V ))</th>
<th>DIC (using ( p_D ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>None</td>
<td>26907</td>
<td>1210.0</td>
</tr>
<tr>
<td>2</td>
<td>Julian day (linear)</td>
<td>16957</td>
<td>524.3</td>
</tr>
<tr>
<td>3</td>
<td>Julian day (harmonic)</td>
<td>16683</td>
<td>385.1</td>
</tr>
<tr>
<td>4</td>
<td>Julian day (harmonic), Juvenile/Adult</td>
<td>7551</td>
<td>393.6</td>
</tr>
<tr>
<td>5</td>
<td>Julian day (harmonic), Sex</td>
<td>8600</td>
<td>382.8</td>
</tr>
<tr>
<td>6</td>
<td>Julian day (harmonic), Juvenile/Adult, Sex, Interaction</td>
<td>10812</td>
<td>382.9</td>
</tr>
</tbody>
</table>

Table 2: Table of models and model selection criteria. The models with the lowest DIC are shown in bold.

Figure 5: Plot of proportion values against Julian day for model 3. The left panel shows estimates for juveniles whilst the right panel shows adults. The solid lines show median estimates, the outer of the polygons show 90% credibility intervals. The Geese appear to focus on Zostera around November before moving on to Enteromorpha and then terrestrial grasses. There is slightly more uncertainty in the juveniles than the adults.
5.3 Case study 3

Our last case study concerns the dietary behaviour of barn swallows (unpublished data). The data are stable isotope ratios of blood plasma samples from birds captured between May and August in 2009. Again, we use Julian day to determine their behaviour over time. However, in this scenario the sources (chironomid midges, other freshwater invertebrates, and terrestrial invertebrates) are also expected to change over time, and may have been observed on different days to that of the swallows. We thus expand our model to include a temporal component on the source vector as well as that of the dietary proportions. In each case, we use P-splines (Eilers and Marx 1996) to explain the suitably flexible behaviour. Other time series methods, such as random walks or Levy processes, may also be appropriate. We write continuous time as \( t \) and may have been observed on different days to that of the swallows. We thus expand our model to include

The sources are now described by a multivariate spline model so that source data pairs, denoted \( s_{jk}(t) \) for the source experimental data at time \( t \) on isotope \( j \) and source \( k \), are given \( N \left( \begin{bmatrix} X^T \beta_1' \\ \vdots \\ X^T \beta_J' \end{bmatrix}^T, \Sigma'(t) \right) \) independently for each source \( k \). The number of observations for each source is likely to be different, and certainly not equal to the number of consumer observations \( N \). Here, the spline parameters \( \beta_j' \) determine the mean behaviour of the sources over time on isotope \( j \). The \( J \times J \) variance matrix \( \Sigma'(t) \) is also allowed to change over time with diagonal elements given log-splines: \( \log(\Sigma_{jj}(t)) \sim N(X^T \beta_{jj}, \kappa_{\Sigma}) \). The cross isotope covariance \( \Sigma_{12} \) is parameterised through a single correlation parameter for each source, denoted \( \rho_k \), and does not change over time. A spline could also be used here (for example on the arctangent of \( \rho_k \)) but this was not found to improve the fit. The source spline model was run for each of the three sources in turn and used to calculate Maximum a Posteriori (MAP) estimates of \( \mu_k(t) \) and \( \Sigma_k(t) \) for all times \( t \) at which consumer...
data were available. An iso-space plot of the swallows data is shown in Figure 7.

Figure 7: Isospace plot of the swallows data. The source spline model has been run to obtain estimates of the source means and covariances throughout the study period. Arrows indicate the direction of movement over time for the sources and the swallows. The Julian day and the 50% standard ellipses are given for Julian days 150, 190 and 220. Note that the chironomids’ δ¹³C values increase over time from the start of the study period. A similar occurrence can be seen in the terrestrial invertebrates. The data are shown as kernel-smoothed estimates of the swallows’ isotope data again over Julian day, starting at 150 and ending on day 220. The swallows can be seen to also increase their δ¹³C values up until around day 180 whereupon the δ¹³C returns towards its original value. This plot should be read in conjunction with Figure 8.

We use the source model predictions in our standard SIMM with the spline formulation on the proportions as outlined above. An alternative model where all parameters are estimated simultaneously would be possible, if rather slow. Instead, we retain the cutting feedback assumption so that the source and dietary proportions are run separately. For both the source and the dietary proportion models we run for 200,000 iterations, removing 20,000 for burn-in and thinning by 90. For both models, we use 25 knots, which seems to cover the flexibility in the data adequately.

Figure 8 shows the posterior dietary proportion estimates over Julian day for the swallows, predicted from the resulting spline parameter estimates. The results clearly indicate that they are feeding on mainly fresh water invertebrates during the early part of the data before concentrating on chironomids around the start of August. Figure 9 shows the predictive distribution of the data under this model. The fit appears satisfactory.

6 Discussion and future directions

The SIMM formulation outlined in Section 2 allows for a rich framework upon which to include a variety of other statistical structures. The basis of such models is a mixture compositional structure applied to the
Figure 8: Plot of dietary proportion values against Julian day for the swallows data of Case Study 3. The solid lines show median estimates, whilst the filled polygons show 90% credibility intervals for each source.

Figure 9: Predictive distribution of the data for the swallows example. The observations are shown as filled circles, whilst the posterior predictive density is represented by the contours.

dietary sources. In the case studies above we have illustrated how some simple regression and smoothing
models might be included. The results seem useful and allow for many interesting findings which would not have otherwise been possible without, e.g. the time series or spline components.

The main challenges in such models are that the source and TEF values are fully and correctly characterised. It is a simple geometrical exercise to verify that if sources are missing from the data set, or that the TEF means and errors are poorly estimated, the dietary proportion estimates will be biased. This bias is compounded by the compositional nature of the problem where, if one dietary proportion is poorly estimated, then the others may well be too. There is no statistical test for missing sources; we rely on the ability of the ecologist to observe the system adequately. This may be particularly important if multiple organisms and sources are to be analysed simultaneously in a dietary network. A single missing source may cause biases across the dietary proportion estimates. Such models are not to be encouraged unless there is very strong evidence that no sources are missing and that TEFs are estimated correctly.

It is reasonably straightforward to expand our approach for other values of $J$ where differing numbers of isotopes may be available for analysis. It should be noted, however, that if $J \geq 3$ it becomes harder to determine model fit, especially as no obvious iso-space plot can be created. The solutions to this problem may lie in closer scrutiny of predictive distributions, and some subjective judgement over the size of the residual covariance matrix $\Sigma$. In such scenarios extra care is required in reporting the output of the SIMM.

There are further opportunities for expansion of the models. Some of these may include:

- The simultaneous inclusion of multiple tissue varieties. It is often the case that different tissues are sampled on the same consumer as they are captured. These differentially replenished tissues will represent the dietary proportions consumed over different periods. For example, whilst blood plasma might represent the immediately sampled food sources, feathers might represent the diet consumed over the previous few months. A long-term data set where multiple tissues are analysed simultaneously may allow for increased precision in the dietary proportions. Alternatively, it may be possible to estimate the time scale over which the tissues are responding.

- Clustering/Mixture models to determine groupings. If there are hidden groupings amongst the organisms it may be possible to discern them using a model-based clustering approach (sensu [Fraley and Raftery, 2002]). Even without such groupings, increased flexibility can be obtained by using mixtures of Gaussian distributions to model non-parametric behaviour.

- Long-tailed multivariate distributions to account for outliers or small sample sizes. Clearly where sample sizes are small the multivariate Gaussian assumption (especially for sources) may be invalid, and thus heavier-tailed distributions may be required. One such which seems to be most easily fitted is the Multivariate Normal-Inverse Gaussian (MNIG [Barndorff-Nielsen, 1997]) which has been used previously in clustering and financial settings.

These are just three of the active areas of research to which SIMMs are being applied by our group. Many others are likely to appear as the field advances. We hope to report soon on these exciting new developments.

References


