UNIVERSITY OF LEEDS

Prior Distributions for Stochastic Matrices

Anastasia Frantsuzova

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

University of Leeds

Faculty of Engineering and Physical Sciences

School of Mathematics

August 2021

Intellectual Property

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

My gratitude goes to my supervisors, John Paul Gosling, Robert Aykroyd and John Kent, for their advice and guidance, the EPSRC and the School of Mathematics for the financial support. I would also like to thank Andrew Crowe and Paul Brown for the interesting collaboration opportunity and insightful discussions, and Javier Palarea-Albaladejo for introducing me to the CoDA community. I am finally grateful to my family and Andrei for their love and support, to my colleagues at St Andrews for accommodating the write-up period, to the colleagues and friends I made at Leeds and to friends who have been with me through the years.

Summary

Right-stochastic matrices are used in the modelling of discrete-time Markov processes, with a property that the matrix elements are non-negative and each row sums to one. If we consider the problem of estimating these probabilities from a Bayesian standpoint, we are interested in constructing sensible probability distributions that can be used to encapsulate expert beliefs about such structures before any data is observed. Through the process of expert elicitation, this uncertainty can be represented in terms of probability distributions. In this thesis, we explore multivariate distributions on the simplex support from the view of expert elicitation. We explore properties and constraints of these distributions, and ways to elicit expert judgement about their parameters. This is interesting both mathematically and from a practical standpoint, particularly where there are many such variables to explore, which can prove cognitively challenging and tiring for the experts.

Similarly, data representing proportions of a whole can be unified into the compositional framework (Aitchison, 1986) with similar non-negativity and unit-sum properties. This thesis also explores the study of compositional data analysis, its problems and modern ways of approaching them. Application of these methods is found in exploring how high resolution imagery obtained over rural areas could be used in order to identify the distribution of tree species found in those areas where monitoring is prohibited.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A key step in Bayesian analysis is the identification and construction of a prior probability distribution. Through the Bayes rule, the prior distribution is combined with the likelihood function to form a posterior distribution, and the latter encompasses the probability of some hypothesis after evidence has been observed. A prior probability distribution is called non-informative if it gives equal consideration (coverage) to all the regions of the probability space. A non-informative prior distribution carries a lack of knowledge about an unknown quantity θ, asserting equal (or maximally equal) probability for the values θ can take. Though one may argue that a truly non-informative prior is a deceitful concept in itself, since the choice of the distribution reflects some prior knowledge and subjective thought (Goldstein, 2006). If an non-informative prior distribution is used, the Bayesian analysis is dominated by the likelihood, and any decisions based on the analysis are be drawn from the evidence as it occurs. This outcome is equivalent to the maximum likelihood estimation in frequentist statistics.

On the other hand, if a considerable body of knowledge is available for a choice of the prior that can reflect uncertainty about θ, the statistician is faced with the goal of specifying parameter values for this subjective prior distribution. This is especially important when little evidence can be obtained for a novel scien-

tific method without testing to destruction or unethical medical practices, for example. Hence, a lot of emphasis would be placed on existing knowledge about the uncertainty of the parameters of the data-generating process. The exercise of translating someone's beliefs and knowledge about an uncertain quantity into a probability distribution is called elicitation. In different fields of application, elicitation can be understood to mean knowledge extraction (Gavrilova and Andreeva, 2012), which is directed primarily towards qualitative insights on the scientific question, or it can have a very structured and rigorous format, such as the iterative consensus-seeking Sheffield Elicitation Framework (O'Hagan et al., 2006).

In this thesis, one of the questions we address is the quantification of expert knowledge. Three key parties play part in the process of elicitation: the experts are the individuals with considerable insight about the scientific question under investigation. The expert is of interest to the statistician, who is the facilitator of the elicitation exercise. Finally, the decision maker relies on the experts' knowledge and the facilitator's analysis to influence a decision or a direction of research (O'Hagan et al., 2006).

Elicitation possesses numerous nuances in decision-making and psychology. Questions like the choice of experts (Granger, 2014; O'Hagan, 2019*a*), whether the elicitation should be conducted individually or a group consensus sought, not to mention psychological phenomena of unconscious biases and anchoring (O'Hagan et al., 2006) all require careful consideration and planning. The latter are the responsibility of the facilitator and play a key part in ensuring that the exercise is fair and the results obtained are a valid representation of the experts' beliefs. Even though these particular topics are not considered in detail in this thesis, a large body of work is available in statistics and psychology, most notably Shanteau (1992*b*,*a*).

After the goal of an elicitation exercise has been specified and the appropriate

experts identified, the facilitator is now presented with the task of quantifying the experts' beliefs into a model or identifying parameter values that would inform the prior distribution. In this thesis, we concentrate on prior distributions in parametric form. During the elicitation exercise, the facilitator would quantify beliefs elicited from the expert into summaries that would specify particular parameter values for pre-specified distributions deemed most appropriate for the problem. The task of expressing beliefs quantitatively has been approached by cognitive psychology (Moody et al., 1996) and more often than not it is an unreliable and difficult process. Hence, the facilitator would not expect the expert to confidently identify the summary statistics of their believed prior distribution, such as the mean or the standard deviation, let alone talk about any possible ranges of parameter values. It must be kept in mind that, even though the facilitator must have a sound understanding of statistics, the experts are assumed not to. Expression of expert judgement through graphical means or addressing the most frequent occurrence (mode) and the boundary values are often less straining for the expert, but run into issues with availability bias just as frequently (O'Hagan et al., 2006). These methods seem especially popular in medicine and health technology (Soares and Bojke, 2018) and constraints on the time and resources available for an elicitation exercise drive the decision maker to conduct individual interviews or an online-based study, the experts being unable to meet face-to-face. Similar execution of an elicitation exercise can occur if the identified group of experts becomes unmanageably large.

Already we may recognise that the fitting of a distribution to a set of expert opinions is uncertain. Expression of uncertainty about each judgement is area of research formalised as non-parametric prior specification (Gosling, 2008). Alternatively we may look towards identifying a hyper-prior distribution to reflect this uncertainty, as explored by Albert et al. (2012).

As mentioned, the goal of an elicitation exercise is to provide some plausible parameter values that can be used to specify a prior probability distribution. The

parameter values reflect experts' knowledge about a particular scientific question. The focus of this thesis is to explore multivariate prior distributions that express uncertainty about a vector of non-negative proportions that must sum to one. A simple example of this set-up in a Bayesian analysis is where the likelihood is the multinomial count model, and the multivariate Beta distribution (Dirichlet distribution) is the conjugate prior. In this situation, parameters of the Dirichlet distribution are the target parameters to gain insight from the experts. The Dirichlet distribution then expresses uncertainty about event probabilities in the multinomial likelihood. The Dirichlet distribution is one of the possible prior distributions for a multinomial count model, as is explored in Chapter 4.

Statistical analysis of vectors of non-negative proportions that sum to unity is an area in statistics known as compositional data analysis (CoDA). Due to Aitchison's fundamental work in this area in the 1980s and increasing computing power, CoDA has grown from a small topic when analysing geological data to gaining its own place in the study of multivariate statistics. A further interest of this thesis is to combine the study and practice of expert elicitation with compositional data. Hence, we consider constructing prior distributions that reflect uncertainty about a set of proportions. Due to the underlying structure of compositional data - the sum-one constraint and strict positivity of each element, we are faced with eliciting multivariate distributions with an inherent negative correlation structure. This structure may not always be appropriate, for example, where spatial dependencies hold, so we explore several recent developments which allow for flexibility of specification of the prior distribution. This too comes with a shortfall, since such distributions contain more parameters, and hence may require more judgements to be elicited from the experts; a challenge with increasing number of proportions considered. We review methods used to specify this covariance structure as well as ways to elicit the multivariate priors.

Separate consideration of compositional data analysis is given through a collaboration with Fera Science, UK. The use of non-standard techniques for regression

modelling for compositional response data are to be considered through the use of random forest regression, as well as parametric approaches of multivariate regression using Aitchison's log-ratio transformations. Further, we explore variable importance and attempt to reduce dimensionality of the problem using a dimensionality reduction technique called Principal Variables (Cumming and Wooff, 2007) that finds close relation to principal component analysis, but has simpler interpretability. We inspect whether augmentation of proportions into classes has an effect on the model's predictive power and transferability to new data sets. Similarly, we outline parallels in issues for modelling aggregated proportions and elicitation of augmented classes of compositions.

The thesis is organised as follows: Chapter 2 presents a view into compositional data analysis and the particulars important for this thesis. Chapter 3 describes statistical modelling techniques of tree species in a rural environment, which stems from collaboration with Fera Science, UK. This work can be summarised by a multiple regression problem with compositional response and a set of continuous explanatory variables as predictors, and the use of principal variables is conducted to highlight spatial structure in the data set and to reduce computational time. Hereinafter, the thesis concentrates on elicitation of prior distributions to describe uncertainty about a set of non-negative proportions summing to unity. Chapter 4 describes families of multivariate distributions on this simplex space, with particular focus on some of the recent developments of the Dirichlet, as well as distributions suitable for modelling transformed compositional data and graphical approaches, such as copulae. Chapter 5 follows on from this - we explore discrete-time Markov chain transition matrices, which adhere to a similar mathematical structure as compositional data. We look towards modelling uncertainty about the elements of the transition matrix. Chapter 6 considers the topic of expert elicitation as a general overview and to serve as an introduction to Chapter 7, where we look at elicitation methods of prior distributions lying in the simplex space. The thesis concludes with a study that explores suitabil-

ity of an approach that attempts to fit a joint Dirichlet distribution to expert judgements directly, without separating the exercise into explicitly questioning the experts about marginal and conditional components of the distribution. This is motivated by attempting to minimise the number of judgements gained from the experts, without sacrificing the information presented in those judgements.

Chapter 2

Compositional data analysis

The study of compositional data stems from the ideas of Ferrers (1886) and Pearson (1897) in the late nineteenth Century. Pearson's work considered spurious correlation due to the sum-to-unity constraint for a vector of non-negative observations. This nuance was previously disregarded in statistical practice and resulted in unreliable inference. In the 1960s, geologist Felix Chayes urged that spurious correlation and also interpretation of correlation between parts in a geochemical composition is not described in an adequate sense (Chayes, 1960) and this idea was further investigated by scholars in the 1970s (Darroch, 1969; Darroch and Ratcliff, 1978; Miesch, 1969; Kork, 1977).

Aitchison formalised the approach of modelling compositional data in his famous work *The Statistical Analysis of Compositional Data* (Aitchison, 1986). He urged that the compositional components should be transformed using the log-ratio approach, and then the logistic Gaussian distribution can be used to liberate this constrained structure and improve modelling approaches that existed at the time. Aitchison's own definition of a compositional data set reads - *D-part vectors, describing quantitatively the parts of some whole, which carry exclusively relative information between the parts* (Aitchison, 1986).

We can formalise this definition by defining a compositional vector of D parts, $\boldsymbol{x} = (x_1, x_2, ..., x_D)$:

Definition 2.1. *(Component) Vector* $\boldsymbol{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D$ *subject to the constraints* $x_i \geq 0$ *and* $\sum_1^D x_i = 1$ *with each* x_i *referred to as component;* $i = 1, ..., D$.

Definition 2.2. *(Simplex) The simplex* Δ^D *is the sample space such that* $\Delta^D = \{\boldsymbol{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D$ *subject to the constraints* $x_i \geq 0$ *and* $\sum_1^D x_i = 1$, *i=1,...,D}.*

Compositional data sets can be found in fields such as economics (Fry et al., 2000), geology (Aitchison, 1986) and the biological sciences (Li, 2015). A very typical example can be observations of rock constitution in terms of percentages of various chemical elements. Alternatively, we may represent a population's hair or eye colour in terms of proportions or percentages when varying survey efforts take place, and yet some relative comparison is desired. Other instances of compositional vectors can occur when arbitrary totals are imposed by physical constraints, such as measuring instruments. One other example is the number of hours in the day available for varying activities - work, sleep or leisure. Overall, any data set comprising of some whole or falling into a finite number of categories can be transformed into the compositional type through the process of normalisation (Aitchison, 1986), and can then be used to represent relative measures. In a compositional data set, the term 'whole' can be understood to mean any fixed constant that is consistent throughout the entire data set. For example, when considering compositions expressed as percentages it would be 100%, or regarding number of hours in the day available for differing activities - 24 to the nearest hour. More often, however, the constant is equal to 1 and the components of the compositional vector are expressed as fractions or finite decimals.

In wider mathematical context, probability (stochastic) vectors can also be described as compositional, since they adhere to the same mathematical constraints. We consider stochastic vectors and stochastic matrices in Chapter 5 of this thesis.

Similarly, brief consideration is given to most current and yet incomplete research areas in compositional data analysis, such as amalgamation of classes.

Compositional data can be presented through simplistic graphical tools such as pie charts or composite bar charts. Usually, a separate colour or pattern represents a unique category; for instance, eye colour compositions for a population of a particular country. More advanced graphical depictions of compositions, especially if the distribution of compositional parts is of interest, is known a ternary plot (Aitchison, 1986). Also known as a simplex plot or a de Finetti diagram, it is used to visualise three compositional parts in a two-dimensional diagram.

Example: eye colour

The following ternary diagram represents three-component compositions of eye colour synthesised to come from a population of six anonymous countries. The original data may have been count data that have been normalised for ease of relative comparison. Here, data are shown as ready percentages for illustration purposes. The data set consists of eye colours Blue, Brown and Other, where Other is an aggregation of all the eye colour classes except Blue and Brown. This choice was arbitrary to provide a simple example of a three-part composition that can be represented with the ternary plot Figure 2.1.

The sum of the eye colour categories for each of the six countries, represented by black dots, is given by unity or 100% as on the simplex plot. Each point is positioned on the barycentric coordinates and the arrows show direction of each of the three axes. The ternary plot allows us to visualise the distribution of this data set and is a useful tool in visual depiction of compositional clustering problems (Aitchison, 1986). Additional features of the ternary plot are presented in this thesis in later chapters as needed.

As described, the two-dimensional ternary plot can be used as a tool for graphical depiction of a three-part composition, since the dimension of the composition is

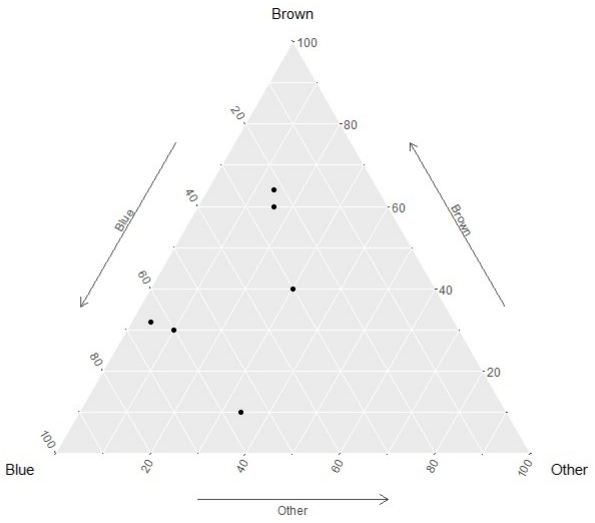


Figure 2.1: Ternary diagram, eye colour example.

$D-1$. Figure 2.2 shows how one may consider a composition with respect to orthogonal axes, the latter representing distinct eye colours. All three-part compositions for eye colours in a particular country are contained inside the two-dimensional equilateral triangle, which joins (100,0,0), (0,100,0) and (0,0,100). In instances where a composition's dimension $D > 3$, one would now consider simplices of higher $D-1$ dimensions with D vertices. Figure 2.3 provides an illustration of higher dimensional simplices.

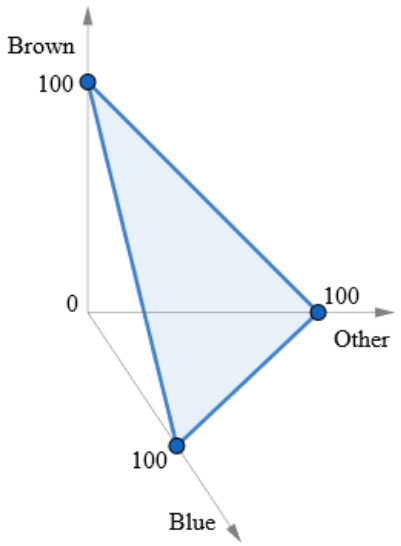


Figure 2.2: 3-part composition plotted in three-dimensional Cartesian coordinates (eye colour example).

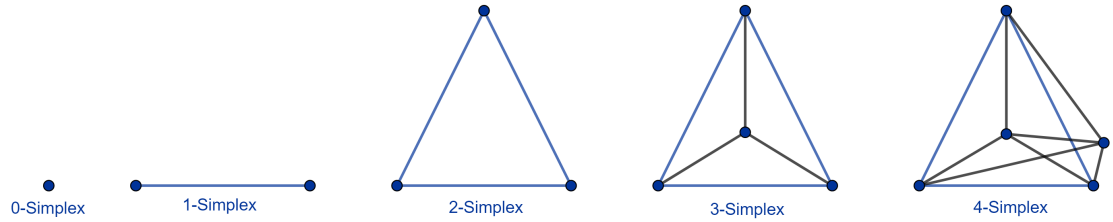


Figure 2.3: Simplices in dimensions 0 to 4.

The 3-simplex is commonly known as a tetrahedron, and the 4-simplex is called a 5-cell. The face of a D-simplex is itself a simplex of dimension $D-1$. It may be obvious to note that depiction of compositions in higher dimensions is of little use to the human eye, although limited insight could be gained from distribution of points in a tetrahedron, provided there are obvious clusters at the vertices or as a central mass in the simplex. Thus, as in the eye colour example, it is common practice to aggregate parts to reduce the dimensionality of the problem. Other approaches akin to principal component analysis (PCA) exist with compositional data (Greenacre, 2018) that tackle high-dimensional compositional data and represent parts as projections onto the Euclidean coordinate plane. The question of representing compositional data is relevant to the topic of expert elicitation, as was outlined in Chapter 1 and is further discussed in the later chapters of this thesis.

In the rest of this chapter, however, we outline the most important concepts in compositional data analysis (CoDA) relevant for this thesis, starting with the work of Aitchison. We explore some mathematical properties and transformations of the compositional structure, as well as higher dimensional representation of the simplex.

2.1 Principles of CoDA

There are three basic principles of compositional data analysis that should be adhered to: scale invariance, subcompositional coherence and permutation invariance.

Scale invariance states that compositional data carry only relative information, so any change of the scale of the original data has no effect. If the original data are multiplied by any scale factor S, for example, a change of units, then the compositional data remain the same after the operation of closure (dividing of raw data by its total to obtain compositional values, which are proportions summing to 1).

Subcompositional coherence is the second principle of CoDA. It implies that any results obtained for a subset of parts of a composition (subcomposition) remain the same as in the composition. For instance, computing the means and variances of parts of a composition does not adhere to this principle. Similar outcome is seen when correlation between parts is computed (Aitchison, 1986).

Permutation invariance is the final principle. It means that results of an analysis do not depend on the order that the parts appear in a composition. Even though in a compositional data set the parts may appear in the same order for each sample taken from this data set, permuting the columns of the data set should be possible without affecting any results derived.

Both scale and permutation invariance appear trivial and accepted principles, however, subcompositional coherence is a principle that often dominates the process of analysing compositional data (Greenacre, 2018).

A fourth principle that has been considered is distributional invariance, which is also one of the principles of correspondence analysis (Greenacre, 2002). This principle relates to amalgamation of parts of a composition. For example, in a chemical setting - the percentage of element A is always a fixed multiple of

element B. Then A and B can be amalgamated, since they essentially contain the same information and the data analysis should not be affected.

2.2 Vector space structure of the simplex

Through Aitchison's work, we can configure the simplex Δ^D as a vector space, which in turn allows to define bases and straight lines. In order to do this, compositional operations known as perturbation and powering of compositions are defined:

Definition 2.3. *(Perturbation) If* $\boldsymbol{x} = [x_1, \ldots, x_D], \boldsymbol{y} = [y_1, \ldots, y_D] \in \Delta^D$ *their perturbation is* $\boldsymbol{x} \oplus \boldsymbol{y} = C[x_1 y_1, x_2 y_2, ..., x_D y_D]$

for some compositional vector C.

Perturbation adheres to the properties of a commutative group operation:

1. internal operation: $\boldsymbol{x} \oplus \boldsymbol{y} \in \Delta^D$

2. commutative: $\boldsymbol{x} \oplus \boldsymbol{y} = \boldsymbol{y} \oplus \boldsymbol{x}$

3. identity element: $\boldsymbol{n} = C[1, 1, ..., 1] = [1/D, 1/D, ..., 1/D]$ such that $\boldsymbol{x} \oplus \boldsymbol{n} = \boldsymbol{n} \oplus \boldsymbol{x} = \boldsymbol{x}$. $\boldsymbol{n}$ is unique.

4. inverse element: given $\boldsymbol{x} \in \Delta^D$ there exists an inverse $\ominus\boldsymbol{x} = C[{x_1}^{-1}, {x_2}^{-1}..., {x_D}^{-1}]$ such that $\boldsymbol{x} \oplus (\ominus\boldsymbol{x}) = \boldsymbol{n}$.

The perturbation operation is equivalent to addition in the real vector space.

Next, let us consider the powering operation in Δ^D which plays the same role as multiplication by scalars in the real vector space.

Definition 2.4. *(Powering) Let* c *be a scalar and* $\boldsymbol{x}$ *is a composition in* Δ^D. *Powering by* c *is defined by* $c \odot \boldsymbol{x} = C[x_1^c, x_2^c, ..., x_D^c]$.

The main properties of the powering operation are:

1. $c \odot \boldsymbol{x} \in \Delta^D$

2. identity element: $1 \in \mathbb{R}$ satisfies $\mathbf{1} \odot \boldsymbol{x} = \boldsymbol{x}$

3. distributive property: $c \odot (\boldsymbol{x} \oplus \boldsymbol{y}) = (c \odot \boldsymbol{x}) \oplus (c \odot \boldsymbol{y})$.

Powering by -1 can be used to define the inverse element: $(-1) \odot \boldsymbol{x} = \ominus \boldsymbol{x}$. Together, powering and perturbation satisfy the properties for Δ^D to have a vector space structure.

Another important notion to aid with enumeration of the compositional elements is that of affine independence:

Definition 2.5. *(Affine independence) Points* $a_1, a_2, ..., a_r \in \mathbb{R}^r$ *are affinely independent if whenever* $\lambda_1 a_1 + ... + \lambda_r a_r = 0$ *with* $\lambda_1 + ... + \lambda_r = 0$ *then* $\lambda_1 = ... = \lambda_r = 0$.

2.3 Higher dimensional simplices

The usual definition of the simplex Δ^D can be regarded from a geometrical perspective to be the D-dimensional polytope which is the convex hull of its $D+1$ vertices. Each D-simplex can be constructed by taking the convex hull of the previous simplex and including one additional point in the D-space such that it is affine-dependent with the preceding simplex. This procedure can be carried out indefinitely, allowing us to define D-simplices for any $D \in \mathbb{N}$.

The simplex is a generalisation of a polytope, which means that a simplex can be decomposed into elements such as its edges, vertices, faces and cells. For instance, an equilateral triangle (2-simplex) has 3 vertices, 3 edges and 1 face. The tetrahedron has 4 vertices, 6 edges, 4 faces and 1 cell.

Definition 2.6. *(d-element) The convex hull of any* $(d+1)$ *vertices of a D-simplex is a d-element of the simplex. For* $d = 0, 1, 2, 3$ *the elements are called vertices, edges, faces and cells respectively.*

Lemma 2.1. *A d-element of a standard D-simplex is itself a d-simplex.*

Proof. A d-element has $(d+1)$ vertices which lie in the $(d+1)$-dimensional

coordinate hyper plane $H \subset \mathbb{R}^{\mathbb{D}+\not\mathbb{F}}$. By omitting the coordinates which take only zero value on the vertices, we can say that $H = \mathbb{R}^{\mathbb{D}+\not\mathbb{F}}$. Now, in H the d-element is the standard d-simplex. □

To proceed, let us define a number which encompasses the elements of the D-simplex

Definition 2.7. *For $D, d \in \mathbb{N}$, let $\Delta(D, d)$ denote the number of d-elements in the D-simplex.*

From previous definition of d-elements, we can see that for a 3-simplex (tetrahedron), $\Delta(3, 0) = 4$ for the number of vertices, $\Delta(3, 1) = 6$ for the number of edges, $\Delta(3, 2) = 4$ for the number of faces and finally $\Delta(3, 3) = 1$ for the number of cells.

We can recognise that the pattern follows the values of Pascal's triangle, and generalises to

Theorem 2.2. *1.* $\Delta(D, 0) = D + 1$.

2. $\Delta(0, d) = 0, d \geq 1$.

3. $\Delta(D, d) = \Delta(D - 1, d) + \Delta(D - 1, d - 1)$.

Proof. 1. The D-simplex has $D + 1$ vertices, from the definition of the simplex.

2. The zero-simplex is a single point in space, so has zero elements of any dimension except the zero-dimension.

3. d-elements of the D-simplex include the d-elements $\in \Delta(D - 1, d)$ and the d-elements which connect the vertex to each $(d-1)$-element of $\Delta(D-1, d-1)$. □

The enumeration of d-elements is concluded with the following theorem

Theorem 2.3. $\Delta(D, d) = \binom{D+1}{d+1}$

Proof. A D-simplex has $D + 1$ vertices. By Definition 2.5 a d-element is the convex hull of any of the $d + 1$ vertices. There exist $\binom{D+1}{d+1}$ ways to choose these

vertices and no two choices should have the same convex hull due to the points being affine-independent. □

2.4 Variance-covariance structure

From earlier remarks of Pearson (1897) and more formally through Aitchison (1982), several issues arise when defining the covariance structure of a composition $\boldsymbol{x} = (x_1, \ldots, x_D)$.

The sum constraint necessarily drives one of the covariances in each row of the variance-covariance matrix to be negative (Aitchison, 1986). This is due to $\text{cov}(x_1, \ldots, x_D) = 0$, so then $\text{cov}(x_1, x_2) + \text{cov}(x_1, x_3) + \text{cov}(x_1, x_D) = -\text{var}(x_1)$, if x_1 is not a constant. This naturally has a restriction on the correlation coefficient. For example, when $D = 2$ the correlation matrix is of the form

$$\text{corr}(x_1, x_2) = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

The second issue, although it receives less attention in applications of CoDA, comes from subcompositional analysis where $\boldsymbol{x}^* = (x^{(1)}, \ldots, x^{(k)})$ is a subvector of k of D parts of $\boldsymbol{x}$. The aim is for any analyis on subcompositions to give a covariance structure that is relatable to the covariance structure of the full composition $\boldsymbol{x}$ (Aitchison, 1986). This nuance too is important when considering possible transformations of data lying in the simplex space onto the Euclidean space in order to be able to employ standard multivariate modelling techniques.

2.5 Log-ratio transformations of compositions

For many practical aspects of compositional data analysis consideration of parts of a whole in their original form can be restrictive in view of accommodating different distributional properties and adhering to assumptions made in modelling.

The most obvious of such being the independence property, which is immediately violated by the unit-sum constraint. One other instance is the violation of the Normality assumption in a linear regression problem with compositional responses (Aitchison, 1986). Also, the principle of sub-compositional coherence dictates that the results in a sub-composition must be the same as those when considering the larger composition, which is not true if we consider solely the original non-transformed parts. However, we find that ratios of parts do adhere to this principle. For example, in two separate chemical studies it would be possible to compare ratios of components which are common to both studies. Usual analysis and even hypothesis testing can be conducted on the ratio values where normality is assumed (Greenacre, 2018). However, it is evident to see that a ratio of non-negative parts cannot be negative, which could introduce problems if the scientist is constructing a confidence interval for the mean, for instance.

In the question of regression analysis, compositional data can form the explanatory variable of interest, the response or both. In any case, as discovered by many statisticians over the course of the last century, it is unwise to model this type of data keeping in mind the usual Gaussian assumptions, such as independence of errors (Aitchison, 1986). This holds even in the instance of multivariate analysis, which relies on the assumption of multivariate normality. If a compositional data set is made up of D components, we only need to know $D-1$ of them to be able to deduce the remaining component, giving some room for measurement error. This brings about the idea of within-component correlation, as Pearson (1897) pointed out in his work *'On a form of spurious correlation which may arise when indices are used in the measurement of organs'*. Working in the usual classical framework, we are assuming the Euclidean-geometric setting in real space, and this is not so suitable for compositional data. To see this, it suffices to consider about how a percentage change 1% to 2% carries different information than percentage change from 91% to 92% percent. In the first instance, as well as an increase of 1%, the original proportion has doubled in size. The same cannot be

said for the second percentage change. This example highlights the fact that the Euclidean distance measure is an unsuitable metric for the compositional space.

For this, Aitchison recommended that compositional data is analysed in such a way that scale invariance is preserved, such that any inference should not be dependant on the scale at use. Secondly, subcompositional coherence is too to be maintained - inference upon a component subset should not depend upon any data outside of that subset. Finally, order of the components should not influence the analysis. These three requisites, as defined above by the Principles of Compositional Analysis, can be summarised as the Aitchison Geometry.

The first formal attempt by Aitchison to make the data more symmetric, the ratios can be logarithmically transformed. For parts x_1 and x_2, the additive log-ratio $\text{alr}_{1,2} = \log(x_1/x_2)$ that generalises to

Definition 2.8. $alr_{i,D} : \log\left(\frac{x_i}{x_D}\right),\ i = 1, ..., D-1.$

The alr transformation relieves the unit-sum constraint and imposes normality following the multivariate Gaussian distribution, as argued by the author (Aitchison, 1986). After carrying out usual regression modelling, it is possible to back-transform the estimated coefficients, however the regression coefficients would have limits in their intepretability. Therein lies also the choice of the denominator in the log-ratio. In the above definition it was taken as the last component x_D, but may not necessarily be so. In fact, the denominator in the alr transformation can often be determined by the experimental set-up or decided according to another criterion, such as a reference variable.

A way to bypass the decision on the reference component can be explored through centred log-ratio transformations, where the geometric mean of the parts is employed instead:

Definition 2.9. $clr_{i,D} = \log\left(\frac{x_i}{(\prod_i x_i)^{1/D}}\right),\ i = 1, ..., D.$

In this way, the log-transformed parts are centred with respect to their mean

across the parts. While the transformed parts themselves are not independent, any subset of a clr set is linearly independent.

The log-transformation is known to convert ratio-scale data into interval-scale data. The ratios are linearized and multiplicative differences become additive on an interval scale. Most statistical methods utilise data on the interval scale, exemplified through the use of means and variances. Hence, the log-transform is the key to convert ratios into the appropriate additive scale for statistical computations and to symmetrise their distribution, as well as reduce the effect of outliers.

2.6 Box-Cox type transformation of compositions

The Box-Cox transformation is an example of a power transform family of transformations suitable for removing compositional constraints and projecting the original compositional data to the Euclidean space.

A general Box-Cox transformation with respect to a power parameter λ for a positive vector $\boldsymbol{y} = (y_1, \ldots, y_n), y_i > 0$ is given by Box and Cox (1964).

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda(\prod_{k=1}^{n} y_i)^{\frac{\lambda-1}{n}}} & \text{if } \lambda \neq 0, \\ (\prod_{k=1}^{n} y_i)\ln(y_i) & \text{if } \lambda = 0. \end{cases} \tag{2.1}$$

The parameter λ is estimated using maximum likelihood techniques so that the vector of transformed variables is approximately multivariate Gaussian. Then standard statistical modelling can be carried out on the set of transformed values. It is trivial to note that the logarithmic transformations defined previously are a special case of the Box-Cox transformations, in which case a small value for λ would be selected with the ratios of compositional parts. As well as the log-ratio approach, the Box-Cox transformation relies on the geometric mean value as a measure of the central tendency of compositional data.

An alternative has been proposed by Sharp (2006) in cases where it is desirable that the measure of central tendency lies directly inside the compositional data set. Sharp highlights that the geometric mean is an unsuitable measure when the compositions lie in a straight line. Instead, it is suggested that a multidimensional extension to the median value is used as suggested by Small (1997) as it is more consistent and intuitive with the compositional representation of variables on a ternary diagram. This is deemed important for applications such as principal component analysis on a compositional data set, and also to be used accordingly with transformations.

The above Box-Cox transformation has a shortfall of being applicable in the instances where $y_i > 0$. In compositional data analysis this meets the issue of essential zeros, where a component of $\boldsymbol{x}$ is exactly zero due to the absence of an observation in a category. For illustration, in the eye colour example from earlier, this could be the absence of people with eye colour other than Brown or Blue in country c_{19} where $\boldsymbol{c} = (c_1, ..., c_N)$ is the set of all countries taking part in the experiment. The strict-positivity constraint was first overcome by Aitchison and augmented by the developments of Tsagris et al. (2011).

Originally, Aitchison (1986) defined the power transformation for compositional vector $\boldsymbol{x} = (x_1, \ldots, x_D)$ and $\alpha \in \mathbb{R}$:

$$\boldsymbol{S} = \left(\frac{x_1^{\alpha}}{\sum_{i=1}^{D} x_i^{\alpha}}, ..., \frac{x_D^{\alpha}}{\sum_{i=1}^{D} x_i^{\alpha}} \right)^T. \tag{2.2}$$

Then the α-power transformation is defined by (Tsagris et al., 2011):

$$\boldsymbol{T} = \frac{1}{\alpha} \mathbb{H}(J\boldsymbol{S} - 1). \tag{2.3}$$

with $\mathbb{H}$ being the Helmert sub-matrix (Lancaster, 1965) and J is a vector of ones (length $|J|$).

The α-transformation is of further interest in modelling compositional data. It can be used to form an alternative method of regression analysis for compositional responses, especially useful when there exist essential zeros in the data set, so log-ratio based transformations are problematic. This is further explored and illustrated in Chapter 3 of this thesis.

2.7 Log-contrast transformation on the simplex

Another transformation of original compositional parts $\boldsymbol{x} = (x_1, \ldots, x_D)$ proposed by Aitchison (1986) is the log-contrast, which, as a linear combination, can be mapped to the Euclidean space.

Definition 2.10. *(Log-contrast) A log-contrast of a composition* $\boldsymbol{x} = (x_1, \ldots, x_D)$ *is a function* $f(\boldsymbol{x}) = \sum_{i=1}^{D} \alpha_i ln(x_i)$*, where* $\sum_{i=1}^{D} \alpha_i = 0; i = 1, ..., D$.

A particularly attractive use for log-contrasts is found in chemistry (Grunsky et al., 2008), for instance, in mass-preserving reactions $\sum_{k=1}^{D} \alpha_k = 0$ holds and α values are known constants. When a chemical reaction reaches equilibrium state, the log-contrast of a set of compositional parts and the α constants find suitable interpretation (Grunsky et al., 2008). Further, log-contrasts can be readily used in linear regression analysis with compositional responses and explanatory variables (Aitchison, 1986).

2.8 Amalgamation in CoDA

It is often observed that compositional data can fall into natural groupings, whether based on physical properties of the data or considerations in the statistical analysis carried out, such as heirarchical clustering. For example, in Chapter 3, we explore tree types found on a study site in Northern Yorkshire. Individual tree types fall into classes and families, and a natural question arises about any benefits or shortfalls of modelling these families of trees, instead of the

individual tree types.

In other instances, where compositional parts are numerous or contain many essential zeros, it may be advisable to consider some amalgamation (summing) of parts for computational ease or to bypass modelling of zeros, especially if the question of inference on individual parts is not acute. Amalgamation can also be an effective technique to achieve dimensionality reduction in a compositional data set. Returning back to the eye colour example at the beginning of this chapter, the class Other is indeed an amalgamation of eye colours such as green, hazel, grey, red, amber, variants of heterchromia or 'no response'. Aggregating these categories into a separate class allowed us to represent the categories Brown, Blue and Other on a ternary plot.

A transformation that exploits this grouping nature is known as the isometric log-ratio transformation (ilr). The definition of the ilr is based on two subsets (subcompositions) D_1 and D_2 of the compositional set (Egozcue et al., 2003). D_1 and D_2 are non-overlapping, such that $D_1 \bigcap D_2 = \emptyset$.

Definition 2.11. $ilr_{D_1,D_2} = \sqrt{\frac{D_1 D_2}{D_1+D_2}} \log\left(\frac{(\prod_{i\in D_1} x_i)^{1/D_1}}{(\prod_{i\in D_2} x_i)^{1/D_2}}\right)$, *where* D_1 *and* D_2 *are the first and second non-overlapping subsets of compositional parts respectively.*

An alternative specification for groupings of parts could be simple amalgamation of ratios, as defined previously. These offer ease of interpretability compared to the isometric log-ratio transform, provided adequate labelling. Comparison of different ilr values is also problematic, as it depends on relative values of the parts in the geometric means that form the transformation, and very differing original compositions can give the same ilr values.

Thus, while interesting from a theoretical standpoint, practical use of isometric log-ratio transformations is limited due to difficulties in their interpretability.

One other important feature is that amalgamation of components is not a linear operation in the simplex space (Aitchison, 1986) and the result of this is Simpson's

paradox (Good and Mittal, 1987), which was earlier recognised by Pearson (1897) and Yule (1903).

Outside of the application to compositional data sets, Simpson's paradox deals with contingency tables of success and failure rates in a population where an experiment took place to determine the aforementioned rates. If the population is divided into distinct classes and a contingency table is created for each separate class, Simpson's paradox can be seen when the performance (success/failure rates) of non-overlapping classes contradict the original contingency table for the whole population. Egozcue and Pawlowsky-Glahn (2008) discuss this phenomenon in the compositional setting of amalgamations, as a natural extension of the contingency table set-up. The authors make an assessment of existing methods to analyse proportions in which Simpson's paradox may occur, and address the question of finding a representative measure of each sub-composition such that the rates (success or failure, for example) can be compared across sub-compositions. Egozcue and Pawlowsky-Glahn (2008) find that an alternative measure to analyse is the geometric mean, rather than the non-linear amalgamation operation.

In contrast to earlier shortfalls of amalgamation of compositions, Greenacre (2018) argues for the use of transformations to consider amalgamated compositional data, deeming it necessary in fields such as geochemistry. He states that the acceptability of the use of a transformation technique should be guided by its benefits to a research question, and not its mathematical properties, such as non-linearity of amalgamations in the simplex, as long as the basic principles of scale invariance and subcompositional coherence are satisfied. On a wider perspective, such bespoke approaches to modelling may be justified on a case-by-case basis, but their transferability to other situations is naturally questionable.

2.9 Other transformations

One other transformation to consider for the purposes of this thesis is the square-root transformation. The square root transforms the simplex space onto the hypersphere, which is a contrast to earlier transformations onto the familiar Euclidean space. Even though this would again deviate from the normality assumption in classical multivariate modelling, the resulting set of variables can be modelled using the Kent distribution and the von Mises distribution (Scealy and Welsh, 2011; Stephens, 1982).

In the usual setting $\boldsymbol{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D$ subject to the constraints $x_i \geq 0$ and $\sum_1^D x_i = 1$, then $\sqrt{\boldsymbol{x}} = (\sqrt{x_1}, \ldots, \sqrt{x_D})$ lie on a hypersphere of dimension $D-1$: $S^{D-1} = \{\boldsymbol{y} \in \mathbb{R}^D : ||\boldsymbol{y}|| = 1\}$.

$\sqrt{\boldsymbol{x}}$ lies on the positive orthant of S^{D-1} and when $\sqrt{\boldsymbol{x}}$ is not on the boundary of the orthant it can be modelled using the Kent distribution. It is due to the unit-sum constraint of compositional data that this transformation and further modelling is possible, since $\sqrt{\boldsymbol{x}}^T\sqrt{\boldsymbol{x}} = 1$ as for directional data vectors that lie on the hypersphere. This transformation similarly preserves essential zeros, and the work of Scealy and Welsh (2011) use the transformation to further investigate regression with a compositional response variable with the help of the Kent distribution. The authors found that this approach to regression is difficult to implement, and that the transformed components exhibit large variance and lie close to the boundary of the orthant. The lines on the simplex (example Figure 2.1) are no longer equidistant on the hypersphere and points are stretched further apart at the boundary. Combined with a large variance of transformed components this can result in a phenomenon called folding. This may result in difficulties of estimation of parameters of the Kent distribution and stability of optimisation (Scealy and Welsh, 2014).

A final consideration of transformations of the simplex space is through a truncation. Butler and Glasbey (2008); Dobigeon and Tourneret (2007) used a trun-

cated Gaussian distribution to model the simplex space. In the former case Butler and Glasbey (2008) employed a latent Gaussian model with the primary aim of handling essential zeros - when a compositional component is exactly zero due to the absence of an observation in a category. These latent variables are assumed to follow the multivariate Gaussian distribution and lie on the unit hyperplane. The transformation is a function that minimises the squared Euclidean distance between the original composition parts and the latent variables, and from this unknown parameters of the multivariate Gaussian distribution are estimated. This approach faces shortfalls, namely the immediate violation of subcompositional coherence and scale invariance, which are two of the main principles of CoDA. Furthermore, estimation of parameters for $D > 3$ is deemed problematic. Butler and Glasbey (2008) thus express that this transformation should be applied for exploratory analysis or diagnostic purposes, or when other transformations are even less appropriate. More recently, other transformations have been explored through the works of Leininger et al. (2013); Tsagris and Stewart (2020) and Scealy and Wood (2020) that use sophisticated folding and scale matching techniques to overcome the modelling difficulties introduced by essential zeros.

2.10 Modelling of compositional data

In this chapter so far, we have addressed constraints on compositional vectors, the effect these constraints may have on the variance structure and possible solutions in the form of transformations. In statistical modelling, compositional structure can be found when data are in the role of predictor variable set, the dependent response set or both. Log-ratio transformed compositional parts are similarly employed in the role of predictors and responses, as detailed in Aitchison's work (Aitchison, 1986). Compared to linear regression modelling, two key assumptions are violated in the compositional framework - the Normality assumption and the independence assumption. This is especially prevalent when compositional parts

are considered as the response variable. However, compositional predictors also would exhibit high multicollinearity, even when the data are transformed. For example, the clr-transformed parts are still not independent. Similarly, if the purpose of the analysis is prediction, problems may arise with predicted values outside of the range $[0, 1]$ or that they do not submit to the sum-one constraint.

Many approaches have been suggested to overcome these difficulties, and with increasing computing power reliable inference has been possible for high-dimensional compositional data as well, where the number of predictors exceeds the number of data samples available.

The approaches to modelling compositional data can be split roughly into two. The first relies on transformation of the original composition, as given in earlier parts of the chapter. Then, any problems that arise therein, for example, high multicollinearity of log-ratio transformed parts, are addressed with existing statistical methods. See Wang et al. (2010) for partial least squares approach, for instance. The second view on compositional regression is to assume that the true and underlying distribution for the responses is the Dirichlet, and similar generalised linear models are built through a logit link function for the mean (Maier, 2014).

Application of Bayesian techniques in the compositional framework can be explored through the work of Iyengar and Dey (1998), who review methods such as Box-Cox transformations for both parametric and non-parametric regression models. Bayesian perspectives can allow to express the regression coefficients in terms of probability distributions. The modeller can include any information already known about the parameters through the prior distribution. The Bayesian model can be applied directly and exactly to fewer samples also, as it does not rely on asymptotic results for validity. Moreover, the issue of overfitting, where the selected model can fit the existing data extremely precisely but may be completely unsuitable for a new data set at hand, can be dealt with

through the use of a prior distribution expressing a penalty term reflecting model complexity. Bayesian regression of compositional data can also accommodate high-dimensional models, where the number of predictors exceeds the number of observed responses. A common choice for prior distributions in a Bayesian regression setting are spike-and-slab priors (Andersen et al., 2014) and horseshoe priors (Carvalho et al., 2012). For the former, regularisation approaches have also been incorporated, as in the work Ročková and George (2018), who describe the spike-and-slab approach combined with the penalisation LASSO.

Bayesian approaches to compositional regression modelling arguably have easier interpretation, than those based on transformations due to the need to back-transform - the meaning of the regression coefficients may not be intuitive in the physical application, and may be difficult to communicate to scientists with limited statistical background. For example, the additive log-ratio transformations is not symmetric with regards to which part is taken as the reference part in the denominator, and interpretation of the estimated regression coefficients should change if a different reference part is taken. Similarly, transformation-dependent modelling relies heavily on the absence (or adequate accommodation) of essential zeros. Careful consideration must be given to modelling compositions that contain these zero values. This area of CoDA has been widely studied (Aitchison and Kay, 2003; Stewart and Field, 2011), and is still of interest in modern statistics.

2.11 Other considerations - multimodality

The relative nature of parts of a composition can give rise to similar problems of bimodality or multimodality as count data. This can be due to mixture of some populations, or, as we will see in Chapter 4, high variances in the univariate Beta distribution, for example, will drive its probability density function to have peaks at the end-points of the $[0, 1]$ interval. Similarly for data sets with high proportion of essential zeros, we may observe modes in the distribution of data, and may look

to an amalgamation to constrain this phenomenon. A log-ratio transformation is not feasible for a zero-part in a composition, however, we may look towards the aforementioned isometric log-ratio transformation to work with amalgamations, or to change the scale of the data set and allow for later modelling with the multivariate Gaussian family. In the Bayesian framework for a likelihood-driven analysis any non-informative prior on the simplex (for example, the Jeffreys prior distribution) is hoped to capture multimodality in a sufficient manner. Alternatively, if a-priori a multimodal structure is strongly believed to exist, we may wish to construct an accommodating prior distribution. As already mentioned, a candidate for such could be a high-variance Beta distribution, which generalises to the Dirichlet distribution with high concentration parameter values. Other distributions that can accommodate instances of multimodality are described in Chapter 4. Similarly, we consider these in light of a prior elicitation exercise in Chapter 7. Firstly, identification for the need of a multimodal distribution must be established through discussion with the experts, and then this decision leads to the selection of an appropriate prior distribution with parameters that accommodate multimodality. Alternatively, should the experts give judgements that imply existence of more than one mode in the consensus prior distribution, this needs to be highlighted in the discussion that follows. All these nuances are described in Chapters 6 and 7.

Chapter 3

Modelling approaches for identification of tree species

In this chapter, we illustrate the nuances of compositional data analysis in a regression setting. The task presents itself as a set of continuous predictor variables and the response is a 10-part composition. We compare approaches after conducting transformations in the log-ratio family with a regression approach that assumes the underlying distribution is the Dirichlet. We also explore non-parametric approaches to modelling the problem using random forest regression and similarly seek to reduce computational time by extracting regressors that explain most of the variance within the set of regressors using principle variable method of Cumming and Wooff (2007). The practical motivation for this approach is to enhance an existing sub-pixel classification method that is used to identify tree types in a woodland area in North Yorkshire, UK. This chapter stems from collaboration with Fera Science, UK, and a part of this work has formed a paper due for submission to the journal Remote Sensing of Environment. Any work in this chapter not carried out by A. Frantsuzova is clearly indicated.

3.1 Introduction

This chapter describes a joint project with Fera Science, UK, with the aim to augment Fera's existing research started in the "Pre-clasp pilot study – Earth observation for the identification of tree species distributions using sub-pixel classification methods" from 2017. Fera Science is a UK-based research organisation that focuses on plant sciences, environmental and agricultural conservation, and food safety. The pilot study is part of a larger project to investigate dieback of ash trees in the UK, interest also lies in investigating methods of detecting tree types where only lower-quality satellite imagery is available. Particular focus is with monitoring amenity woodlands, which comprise of private and agricultural land, smaller publicly-owned blocks, as well as individual trees in peri-urban areas (DEFRA, 2018*a*). This is also of wider importance to the UK Tree Health Resilience Strategy (DEFRA, 2018*b*) which highlights the importance of, and key steps for, managing a healthy and resilient treescape. Therefore, there is a need to enhance existing methodology to identify and monitor tree species in urban and peri-urban areas.

In some instances, only lower-resolution land cover data from a satellite, such as ESA Copernicus programme Sentinel 2 satellites, may be available for statistical modelling. Another source of data are higher resolution Unmanned Aircraft Systems (UAS) or 'drones'. Drones provide a cost-effective and efficient way of gaining high resolution imagery over a small area. However, the use of drones in peri-urban and urban areas is subject to aviation regulations and their use in highly populated areas may require specific permission. A potential solution to enhance information that can be gained about tree species distribution from lower-quality satellite data without resorting to the use of drones is to apply sub-pixel classification algorithms trained on high-resolution drone imagery in rural locations to lower resolution Sentinel 2 satellite imagery.

There exist several variants of sub-pixel classification algorithms, for instance,

Sood and Gupta (2018) review the linear spectral mixture model, support vector machine models, methods based on maximum likelihood classification, as well as others. We concentrate on estimating the proportional composition of pixels that have mixed spectral characteristics. Details of data collection and spectral characteristics are described in the next section.

Fera's pilot study used the method of random forest regression to classify tree types in each mixed pixel. This was motivated by random forest's non-parametric property, allowing us to bypass the assumption of a particular underlying probability distribution. The random forest algorithm is also robust to any outliers in the data, and has proven to be a well-established method in remote sensing and land map modelling (Gislason et al., 2006; Rodriguez-Galiano et al., 2012; Pelletier et al., 2016). In this collaboration, the role of A. Frantsuzova and J. P. Gosling was to augment the modelling approach to increase prediction accuracy, measured using root mean squared error (RMSE). Similarly, the link with compositional data analysis is evident, as the response variable is expressed as a 10-part composition. It is interesting to also compare the random forest algorithm with techniques arising from developments in CoDA.

3.2 The data set and exploratory analysis

The data set was collected and compiled by Fera Science, UK: Paul Brown carried out drone imagery capture, image processing and high-resolution classification; Lee Butler carried out drone imagery capture and image processing; and Simon Conyers (University of Newcastle) was responsible for the species identification for classification training of the drone imagery. Data collection was carried out primarily for the pilot study (2017), and no further additions to the data set by the above persons were made for the purposes of collaboration with A. Frantsuzova and J. P. Gosling.

Data was collected over an area of interest (AOI) presented in Figure 3.1 below

located in North Yorkshire, UK. The area is classified as a small mixed woodland sized 34.94 ha and contains natural and plantation woodland. A drone (UAS) was used to collect imagery over the AOI for leaf-off (February 2016) and leaf-on (August 2016) conditions. Details of the UAS specifics are omitted here, but one camera captured near infra-red imagery also. Another data set was compiled by Fera Science, UK and consisted of a ground survey of the AOI. A random forest classifier was used on high-resolution drone data compared against data from the ground survey. Overall classification accuracy of the UAS imagery was 68%, and accuracy varied across species.

Lower resolution satellite imagery was then taken by Fera Science UK from the Sentinel-2 satellite downloaded from the European Space Agency (ESA) Copernicus Open Access Hub (Scihub 2017). The dates for data collection were 16th April 2016, 19th July 2016 and 26th December 2016. This data has four 10 metre spatial resolution bands for visible-NIR (VNIR; blue (490nm), green (560nm), red (665nm), NIR (842nm)) and 20 metre spatial resolution for four bands dedicated to the red edge (705nm, 740nm, 783nm, 865nm) and for two bands in the short-wave infra-red (SWIR; 1610nm, 2190nm). 20m spatial resolution bands were resampled to 10m in order to match the spatial support of the VNIR bands. The data was further corrected and geo-referenced for alignment with UAS hard classification.

The AOI tree classification was converted to a set of area fraction images (AFIs) to be used to train the classification algorithm for the lower-quality Sentinel-2 data (Verbeiren et al., 2008; Heremans et al., 2011). The AFIs were created using a 10m spatial resolution frame derived from the Sentinel-2 imagery for individual grid cells in the AFI to match to a single pixel in the Sentinel-2 imagery. An AFI was created for each tree species and for the area of shadow before being combined into a single GIS layer describing the species composition for each pixel. Table 3.2 shows the mean proportion of each tree class for the entire data set. Grid cells that contained non-classified area (i.e. areas which are not tree or shadow)

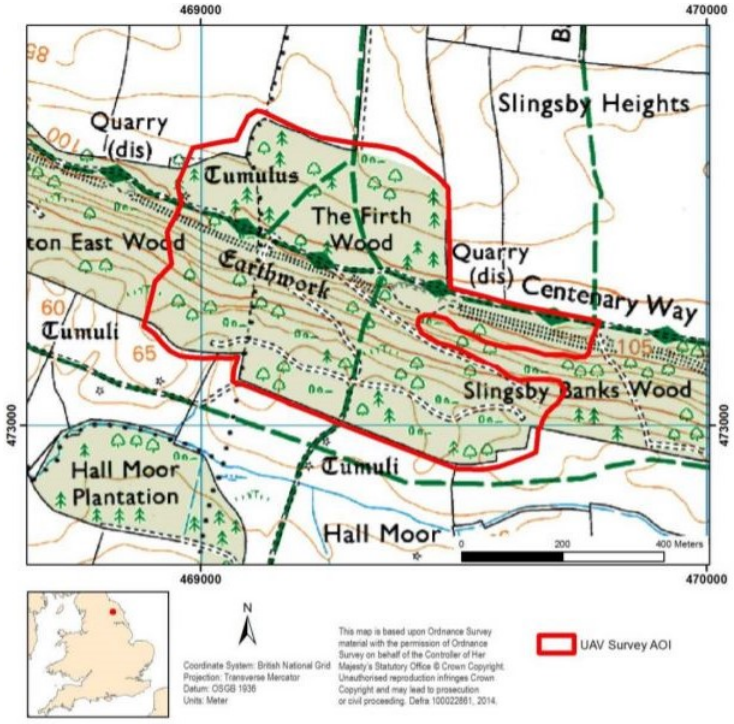


Figure 3.1: Slingsby Bank AOI. Red outline shows the area of UAS imagery classification. Source: "Pre-clasp pilot study – Earth observation for the identification of tree species distributions using sub-pixel classification methods", Fera Science, 2017.

were removed from the AFI data set, so that it only contains pure woodland cells, where the AFI proportions in each cell sum to one. The 10m resolution data set of 30 spectral bands was spatially joined to the AFI image and the combined attribute data for each grid cell was extracted to produce the data set on which the sub-pixel classification is performed.

Band	Wavelength (nm)	Band type	Spatial resolution (m)
B2	490	Blue	10
B3	560	Green	10
B4	665	Red	10
B8	842	NIR	10
B5	705	NIR	20
B6	740	NIR	20
B7	783	NIR	20
B8a	865	NIR	20
B11	1610	SWIR	20
B12	2190	SWIR	20

Table 3.1: Spectral bands of MSI sensor on-board Sentinel-2 satellite.

Tree Type	Mean proportion
Ash	0.021
Beech	0.049
Larch	0.186
Oak	0.189
Scots Pine	0.052
Shadow	0.145
Silver Birch	0.092
Sitka Spruce	0.068
Sweet Chestnut	0.071
Sycamore	0.126

Table 3.2: Mean proportion of tree class for entire dataset.

The AOI tree types are: ash, beech, larch, oak, scots pine, silver birch, sitka spruce, sweet chestnut, sycamore and a shadow class. Furthermore, the British National Grid easting and northing for each grid cell was added to the predictor data set, to allow for exploration of spatial patterns. The study site was dissected into 4 quadrants according to midpoint of the eastings and northings of the cells covering the woodland, as illustrated in Figure 3.2. The eastings for the cells range between 468,906m and 469,756m, with the mid-point being 469,331m. Similarly for northings, we have a range of 472,927m to 473,597m with the mid-point being 473,262m. Random forests were created using training data from three of the four quadrants and the accuracy of the regression calculated using the cells in the remaining quadrant as the testing data set. This approach was iterated until all the quadrants were used in the testing step.

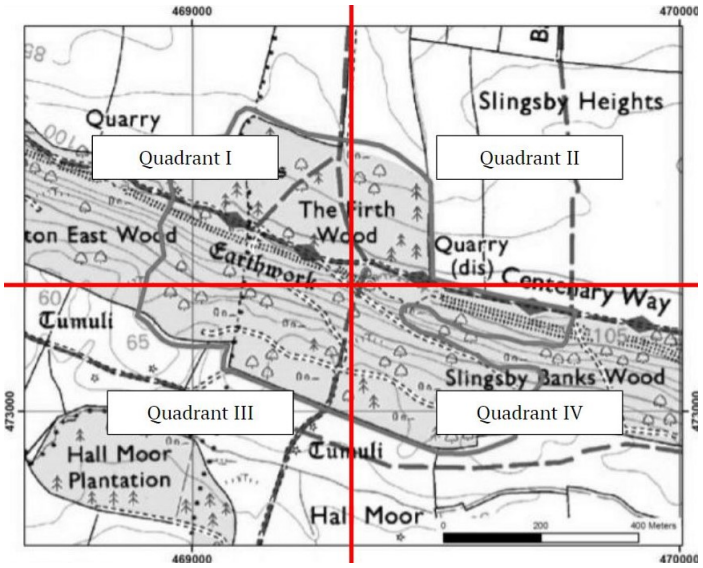


Figure 3.2: Study site quadrant dissection.

Tree Type	% essential zeros
Ash	87.7
Beech	81.1
Larch	51.5
Oak	45.2
Scots Pine	75.3
Shadow	6.32
Silver Birch	64.0
Sitka Spruce	78.3
Sweet Chestnut	70.9
Sycamore	54.3

Table 3.3: Percentage of essential zeros recorded by tree type.

The data set contains 2153 records (rows), 10 tree species (including shadow class) as the response variable, two easting and northing coordinates for each grid cell, and the set of predictor variables comprises of 30 spectral bands (ten bands B2 to B12 repeated for each season: winter, spring and summer) as defined in Table 3.1. No missing values are contained in the data set. The 10-part compositional response of tree types contains finite decimal places (to 2 d.p.) and a large proportion of essential zero values. Table 3.3 summarises zero contributions to individual tree types, and the overall proportion of essential zeros for the tree types is 61.5% to three significant figures.

Spatial dependency was explored through pairwise correlation plots for pixel val-

ues for individual seasons, found in Figures 3.3 to 3.8. Similarly, pairwise scatter plots between the spectral bands for each season show a clear linear relationship of varying strength. As an example, for the season Spring spectral bands $B_{6,spr}$ and $B_{7,spr}$ exhibit an overall stronger linear relationship with the other bands for this season than the band $B_{2,spr}$. Furthermore, $B_{6,spr}$ and $B_{7,spr}$ are both strongly linearly associated with $B_{8A,spr}$ which is a NIR band of a similar wavelength. We can see that there is much more evident linear correlation between individual spectral bands for the seasons of Spring and Summer, and the weakest associations can be seen in the Winter-Summer plot. This hints at the temporal component in the data set, with changing leaf colour by season. We can also detect highly correlated spectral bands and account for this relationship in order to help us reduce the number of dimensions in our set of explanatory variables. On the other hand, this clearly suggests the existence of multicollinearity, which occurs when one explanatory variable in a multiple regression model can be linearly predicted using the other variables, which can lead to sensitivity of coefficients in the model and, hence, predictions. This may also play a minor role when conducting the random forest classification due to the number of correlated variables present. We can see that there is consistent positive correlation between Bands 11 and 12, and where there was negative correlations for the Winter season in Bands 6, 7 and 8, this correlation becomes positive for the seasons Spring and Summer.

Figure 3.3: Correlation plot for pixel values in seasons Spring and Summer.

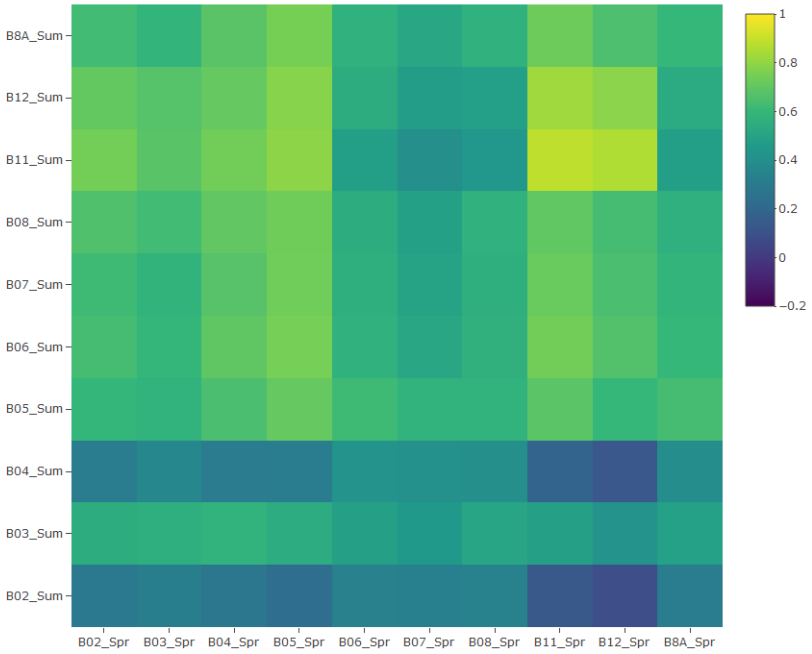


Figure 3.4: Correlation plot for pixel values in seasons Spring and Winter.

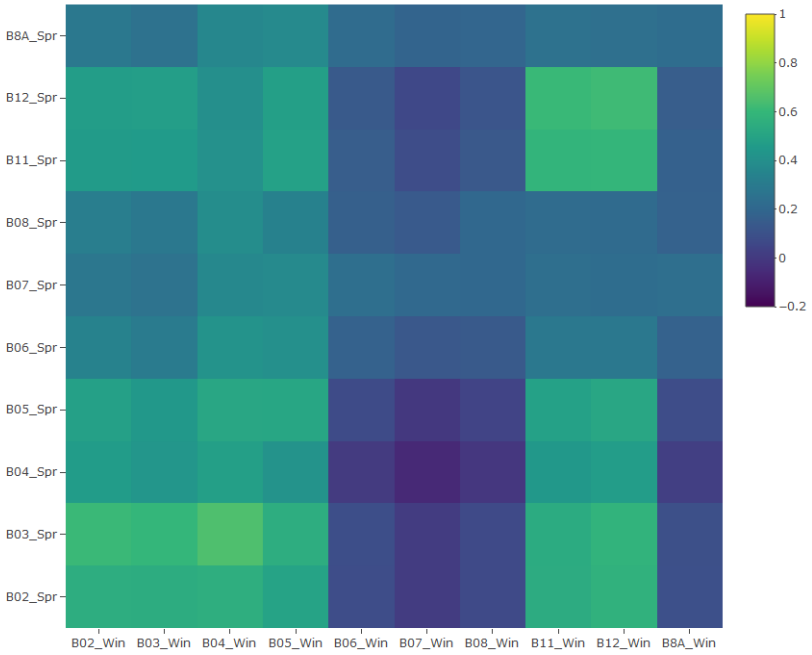
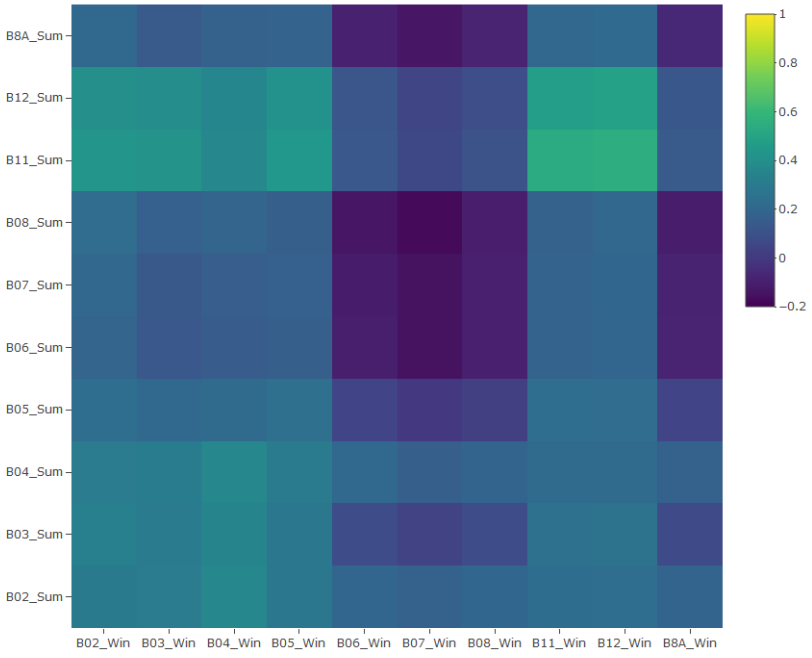


Figure 3.5: Correlation plot for pixel values in seasons Winter and Summer.



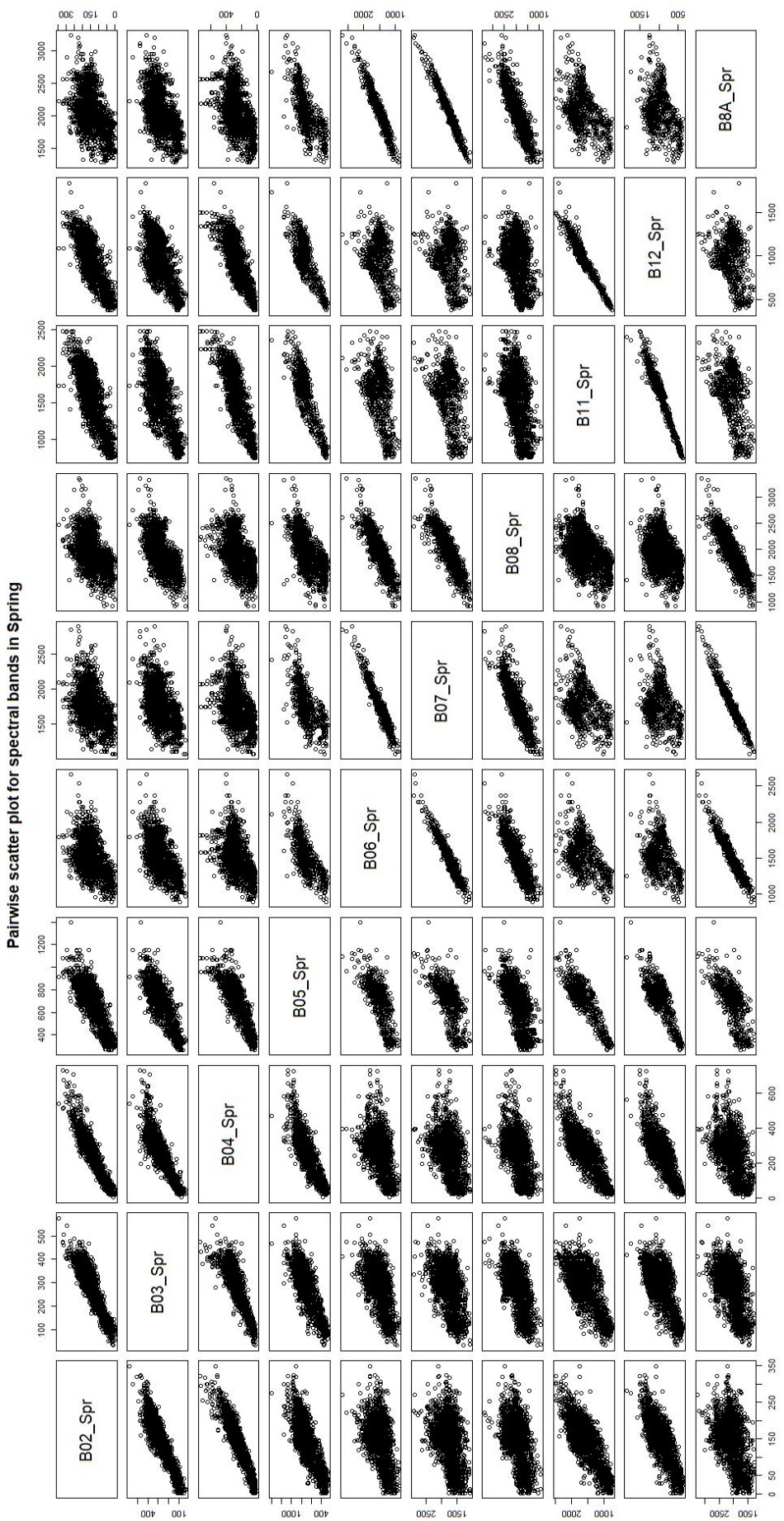


Figure 3.6: Pairwise plot for spectral bands in Spring.

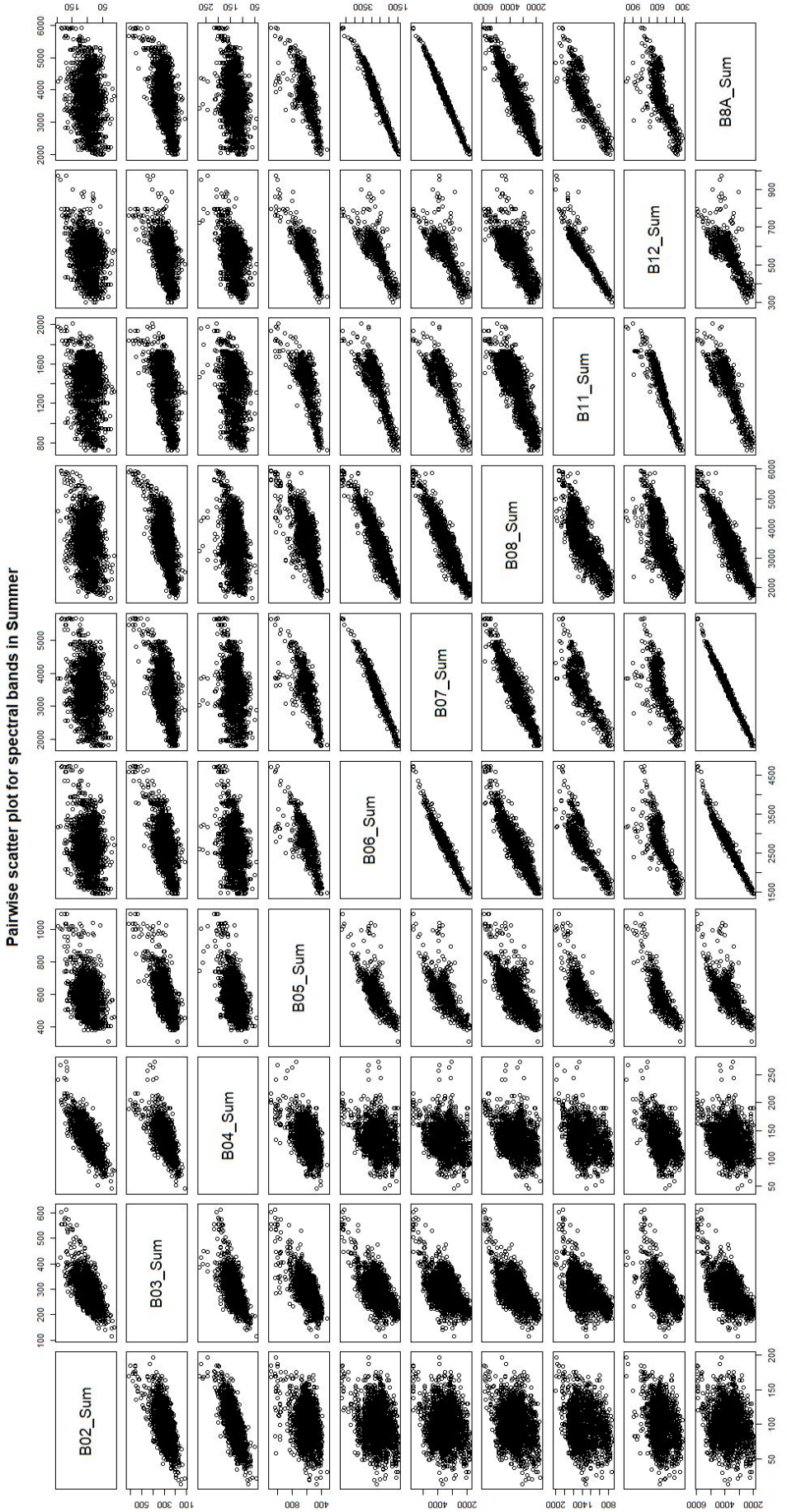


Figure 3.7: Pairwise plot for spectral bands in Summer.

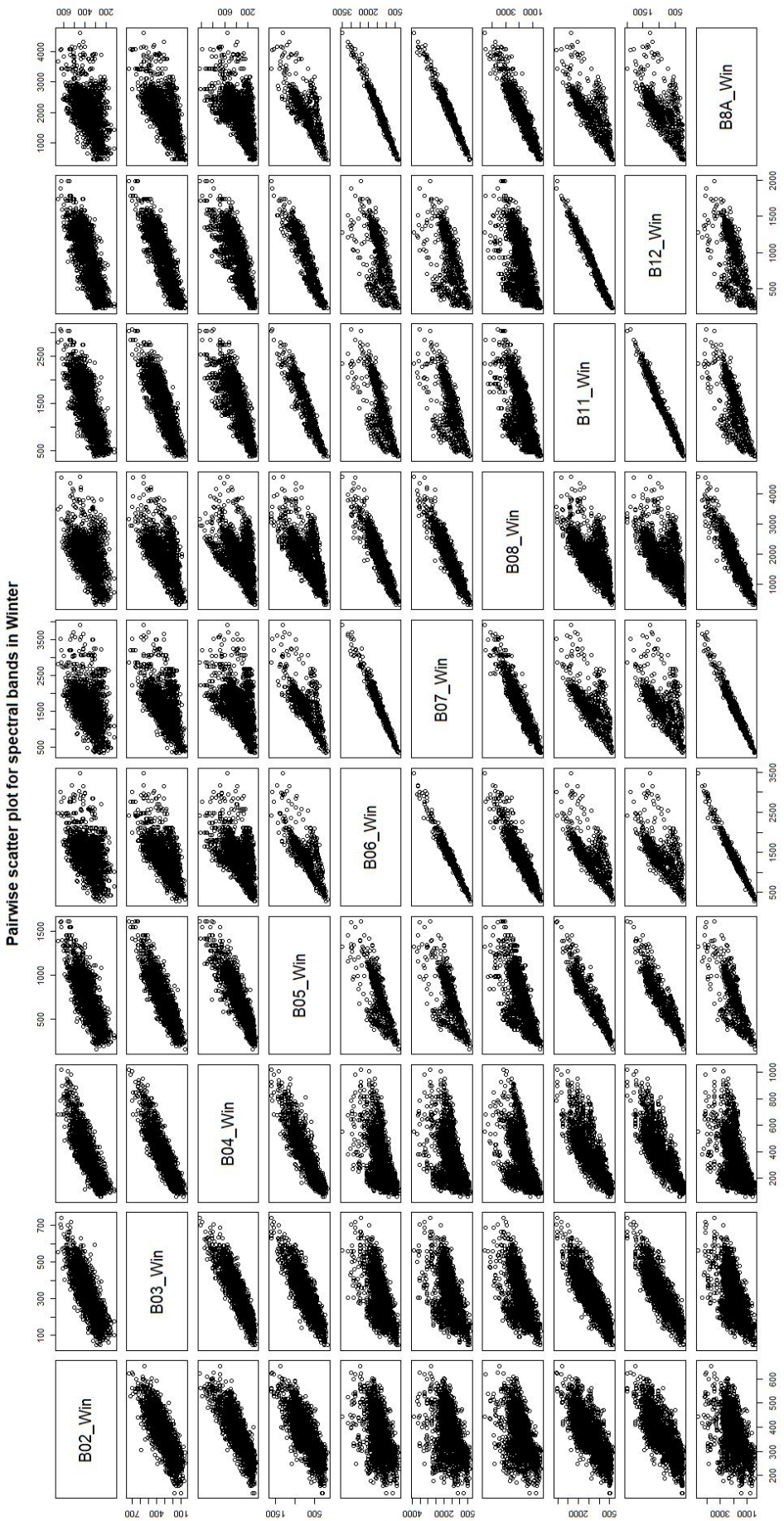


Figure 3.8: Pairwise plot for spectral bands in Winter.

Results from the hard UAS classification can be plotted to give an indication of the spatial spread of each tree type over the area of interest, presented in Figure 3.9 and Figure 3.10. In the later plots, for example, we can see clear regions (south Y coordinate) in both instances where clusters of Larch and Sitka Spruce dominate the pixels. Similar regions are highlighted in the adjacent plots.

Figure 3.9: Scatter plot of proportion of tree type per pixel with coordinate location. Clockwise: Ash, Beech, Shadow class and Oak.

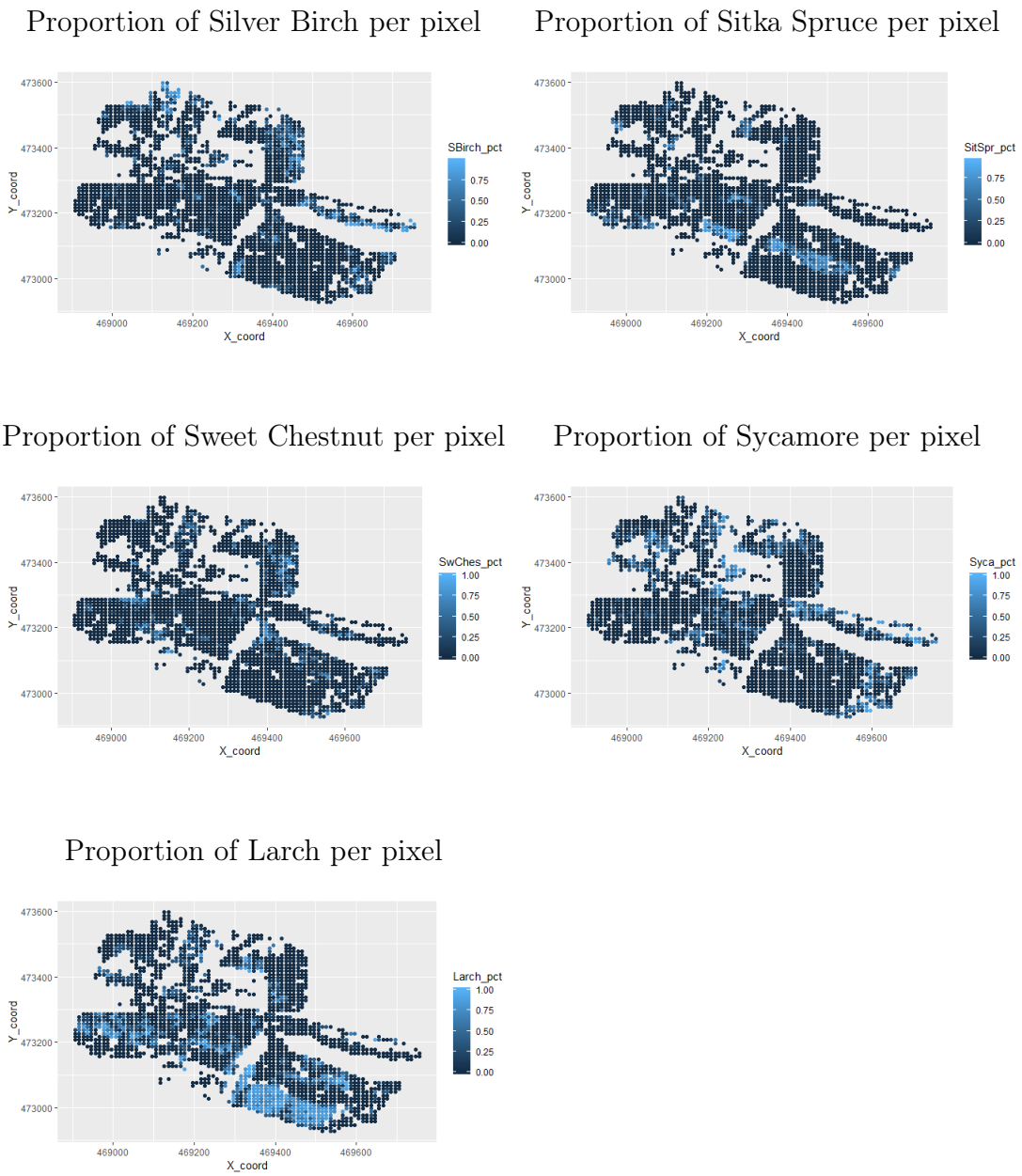


Figure 3.10: Scatter plot of tree type with coordinate location. Clockwise: Silver Birch, Sitka Spruce, Sycamore, Larch and Sweet Chestnut.

3.3 Methodology and results

This section presents the methodology employed for sub-pixel classification modelling of tree types in a mixed pixel. Each subsection concerns a separate method, starting with random forest (RF) regression, which was preferred by our collaborators at Fera. Then follow other approaches in compositional data analysis, which were considered by A. Frantsuzova as being relevant to this problem. For the purposes of coherence each subsection describes each method and presents results driven by the method, and then an overall comparison is conducted in the Discussion section of this chapter.

3.3.1 Regression tree and random forest regression for sub-pixel classification

To perform the sub-pixel classification, a regression tree or random forest regression for each of the nine tree species and the shadow class was constructed (Huguenin et al., 1997). The intensity values of pixels in the 30 bands are treated as predictors and the species composition in a pixel from the UAS classification became the response variable of each regression tree or random forest regression. Recall that each tree species (and the area of shadow) is constructed from an area fraction image before being combined into a single GIS layer describing the species composition for each pixel. The algorithm for building the regression trees is shown in Figure 3.11.

Random forest (RF) due to Breiman (2001) is an process that augments multiple decision trees to form a larger structure, and combines the bagging method (Breiman, 1996) with random variable selection. For each decision tree of the RF bagging sampling is used to binary split the data using different rules, depending whether the problem is regression or classification. For our purposes to build a non-parametric regression model using RF, the splitting criterion minimises the sum of squares of mean deviations used for training each tree model.

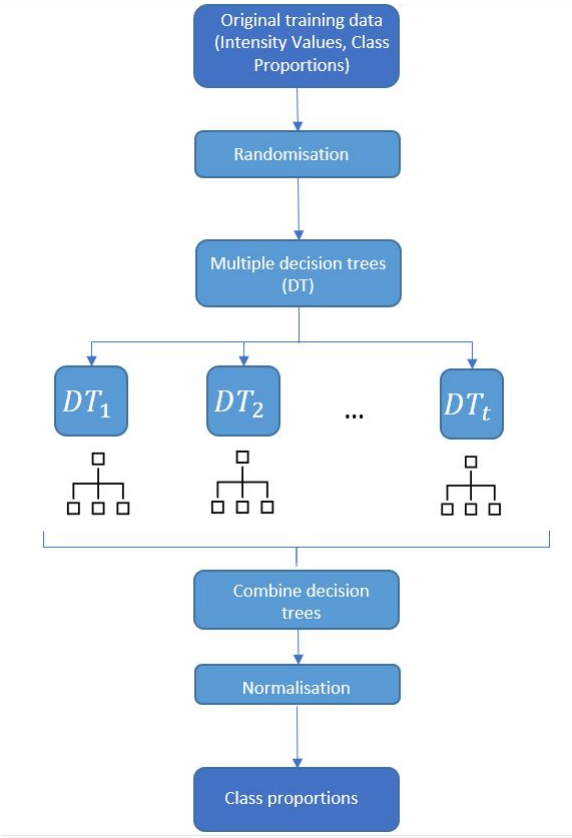


Figure 3.11: Decision tree regression approach for the sub-pixel classification of remote sensing data.

As the proportion for each species is estimated independently, the species proportions for each pixel were normalised according to Equation 3.1 so that they sum to 1:

$$P(i) = \frac{DT(i)}{\sum_i DT(i)}. \tag{3.1}$$

where $DT(i)$ presents decision tree i, $i = 1, ..., t$, as in Figure 3.11.

The pixel data was split into training (80%) and testing (20%) data sets. Regression trees were created using the `rpart` package in `R` and random forest regression was performed using the `randomForest` package in `R` with number of trees in the random forest set at 250, 500 and 1000, 1500. Performing the analysis with random forest size 500 gave no significant reduction in fit accuracy compared to using forests of larger size 1000 or 1500. For both regression trees and random forest regression the branches of the trees are split to maximise the between group sum of squares, with the accuracy of the model defined by the root mean square error as given in Equation 3.2.

$$RMSE = \sqrt{\frac{\Sigma_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}. \tag{3.2}$$

where $\hat{y}_t$ is the predicted value of variable y_t, and n is the sample size.

Additional testing was carried out to explore whether choosing the training data geographically, rather than by randomly assigning training and testing points in 20-80 proportion, affected the performance of the random forest regression. To achieve this, we dissected the study site into 4 quadrants according to midpoint of the eastings and northings of the cells covering the woodland, as demonstrated in Figure 3.2. Random forests were created using training data from three of the four quadrants and the accuracy of the regression calculated using the cells in the final quadrant as the testing data set.

Analogous to the random forest algorithm, classification and regression tree analysis (CART) is an algorithm based on building trees, however, only single trees

Tree Type	Random forest RMSE	Random forest RMSE with spatial (coordinate) explanatory variable	CART RMSE
Ash	0.080	0.084	0.096
Beech	0.120	0.121	0.168
Larch	0.196	0.186	0.312
Oak	0.216	0.208	0.295
Scots Pine	0.103	0.102	0.150
Shadow	0.101	0.110	0.176
Silver Birch	0.157	0.148	0.202
Sitka Spruce	0.115	0.115	0.200
Sweet Chestnut	0.138	0.135	0.169
Sycamore	0.176	0.170	0.253

Table 3.4: Random forest size 500 and 20 repeated simulations, with test data set for prediction purposes.

are considered in the exercise. `R`'s version of this algorithm is named Recursive Partitioning And Regression Trees (RPART) and has been implemented. It rests on recursive splitting of the data set until a stopping criterion is reached. With each movement, a split is performed based on the the explanatory variable which gives the greatest reduction of the dependent variable. We also consider inclusion of two further predictor variables, namely the X and Y coordinates of the pixel location, as it is likely that there exists some spatial dependency, as has been hinted by earlier exploratory plots.

Results from Table 3.4 indicate that the inclusion of a spatial (X-Y) variables did not improve the RMSE scores for all the tree types equally. Using two-sample $t-$tests with 50 repeated runs of the random forest models (mean RMSE estimates were identical to those given above, to 3 significant figures) yielded varied significance results for tree species using different random forest set-ups. Compared to the base scenario with seasonal predictors only, adding the two spatial variables showed a significant difference (at 5% level) in RMSE scores for tree species Beech, Larch, Oak, Shadow class, Birch and Sycamore. In comparison, CART yielded significantly higher RMSE scores across all tree types. Using variable importance plots from the random forest procedure can be found

in Appendix A (9.1). However, it is clear that RF strongly outperforms CART in terms of standard deviation of unexplained variation around the fitted model. It is worth noting the ranges of RMSE scores for the individual tree types. Larch and Sitka Spruce exhibit larger ranges of RMSE between the different approaches (random forest and CART) than the broad-leaf species. This larger range, especially for the quadrant analysis may further suggest an underlying spatial structure, which is highlighted by the block-planting approach used for species such as Larch and Sitka Spruce. Considering RMSE rank of the quadrants in each tree species in Table 3.5 yields us a further indicator that the species are distributed through the woodland - Quadrant IV has the highest mean rank of 2.88, followed by Quadrant III (mean rank:2.44) and Quadrants I and II (mean rank: 2.33). We can further relate RMSE scores with woodland composition. Table 3.2 provides the proportion of each tree type in the entire data set. For the purposes of training and testing the random forest models, a model for each tree type was simulated 20 times, and so the training and testing data sets were selected 20 times for each species. This resampling technique, akin to bootstapping, can provide us with mean RMSE scores representative of those for the whole data set. So a comparison of RMSE scores with the mean proportions in Table 3.4 are reasonable, instead of tracking the proportions of each tree type 20 times each model is trained. Thus, we can make relative comparison of RMSE scores across the tree species - for example, the RMSE for Ash is approximately four times the magnitude of the mean (0.02), whereas for Larch it is only just over one times the magnitude of the mean (0.18). As previously, the differences between RMSE scores between the quadrants for each tree type are significant at the 5% level for all tree types considered.

We consider a similar exercise to the above using multiple response data (tree type compositions), in contrast to performing random forest regression for each tree type in turn and performing normalisation. Due to potential physical dependencies between the tree types, we may gain further insight through simultaneous

Tree Type	Quadrant I	Quadrant II	Quadrant III	Quadrant IV
Ash	0.078	0.101	0.088	0.102
Beech	0.146	0.088	0.169	0.130
Larch	0.197	0.178	0.267	0.302
Oak	0.276	0.283	0.238	0.230
Scots Pine	0.131	0.110	0.156	0.099
Shadow	0.132	0.179	0.119	0.115
Silver Birch	0.192	0.207	0.145	0.215
Sitka Spruce	0.120	0.061	0.188	0.139
Sweet Chestnut	0.154	0.208	0.131	0.155
Sycamore	0.247	0.188	0.162	0.210

Table 3.5: Random forest RMSE with spatial (coordinate) explanatory variable for each respective quadrant used as the data set for prediction and RMSE calculation.

analysis of the responses. In practical terms, similar to the random forest, here we utilise bootstrapping and a set number of randomly selected explanatory variables at each split. Using the `R` package `MultivariateRandomForest` this is a simple exercise, though computationally requires approximately half an hour for one run. Results are presented in Table 3.6.

Another approach used for comparison is kriging, which originates from the field of geostatistics and can be regarded as a generalisation of a univariate (or multivariate) linear regression model as a method of interpolation. Kriging incorporates spatial correlation between the values taken as a sample, hence, no particular model of the spatial structure is presumed a-priori. Kriging is popular in fitting smooth functions to data that is spatially or temporally correlated (Matheron, 1963) For each value interpolated there corresponds also a level of uncertainty about it. For our purposes, the `R` package `DiceKriging` was used to perform a very similar analysis to the above techniques, with the inclusion of spatial X and Y coordinates and default nugget size. RMSE values between the predicted values and the test data set are given in Table 3.6. Compared to the baseline random forest RMSE scores in Table 3.4, the results given by the multivariate random forest and kriging approaches yield significantly higher RMSE scores at 5% level.

Tree Type	Multivariate random forest RMSE	Simple kriging RMSE
Ash	0.1053	0.0940
Beech	0.1503	0.1606
Larch	0.2465	0.3474
Oak	0.2619	0.3105
Scots Pine	0.1136	0.1573
Shadow	0.1356	0.1817
Silver Birch	0.2031	0.2228
Sitka Spruce	0.1375	0.2280
Sweet Chestnut	0.1471	0.1667
Sycamore	0.2121	0.2534

Table 3.6: Multivariate random forest and kriging RMSE scores, with test data set for prediction purposes.

Model diagnostics were considered for the methods under consideration. The random forest and kriging methods do not make distributional assumptions on the residuals, but it is still worth to explore residual homoscedasticity. Figures 9.1-9.6 in Appendix A depict (raw) residual diagnostic plots for the primary random forest approaches, with and without the inclusion of spatial variables. Very similar patterns were found for the CART, kriging and when considering individual quadrants for prediction purposes. Specifically for the residual-fitted plots, we can observe behaviour characteristic of potential model-misspecification for zero-inflated data. In our modelling of mixed pixels, this could prove an important feature (see Table 3.3) due to high proportion of essential zeros in the data set. Future work on this topic could consider the use of zero-inflated models for compositional data to address this issue, see for example (Salter-Townshend and Haslett, 2006).

3.3.2 Tree type amalgamation

Further thought was given to the classes of tree types presented in the data set. Considering that the spatial explanatory variables showed an improvement in RMSE scores, we decided to adopt an amalgamation to the ten response tree type categories into families of trees, in order to see whether there was any further improvement. In order to accomplish this, we consulted the lists of tree genera, freely available online (*List of Latin Botanical Tree Names, Genus and Species*, 2020). From this, two separate families were formed - the Beech Family consisting of Beech, Oak and Sweet Chestnut; and the Oak Family which includes Larch, Scots Pine and Sitka Spruce. The other four categories: Ash, Shadow, Silver Birch and Sycamore remained solitary. The aggregated categories were obtained by simply adding together the relative individual proportions. To follow, the standard random forest approach from earlier, with the inclusion of spatial co-ordinates, was performed. An issue was faced with retaining consistency with previous results due to the reduction of the total number of data set columns used. We thus adjusted the number of rows used also for this task, reducing the total data set by around 100 rows, then splitting into training and testing data sets in the ratio 80:20, as previously. The rows were selected using simple random sampling without replacement, for each iteration of the random forest algorithm. This procedure was repeated twenty times, thus yielding twenty size-adjusted data sets to be split into training-testing data and then the classification was performed. Results can be seen below in Table 3.7, which portray the arithmetic mean of the twenty runs of the algorithm.

Even though we do see an improvement in RMSE scores for the categories left solitary, limited meaningful comparison with previous results can be carried out for the two families of trees, unless we state that the tree family RMSE score spans across the individual tree types that make it up. Our reservations also lie with whether selecting other arbitrary partitions, for example, grouping tree types by alphabetical order, of the categories would yield a similar outcome.

Tree Type	Random Forest RMSE
Ash	0.07988
Birch Family (Beech, Oak, Sweet Chestnut)	0.2341
Pine Family (Larch, Scots Pine, Sitka Spruce)	0.2165
Shadow	0.1068
Silver Birch	0.1492
Sycamore	0.1719

Table 3.7: Random forest RMSE for aggregated categories, with inclusion of spatial coordinates.

It may also help us to investigate any clusters present in the compositional response tree types. For this we portray a cluster dendrogram of the variables, which portrays any hierarchical relationships in the data. Since the response tree types are compositional and contain a significant proportion of essential zero values, the original data was transformed using the centred log-ratio transformation as defined in Chapter 2, for comparative ease of interpretation with the additive log-ratio transformation. Using the clr we have sustained all ten classes of the composition, and this similarly eases interpretation of the hierarchical structure. If we use an alr transformation with the Shadow class as the reference variable, a similar dendrogram structure is produced and can be seen in Appendix A. In this case the ilr transformation was not considered, since it rests on the choice of subsets of the compositional vector, which in itself is determined by a clustering procedure.

The transformation was also used to facilitate the Ward clustering algorithm used in construction of the dendrogram (Pawlowsky-Glahn and Egozcue, 2011), and to allow for the distance metric between the compositional parts to be Euclidean. Unfortunately, however, the centered log-ratio transformation is not completely free of inherent variance structure introduced by compositional closure. The vertical axes represents the total within-cluster error sum of squares, and the clusters are evaluated based on the sum of the squared Euclidean distances between the variables. Compared to the previous naive method of clustering based on pheno-

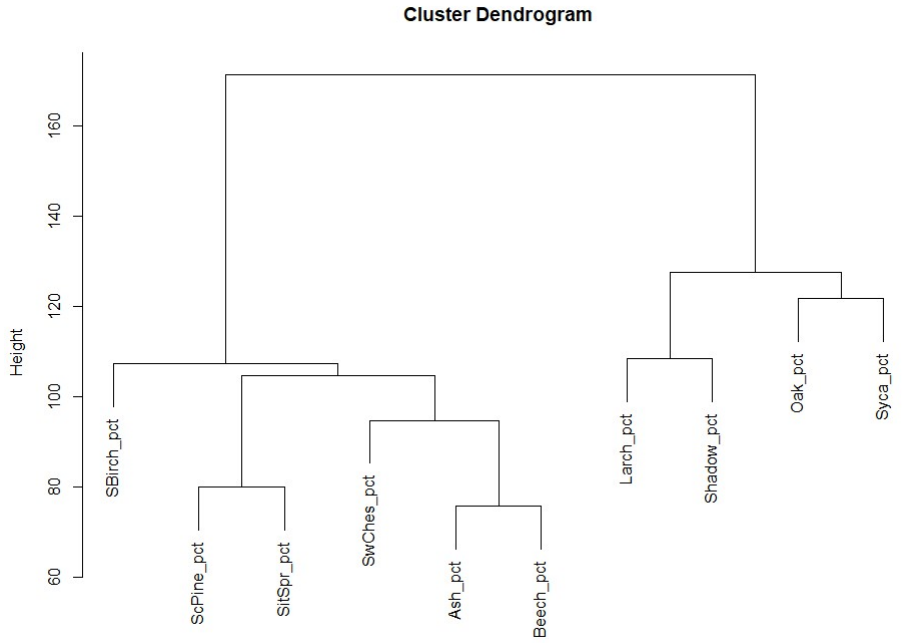


Figure 3.12: Hierarchical cluster dendrogram of *clr*-transformed tree type compositions.

logical insights, the dendrogram's branches do not split at a similar dissimilarity value (height). Perhaps only Silver Birch and the cluster Larch and Shadow class are comparable, however the former seems an outlier with the splits that follow. There also seems no insight to why Larch would be associated with the Shadow class (Figure 3.9 and Figure 3.10). Therefore, from insights into class amalgamations into tree families and more formal hierarchical cluster analysis of the response compositions, there does not seem any motivation to model tree types in groups.

3.3.3 Dimensionality reduction

Having seen no evidence to support dimensionality reduction in the compositional responses, we may now turn towards the explanatory set of variables for the same task. Previously, a strong inter-seasonal correlation was identified: this hints at a temporal component in the data set, with changing leaf colour by season. Alternatively, this could also suggest the existence of multicollinearity, which occurs when one explanatory variable in a multiple regression model can

be linearly predicted using the other variables, which can lead to sensitivity of the model and, hence, predictions. To reduce the number of explanatory variables in the regression and explore the strength of contribution of each spectral band in this multivariate data set to the overall variance, the method of principal variables (Cumming and Wooff, 2007) was implemented. This method can be compared and contrasted with principal component analysis (PCA). The aim of both is to reduce the set of predictor variables, while still maintaining the variability of the data set. In their survey of image classification methods Lu and Weng (2007) provide insight on other methods of dimensionality reduction, and highlight that PCA is advantageous in preserving spectral integrity of the input data set.

In PCA, data points are mapped onto new coordinate axes constructed as orthogonal linear combinations of the predictor variables, and this can make post-hoc interpretation difficult. For principal variables (PV) the sums of squared correlations h_j (for $j = 1, ..., 30$) between a variable of choice v_j and the remaining variables, are examined for each predictor variable in turn. The variable with the highest h_j is chosen and a partial correlation matrix for the remaining variables is calculated, controlling for the contribution of the selected predictor to the correlations between the remaining variables. This process of calculating the partial correlation matrix and selecting the predictor variable with the highest h_j is repeated until a threshold for the proportion of the total variance in the predictor is reached. Akin to PCA, this threshold is selected by the analyst. Then, the predictor variables retaining that cumulative threshold level are retained as the reduced set of variables, and are fed into the random forest regression algorithm.

In the first instance, the analysis was performed on the entire set of spectral bands for all the seasons, with $B_{5,spr}$ giving the largest variance contribution of 41 percent. From then onwards, as seen in Figure 3.13 the bands $B_{6,win}$, $B_{2,sum}$, $B_{7,spr}$, $B_{6,sum}$, $B_{2,win}$, $B_{4,sum}$, $B_{3,sum}$ together explain 90 percent of the total variance.

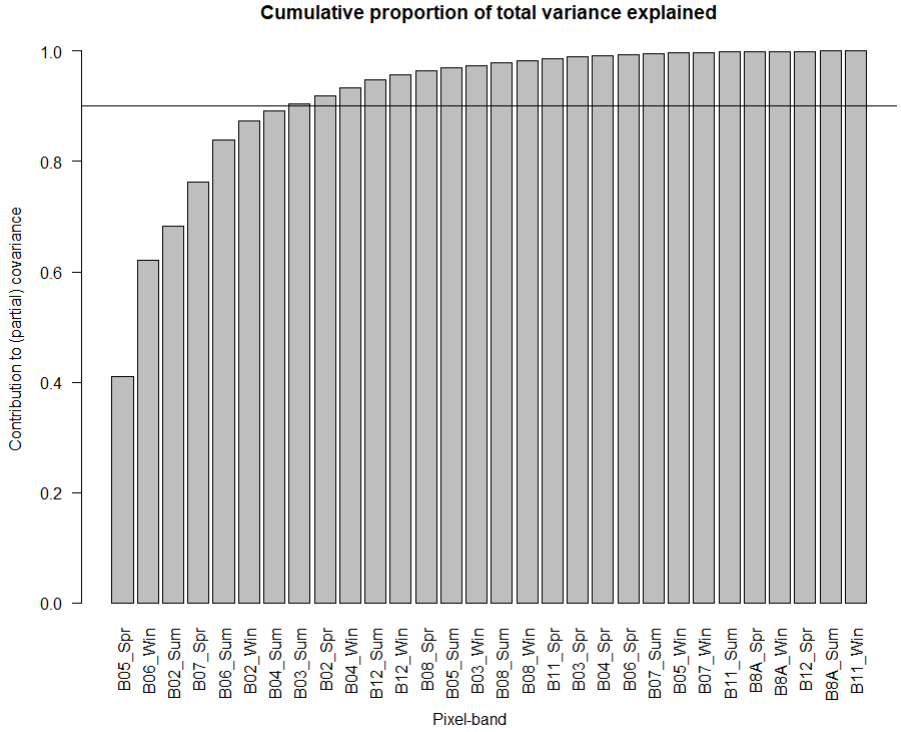


Figure 3.13: Principle variable analysis for entire set of spectral band variables.

For the individual seasons of Winter, Spring and Summer, very similar individual-band contributions could be seen, so the simpler approach driven by Figure 3.13 was used. From Tables 3.8 and 3.9, it can be seen that the new reduced set of variables performs sufficiently well in terms of RMSE, if contrasted with the full data set, and the differences are not significant at 5% level.

Tree Type	Random forest coordinate RMSE (PV 90%)	Random forest RMSE (PV 90 %)
Ash	0.085(0.001)	0.089(0.009)
Beech	0.121	0.124(0.004)
Larch	0.192(0.006)	0.212(0.017)
Oak	0.211(0.003)	0.228(0.002)
Scots Pine	0.102	0.110(0.007)
Shadow	0.107(-0.003)	0.112(0.011)
Silver Birch	0.155(0.007)	0.168(0.011)
Sitka Spruce	0.114(-0.001)	0.119(0.004)
Sweet Chestnut	0.137(0.002)	0.146(0.008)
Sycamore	0.175(0.005)	0.191(-0.062)

Table 3.8: 20 repeated simulations, with test data set for prediction purposes. Numbers in brackets represent the increase (decrease indicated by minus sign) in RMSE to 3 decimal places from the full set of predictors in Table 3.4 with the inclusion of the spatial covariate, where indicated.

Tree Type	RF co-ord RMSE (PV 80%)	RF co-ord RMSE (PV 85%)	RF co-ord RMSE (PV 95%)	RF co-ord RMSE (PV 99%)
Ash	0.082(-0.002)	0.080(-0.004)	0.085(0.001)	0.081(-0.003)
Beech	0.116(-0.005)	0.120(-0.001)	0.119(-0.002)	0.122(0.001)
Larch	0.188(0.002)	0.186	0.196(0.01)	0.192(0.006)
Oak	0.209(0.001)	0.211(0.003)	0.213(0.005)	0.214(0.006)
Scots Pine	0.102	0.103(0.001)	0.104(0.002)	0.104(0.002)
Shadow	0.109(-0.001)	0.109(-0.001)	0.106(-0.004)	0.107(-0.002)
Silver Birch	0.149(0.001)	0.151(0.003)	0.153(0.005)	0.155(0.007)
Sitka Spruce	0.115	0.116(0.001)	0.112(-0.003)	0.116(0.001)
Sweet Chestnut	0.142(0.007)	0.139(0.004)	0.138(0.003)	0.141(0.006)
Sycamore	0.172(0.002)	0.171(0.001)	0.175(0.005)	0.173(0.003)

Table 3.9: 20 repeated simulations, with test data set for prediction purposes. Numbers in brackets represent the increase (decrease indicated by minus sign) in RMSE to 3 decimal places from the full set of predictors in Table 3.4 with the inclusion of the spatial covariate, where indicated.

3.3.4 Multivariate regression on transformed response data

A classical approach to performing a regression analysis on compositional data relies on transforming the appropriate compositional variables to lie on the Euclidean space and not on the simplex space. This is done by performing one of the log-ratio transformations on the data, of which there are the isometric log-ratio (ilr), the additive log-ratio (alr) and centred log-ratio (clr), defined in Chapter 2. In basic analysis of compositional data the log-ratio transforms are taken to follow the multivariate Gaussian distribution, and so we can proceed with the multivariate regression (Aitchison, 1986).

A further challenge to tackle in this case before performing the log-ratio transforms is the presence of essential zeros in the response variable (tree types), as they constitute a large proportion of each variable, as presented in Table 3.3. Those components are not zeros due to measurement error, but by absence of a characteristic leaf colour in a certain pixel. For the moment we consider a naive strategy of non-zero replacement, whereby we add a small quantity (0.005) to each of entry of tree type response, and a similar replication exercise with a 80 percent training subset as in the above techniques.

One of the examples of Bayesian approaches to variable selection in a regression setting can be the Bayesian Lasso (Ročková and George, 2018), which incorporates a penalisation factor in the regression specification with a Laplace prior over the regression coefficients β. However, again therein lies a strong assumption of the underlying Gaussian distribution, so a choice of transformation on the compositional parts is required.

In any case, let us suppose that a log-ratio transformation has been carried out on the original compositional responses. We can define the transformed variable $\boldsymbol{Y_T}$ and assume that now the multivariate Gaussian assumptions behind regression analysis hold

$$\boldsymbol{Y_T} = \boldsymbol{XB} + \boldsymbol{E}. \tag{3.3}$$

where $\boldsymbol{X}$ is the matrix of real-valued predictors, $\boldsymbol{B} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_D)$ are the regression coefficients and $\boldsymbol{E} \sim N(\boldsymbol{0}, \mathrm{ID})$ is the residual matrix. As usual D denotes the dimension of the problem, in this scenario $D = 10$ and the dimensionality decreases by one unit when the alr transformation is performed. Tsagris (2015) recognised that with the alr transformation the modelling problem above is reduced to a multivariate regression, where the first transformed component is like an offset term with a β coefficient equal to 1. To note, the ilr transformation is also a typical choice in regression problems with compositional response, however, in this scenario it was deemed less appropriate due to lack of clear hierarchical structure in the tree types, as explored in the previous sections.

Regression analysis was carried out subject to the alr transform of the compositional tree type responses, and the Shadow class used as the reference component. This was done for several reasons: firstly, little information is carried in the shadow class, apart from the absence of detection of any other tree type at the AOI. Secondly, the Shadow category has the fewest essential zero values than the other classes, which is advantageous to this transformation approach. Nevertheless, a zero enhancement was performed to the set of response tree types, as described earlier. A multivariate Gaussian model fitted to the data dissected into 80-20 train-test ratio as previously and 20 runs of the model gave RMSE scores in Table 3.10. Residual analysis (Appendix A, Figures 9.7-9.12) was also carried out and did not find significant autocorrelation. The normality of alr-transformed data and the resulting model errors (residuals) is also considered. It can be seen that the transformed tree types are right-skewed, which is driven by the high proportion of essential zeros in the original data. This effect is similarly seen in the residual vs. fitted values plots, although there does not seem evidence for an obvious lack of homoscedasticity in residuals. Also, very few of the estimated β coefficients were deemed statistically significant at the 5% level. Hence, it is not certain that the multivariate Gaussian fit to the tree proportion data including the spatial X-Y coordinates is deemed a suitable parametric modelling approach,

and the RMSE scores are higher if compared with the random forest regression for tree types like Ash.

Zhang and Shi (2020) found that random forest in conjunction with alr-transformed compositional data outperformed other machine learning techniques, albeit their compositional data-set was three-part and did not contain essential zero values. Here we too performed an alr-transformation followed by random forest regression with spatial co-ordinate variables and all seasonal spectral bands, with results found in Table 3.10. Again, compared with the original random forest set-up, this approach did not show improvement in RMSE scores for several tree types. This could be driven by the zero-imputation procedure to relieve the number of essential zeros in the compositional parts.

3.3.5 $\alpha - k$-nearest neighbours regression

Another transformation of interest for the application of tree type modelling is the α-power transformation defined in Chapter 2, Section 6. We have seen that the parametric models relying on log-ratio transformations and imputation of zero values do not perform as well in terms of prediction, as non-parametric models like random forest. Tsagris et al. (2021) link a further non-parametric regression model, the k-nearest neighbours k-N N smoother due to Fix and Hodges (1951) with the α-power transform that is accommodating to essential zeros in a compositional variable. Like random forests, the k-N N smoother can be applied to both regression and classification tasks, with our interest lying in the former. The algorithm works by again splitting the original data set into training and testing samples, labelled X_{TR}, X_{TEST}, Y_{TR} and Y_{TEST}. Then, a distance measure is computed between X_{TR}, X_{TEST}, usually the Euclidean or the Mahalanobis norm. The k smallest distances are selected, and k is determined by cross-validation procedure. These smallest distances are associated with X_{TR} and corresponding Y_{TR} values. Then, an average value of those k-smallest associated Y_{TR} values is taken as a prediction for Y_{TEST}. The original k-N N rests on the sample mean as

a measure of the average, whereas $\alpha - k$-N N extends this to the Fréchet mean (Tsagris et al., 2021) and this definition of the average measure works for the α-power transformation. The Fréchet average stems from the Fréchet distribution - one of the distribution choices used for modelling extreme values in the tails of distributions. The Fréchet distribution has some associations with the simplex space, and as the power α approaches zero, the Fréchet mean converges to the geometric mean, a familiar measure in compositional data analysis. $\alpha - k$-N N regression exploits this relationship with compositional data and allows us to bypass log-ratio modelling of compositional parts. More details about $\alpha - k$-N N regression can be found in Tsagris et al. (2021).

A natural question may arise on the values of α and k that are needed for this regression to reach peak performance for a particular data set. The specification of α and k is done through a cross-validation procedure, where several candidate values are compared and the ones driving the smallest error measure are selected. In this applied example, analysis was carried out using the `Compositional` package in `R`, and the cross-validation procedure yielded $k = 2$ closest neighbours to be considered, along with $\alpha = 1$ which means that the original compositional tree types are not power-transformed. RMSE scores for this non-parametric approach can be found in Table 3.10. Compared to the other methods presented in Table 3.10, α-k-N N RMSE scores are significantly different at the 5% level, except for Ash and Sweet Chestnut in comparison to clr-transformed MVN regression RMSE. A plausible reason for this improvement is k-N N's use of neighbouring pixel points for prediction purposes, thus incorporating a spatial clustering element, and tuning of the power transform α and k to yield smaller error.

Tree Type	alr multivariate Gaussian regression RMSE	clr multivariate Gaussian regression RMSE	alr random forest RMSE	α-k-N N regression RMSE
Ash	0.313	0.084	0.280	0.079
Beech	0.148	0.147	0.151	0.133
Larch	0.318	0.329	0.332	0.220
Oak	0.334	0.310	0.362	0.235
Scots Pine	0.240	0.147	0.292	0.121
Shadow	0.173	0.252	0.175	0.128
Silver Birch	0.198	0.200	0.150	0.183
Sitka Spruce	0.209	0.208	0.111	0.143
Sweet Chestnut	0.152	0.152	0.132	0.158
Sycamore	0.234	0.234	0.168	0.206

Table 3.10: 20 repeated simulations, with test data set for prediction purposes. RMSE scores for alr-transformed tree types followed by multivariate Gaussian regression and random forest regression, and the α-power transform k-nearest neighbour approach.

3.4 Discussion and conclusions

This work has demonstrated machine learning and compositional regression techniques in the field of sub-pixel classification. Namely, the role of data set dimensionality reduction through principal variables has become evident in order to increase efficiency of the random forest algorithm and introduce a spatial element into the analysis. Using spectral bands from different seasons aids in capturing phenological differences between species and the bands chosen fit well with other studies - Ottosen et al. (2020) selected bands 2, 3, 6, and 12 from Sentinel 2 for mapping tree cover. A further extension to this work is to incorporate species composition from multiple sources.

The results from comparison of parametric and non-parametric methodology on compositional responses highlighted the problem of essential zero values, which have been abundant in this data set. Zero-replacement did not show an improvement of RMSE scores compared with non-parametric methods. Inference for the multivariate Gaussian regression did not highlight particular pixel bands as significant predictors of the responses, however, residual analysis did not highlight any issues with the underlying Normality assumption. While performing random forest on the full set of predictors took a notable amount of time, it was possible to increase turnaround speed with the use of principal variables to reduce the size of the explanatory variable set. Attempts at clustering of compositional tree types was also performed and did not find an obvious pattern. This is also confirmed by spatial plots of the UAS classification, as some tree types (for example, Larch) occupy significant regions of the southern Y-coordinate space and are not met by an equal proportion of another tree type in the same pixels.

Overall, we have demonstrated that lower-resolution Sentinel-2 data can be used in order to predict tree type based spectral band colours in mixed pixels, with reasonable accuracy. Further attempts were made to construct a regression model in the simplex space using the Dirichlet distribution - this approach is known

as Dirichlet regression (Hijazi and Jernigan, 2009) and has been demonstrated to be successful when the regression model contains compositional explanatory variables and a univariate response vector. Further work is required in this area for modelling multivariate compositional responses. The authors similarly note that the training data is limited to one site only over five temporal measurements, however the analysis conducted shows results close to that of the UAS hard classification and illustrates the issues of sub-pixel classification using an area of contiguous woodland. Further exploration on a more complex example involving multiple woodlands across a larger extent in both geography and time periods covered would be the next step in assessing the impact of spatial and statistical properties of sub-pixel methods in prediction.

Chapter 4

Multivariate distributions on the simplex

4.1 Introduction

The aim of this chapter is to provide an overview of multivariate probability distributions on the simplex support. These distributions are considered in light of being used in an expert elicitation exercise, to yield possible parameters for the construction of prior distributions in a Bayesian analysis. Hence, our interest lies especially in highlighting the following areas: firstly, number of parameters required for the definition of each distribution, since this is directly linked to the complexity and length of the elicitation exercise and hence, and cognitive strain on the experts. In Chapter 7, we discuss in more detail the balance between using a more flexible distribution in an elicitation exercise, which may carry more parameters, and any possible improvement on reflecting the expert opinions. Second, in this chapter, we specify marginal and conditional probability distributions for each multivariate distribution we consider on the simplex, again, to lead up to the task of elicitation exercise. Furthermore, we evaluate the ease of computation of quantiles and measures of dispersion of these multivariate distributions.

4.2 The Dirichlet family of distributions

The first family of distributions under consideration in this chapter is the Dirichlet family - conjugate to multinomial likelihood, which makes it a straightforward choice for modelling uncertainty about compositional data within a Bayesian framework. Before Aitchison's work in the 1980s, this class of distributions was considered to be the sole appropriate way to model compositional data. Its main drawback, however, is its inherent negative correlation structure, and further generalisations of the Dirichlet distribution have been sought to accommodate flexibility in expressing dependence between the random variables. Such distributions are similarly presented in this section, starting from Connor and Mosimann's work, through to more recent developments that are useful in situations where separate parts of the simplex need to be modelled.

4.2.1 Dirichlet distribution

The definition of the $D-$simplex Δ^D allows us to consider the Dirichlet distribution, which is part of the exponential family of distributions and is conjugate to multinomial likelihood. It is considered as a multivariate case of the continuous Beta distribution and defined on the simplex support. The Dirichlet distribution can be derived from two Gamma distributed variables, as follows:

Let two random variables Y_1 and Y_2 follow the Gamma distribution with a shared scale parameter β : $Y_1 \sim \text{Gamma}(\alpha_1, \beta), Y_2 \sim \text{Gamma}(\alpha_2, \beta)$. The probability density function of Y_1, for example, is $f(y_1; \alpha_1, \beta) = \frac{{y_1}^{\alpha_1 - 1} e^{-y_1/\beta}}{\beta^{\alpha_1}\Gamma(\alpha_1)}$ for $y_1 > 0$ and $\alpha_1, \beta > 0$, where Γ denotes the gamma function.

Then, setting $X = \frac{Y_1}{Y_1+Y_2}$ yields that $X \sim \text{Beta}(\alpha_1, \alpha_2)$. To extend this further, if we let $Y_1, ..., Y_{D-1}$ be independent Gamma variables of the form above and set $X_i = \frac{Y_i}{Y_1+...+Y_{D-1}}$ for $i = 1, ..., D-1$, then, jointly, the vector $\boldsymbol{X}$ follows the

probability density

$$\pi(x_i) = \frac{\Gamma(\sum_{i=1}^{D} \alpha_i)}{\prod_{i=1}^{D} \Gamma(\alpha_i)} \prod_{i=1}^{D} x_i^{\alpha_i - 1}. \tag{4.1}$$

Conjugacy of the Dirichlet distribution to the multinomial likelihood can be seen through the following:

Let $n_1, ..., n_D$ be the frequencies for D distinct categories, where $n_i \in \mathbb{Z}_+^D$ and $\sum_{i=1}^{D} n_i = N, i = 1, ..., D$. Then let $x_1, ..., x_D$ be the probabilities of obtaining the respective categories, and $x_i \in [0, 1]$ and $\sum_{1}^{D} x_i = 1$. It follows that the distribution of $n_1, ..., n_D$ is multinomial with the following probability mass function

$$f(n_1, ..., n_D | x_1, ..., x_D, N) = \frac{N!}{\prod_{i=1}^{D} n_i!} \prod_{i=1}^{D} x_i^{n_i} \propto \prod_{i=1}^{D} x_i^{n_i}. \tag{4.2}$$

In the multinomial distribution, x_i is a probability for each class i the data can be assigned to. If the prior distribution has Dirichlet kernel such that

$\pi(\boldsymbol{X}|\boldsymbol{\alpha}) \propto \prod_{i=1}^{D} X_i^{\alpha_i - 1}$, then the posterior distribution follows

$$\begin{aligned} \pi(\boldsymbol{X}|N, \boldsymbol{\alpha}) &\propto \mathbb{P}(N|\boldsymbol{X}) \pi(\boldsymbol{X}|\boldsymbol{\alpha}) \\ &\propto \prod_{i}^{D} X_i^{\alpha_i + n_i - 1}. \end{aligned} \tag{4.3}$$

which we recognise as another Dirichlet kernel, so the posterior distribution follows Dirichlet$(\alpha_1 + n_1, \alpha_2 + n_2, ...\alpha_D + n_D)$.

The parameter $\boldsymbol{\alpha} = (\alpha_1, .., \alpha_D) > 0$ is known as the concentration parameter. Further, the sum of the parameters $\alpha_0 = \sum_{i=1}^{D} \alpha_i$ is known as the standardised Dirichlet precision factor and governs how concentrated the distribution is around its mean vector. Hence, a high α_0 value indicates a high peak of the distribution centred around the mean vector and in the cases where $\alpha_i < 1$ the concentrations would be found at the corners of the simplex. Generally, each α_i determines which regions of the D-simplex have the most probability mass, so the α_i values are relative to one another.

Plots in Figure 4.1 depict a graphical representation of the Dirichlet distribution over a simplex with varying values of $\boldsymbol{\alpha}$. Where values of α_i are equal, we are imposing that the three outcomes have equal probability of success, so should be symmetric over the simplex. For $\alpha >> 1$, the values of the distribution approach the geometrical centre of the simplex. The last plots show a case of asymmetry, weighted by the greater value of α.

If $\boldsymbol{X} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, then

$$\text{mode}(\mathbf{X}) = \frac{\alpha_i - 1}{\alpha_0 - D}, \alpha_i > 1. \tag{4.4}$$

For the expected value of $\boldsymbol{X} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, let us show the derivation for $\mathbb{E}(X_1)$, which generalises to $\mathbb{E}(X_i) = \frac{\alpha_i}{\alpha_0}, i = 1, ..., D$.

$$\begin{aligned}
\mathbb{E}(X_1) &= \int ... \int x_1 \frac{\Gamma(\sum_{i=1}^{D} \alpha_i)}{\prod_{i=1}^{D} \Gamma(\alpha_i)} \prod_{i=1}^{D} x_i^{\alpha_i - 1} dx_1 ... dx_D \\
&= \int ... \int \frac{\Gamma(\sum_{i=1}^{D} \alpha_i)}{\prod_{i=1}^{D} \Gamma(\alpha_i)} x_1 x_1^{\alpha_1 - 1} \prod_{i=2}^{D-1} x_i^{\alpha_i - 1} (1 - \sum_{i=1}^{D-1} x_i)^{\alpha_D - 1} dx_1 ... dx_{D-1} \\
&= \frac{\Gamma(\sum_{i=1}^{D} \alpha_i)}{\Gamma(\alpha_1) \prod_{i=2}^{D} \Gamma(\alpha_i)} \frac{\Gamma(\alpha_1 + 1) \prod_{i=2}^{D} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{D} \alpha_i + 1)} \\
&= \frac{\Gamma(\sum_{i=1}^{D} \alpha_i)}{\Gamma(\sum_{i=1}^{D} \alpha_i + 1)} \frac{\Gamma(\alpha_1 + 1)}{\Gamma(\alpha_1)} \\
&= \frac{\alpha_1}{\sum_{i=1}^{D} \alpha_i} = \frac{\alpha_1}{\alpha_0}.
\end{aligned} \tag{4.5}$$

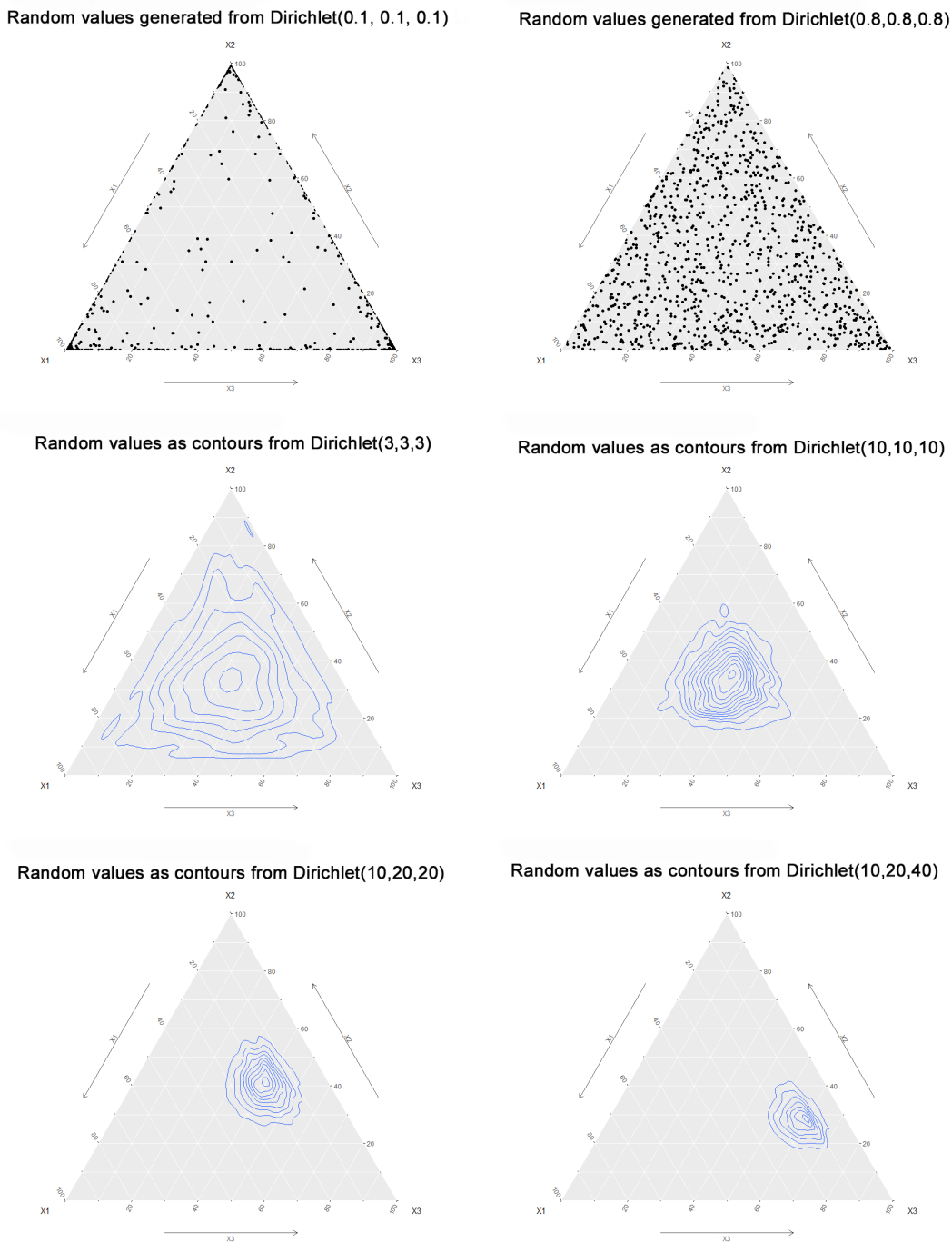


Figure 4.1: Contours of the Dirichlet density for 3 variables and varying $\boldsymbol{\alpha}$ parameter vectors.

and through similar arguments and definition of the second moment is

$$\mathbb{E}(X_i^2) = \frac{\alpha_i(1+\alpha_i)}{(\sum_{i=1}^{D}\alpha_i + 1)\sum_{i=1}^{D}\alpha_i}, \tag{4.6}$$

$$\mathrm{Var}(X_i) = \frac{\frac{\alpha_i}{\alpha_0}(1-\frac{\alpha_i}{\alpha_0})}{1+\alpha_0}, \tag{4.7}$$

$$\mathrm{Cov}(X_i, X_j) = \frac{-\alpha_i\alpha_j}{\alpha_0^2(1+\alpha_0)}; i \neq j, \tag{4.8}$$

where $\alpha_0 = \sum_{i=1}^{D}\alpha_i$.

This is due to

$$\mathbb{E}(X_iX_j) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+2)}\frac{\Gamma(\alpha_i+1)\Gamma(\alpha_j+1)}{\Gamma(\alpha_i)\Gamma(\alpha_j)} = \frac{\alpha_i\alpha_j}{(\alpha_0+1)\alpha_0}.$$

Since $Cov(X_i, X_j) = \mathbb{E}(X_iX_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)$, we obtain

$$Cov(X_i, X_j) = \frac{\alpha_i\alpha_j}{(\alpha_0+1)\alpha_0} - \frac{\alpha_i\alpha_j}{\alpha_0^2} = \frac{-\alpha_i\alpha_j}{\alpha_0^2(1+\alpha_0)}.$$

The marginal distribution can be derived by considering the joint density of $x_1, ..., x_D$, yielding

$$f(\boldsymbol{x}) = f_1(x_1)f_2(x_2|x_1)...f_{D-1}(x_{D-1}|x_1, ..., x_{D-2}),$$

where $f(\cdot)$ is the Dirichlet density function.

Then marginally, $X_i \sim \mathrm{Beta}(\alpha_i, \alpha_0 - \alpha_i)$, and the summary statistics follow from the standard Beta distribution. Otherwise, determining the median values for the Dirichlet case is not a trivial matter, since symmetry does not seem to hold, hence it is infeasible to consider intersection of the marginal hyperplanes (Small, 1997).

We could consider solely the vector of medians arising from the D marginal

Beta distributions without the interplay into multivariate distributions. All these nuances are again reflected in the process of expert elicitation, as is depicted in Chapter 7, as we explore the idea of elicitation of multivariate distributions even without the simplex constraints.

For the moment, however, let us reflect on the Dirichlet distribution. This distribution has D parameters, and from the definitions above we see that the parameters are only constrained to $\mathbb{R}^+$. However, due to the unit-sum constraint on proportions, the final parameter α_D is deterministic. On the other hand, adhering to principles of compositional data analysis, the order of compositions need not be fixed, which, in the exercise of expert elicitation, would allow for some degree of variability of the final parameters values.

The Dirichlet distribution has proven to be a classic choice for modelling on the simplex, as seen in numerous works (Mateu-Figueras and Tolosana-Delgado, 2006; Ng et al., 2011). However, Aitchison criticised this choice, deeming it "inadequate for the description of the variability of compositional data" due to the Dirichlet's implied independence structure between some compositional parts and considered this distribution to struggle to accurately model compositions whose components possess even weak forms of dependence. Aitchison advised towards the use of log-transformations and the logistic Gaussian distribution, which has more parameters for tuning the covariance structure between components.

Frigyik et al. (2010) introduce two further distributions closely related to the Dirichlet, which are used to model subsets of the simplex. They motivate such applications where the likelihood is only relevant to a particular region of the simplex. An example could be the analysis of one of the categories which exceeds some threshold proportion, for instance, it is greater than the value of 0.5. They consider a previous approach trialled by Nallapati, Ahmed, Cohen and Xing (2007), which is achieved by a separate normalisation of the Dirichlet distribution by restricting the support of the Dirichlet from the full simplex Δ to a restricted

simplex region $\tilde{\Delta}$, and then the Dirichlet probability density is re-normalised over $\tilde{\Delta}$. One problem arising from this is the computational cost of finding the normalisation factor, and further, the summary statistics (mean, covariance, mode) have no closed form, which makes them troublesome for modelling, let alone an exercise such as expert elicitation. Similarly, the estimation of $\boldsymbol{\alpha}$ parameters is costly through the Maximum Likelihood approach.

4.2.2 Connor & Mosimann (generalised Dirichlet) distribution

A more generalised version of the Dirichlet was introduced by Connor and Mosimann (1969). It is equally known as the Connor-Mosimann distribution and can be constructed in the following way:

Definition 4.1. *(Generalised Dirichlet distribution) If we let* $Z_j \sim Beta(\alpha_j, \beta_j)$ *for* $j = 1, ..., D-1$ *and for the remaining component* $Z_D = 1$*, then for* $j = 2, ..., D$*:* $P_1 = Z_1$,

$$P_j = Z_j \prod_{i=1}^{j-1}(1 - Z_i). \tag{4.9}$$

and its probability density given by the following (Connor and Mosimann, 1969):

$$\pi(x_1, ..., x_D) = \prod_{i=1}^{D-1} \left[\frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} x_i^{\alpha_i - 1} (\sum_{i=1}^{D} x_j)^{\beta_{i-1} - \alpha_i - \beta_i} \right] x_D^{\beta_{D-1}-1}. \tag{4.10}$$

The covariance structure of the Connor-Mosimann distribution is more flexible than that of the Dirichlet distribution, for which a negative correlation between any pair of compositions is imposed. In the case of the Connor-Mosimann distribution, however, this is only true for pairing of the first composition with any other one in the set $\boldsymbol{X}$, X_1 and X_j, where $j = 2, ..., D$. This holds due to $\text{Cov}(X_1, X_j) = \frac{\mathbf{E}(X_j)}{\mathbf{E}(1-X_1)}\text{Var}(X_1)$ for $j = 2, ..., D$.

However, for other successive covariates, this correlation can be positive, and

more generally sign(Corr(X_j, X_m))=sign(Cov(X_j, X_{j+1})) for $1 < j < m \leq D$ (Connor and Mosimann, 1969).

4.2.3 Modified Connor-Mosimann distribution

In the Dirichlet distribution, the number of parameters is the same as the number of dimensions. The Connor-Mosimann distribution has one fewer parameter, each of which is composed of hyperparameters. When exposed to multinomial likelihood this distribution is conjugate. Further work by Wilson (2017) proposes a modification to the Connor-Mosimann to increase flexibility by utilising scaled Beta distributions, which yields four times the previous number of hyperparameters. It is defined in a similar manner to the above, only now Z_j has two additional parameters in the scaled Beta distribution, A_j and B_j; $j = 1, ..., D-1$, $0 < A_j, B_j < 1$. Formally, the probability density function of the scaled Beta distribution is

$$\pi(x_j) = \Big|\frac{1}{B_j - A_j}\Big|\frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)}\Big(\frac{x_j - A_j}{B_j - A_j}\Big)^{\alpha_j - 1}\Big(1 - \frac{x_j - A_i}{B_j - A_j}\Big)^{\beta_j - 1}; x_i \in [A_j, B_j]. \tag{4.11}$$

where $|\cdot|$ is the modulus function. The marginal distributions of the modified Connor-Mosimann distribution function follow the scaled Beta distribution, which relates to the unscaled Beta as follows:

$$X \sim \text{ScaledBeta}(\alpha, \beta, A, B) = A + (B - A)\text{Beta}(\alpha, \beta); \tag{4.12}$$

where parameters $A, B \in \mathbb{R}$ are used to re-scale the Beta distribution to the interval $[A, B]$. This modified Connor Mosimann form is not conjugate when exposed to the multinomial likelihood.

This likelihood has the form as in Equation 4.2 and hence the posterior is proportional to

$$(\frac{x_i - A_i}{B_i - A_i})^{\alpha_i - 1} x_i^{n_i} (1 - \sum_{i=1}^{D} x_i)^{n_i} (1 - \sum_{i=1}^{D} \frac{x_i - A_i}{B_i - A_i})^{\gamma_i}.$$

where $\gamma_m = \beta_i - \alpha_{m+1} - \beta_{m+1}$ for $m = 1, ..., D-1$ and $\gamma_D = \beta_D - 1$.

4.2.4 Shifted-Scaled Dirichlet distribution

The scaled Dirichlet distribution is a further generalisation of the Dirichlet distribution and is due to Dickey (1968). It finds close relation to the multivariate scaled Beta distribution as outlined in Equation 4.11. The probability density function of the scaled Dirichlet distribution can again be achieved through normalisation of appropriately scaled Gamma-distributed variables, but its analytic form is given by

$$\pi(x_1, ..., x_D) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{D} \Gamma(\alpha_i)} \frac{\prod_{i=1}^{D} {\beta_i}^{\alpha_i} {x_i}^{\alpha_i - 1}}{(\sum_{i=1}^{D} \beta_i x_i)^{\alpha_0}}. \tag{4.13}$$

where $\alpha_0 = \sum_{i=1}^{D} \alpha_i$. $\pi(X_1, ..., X_D)$ as expressed above is denoted by authors as $\boldsymbol{X} \sim \boldsymbol{SD}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$. $\boldsymbol{X}$ can be reduced to the Dirichlet distribution if $\boldsymbol{\beta} = (1, 1, ..., 1)$. The scaled Dirichlet distribution has $2D$ parameters and the authors note that the scaled Dirichlet distribution is translation of the Dirichlet density in the simplex space, and thus belongs to the same family of distributions. The marginal distributions for $\boldsymbol{X} \sim \boldsymbol{SD}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are defined in a similar sense, and for a bivariate case $D = 2$ are given by

$$\pi(x) = \frac{1}{B(\alpha_1, \alpha_2)} \frac{\beta_1^{\alpha_1} x^{\alpha_1 - 1} \beta_2^{\alpha_2} (1 - x)^{\alpha_2 - 1}}{(\beta_1 x + \beta_2 (1 - x))^{\alpha_1 + \alpha_2}}. \tag{4.14}$$

The scaled Dirichlet distribution is not conjugate with the multinomial likelihood, similarly its covariance structure is not of closed form (Monti et al., 2011).

There is also no closed form for mode($\boldsymbol{X}$), $\mathbb{E}(\boldsymbol{X})$ and expressions for cumulative distribution function and quartiles with respect to the Lebesgue measure in real space. Estimation of parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is completed through application of the Expectation-Minimisation algorithm (Alsuroji, 2018) until convergence is achieved. Monti et al. (2011) further investigate a distribution closely linked to the aforementioned, but containing another vector of parameters that is used as a scaling factor of the original Dirichlet distribution. Appropriately named the shifted-scaled Dirichlet distribution, it involves applying the powering operation, as well as the perturbation operation to a Dirichlet random composition.

Definition 4.2. *A random vector* $\boldsymbol{X}$ *has a shifted-scaled Dirichlet distribution* $\boldsymbol{X} \sim \boldsymbol{pSD}^D(\boldsymbol{\alpha}, \boldsymbol{p}, a)$ *with parameters* $\boldsymbol{\alpha} \in \mathbb{R}_+^D, \boldsymbol{p} \in \Delta^D$ *compositional vector representing a shift parameter, and* $a \in \mathbb{R}_+$*, if its probability density function takes the form*

$$\pi(x_1, ..., x_D) = \frac{\Gamma(\alpha_0)}{a^{D-1} \prod_{i=1}^D \Gamma(\alpha_i)} \frac{\prod_{i=1}^D p_i^{-\alpha_i/a} x_i^{-1+(\alpha_i/a)}}{(\sum_{i=1}^D (\frac{x_i}{p_i})^{1/a})^{\alpha_0}},$$

where $\alpha_0 = \sum_{i=1}^D \alpha_i$.

The marginal distributions for $\boldsymbol{X} \sim \boldsymbol{pSD}^D(\boldsymbol{\alpha}, \boldsymbol{\beta}, a)$ are defined in a similar sense, and for a bivariate case $D = 2$ are given by

$$\pi(x) = \frac{1}{aB(\alpha_1, \alpha_2)} \frac{\frac{x^{(\alpha_1/a)-1}}{p_1^{\alpha_1/a}} \frac{(1-x)^{(\alpha_2/a)-1}}{p_2^{\alpha_2/a}}}{((x_1/p_1)^{1/a} + ((1-x)/p_2)^{1/a})^{a_0}}. \tag{4.15}$$

As for the scaled Dirichlet distribution, no closed form solution exists for $\mathbb{E}(\boldsymbol{X})$, mode($\boldsymbol{X}$), nor the covariance structure. The authors instead suggest numerical integration techniques. Similarly, it is not conjugate to the multinomial likelihood and this is again due to the terms in the denominator of the probability density function. The motivation for construction of these probability distributions was to increase flexibility in modelling regions of the simplex, where a classical Dirichlet distribution is not sufficient. This was achieved through using

the known perturbation (translation) and powering (scaling) operations on the simplex, as defined by Aitchison (1986). Even though joint and marginal probability density functions for these distributions can be expressed in the Lebesgue measure, the moments and covariance structures require numerical integration. No use of these distributions in a Bayesian setting has been reported, apart from being included in a Dirichlet mixture model with application to Covid-19 data (Bourouis et al., 2021).

4.2.5 Extended Flexible Dirichlet distribution

Another approach to generalise the Dirichlet distribution is done through a finite mixture (linear combination) of Dirichlet-distributed variables. The probability density function for the Flexible Dirichlet (FD) distribution can be expressed as

$$\pi_{FD}(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{p},\tau)=\frac{\Gamma(\alpha_0+\tau)}{\prod_{h=1}^{D}\Gamma(\alpha_h)}\Big(\prod_{h=1}^{D}x_i^{\alpha_h-1}\Big)\sum_{i=1}^{D}p_i\frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i+\tau)}x_i^{\tau}. \tag{4.16}$$

where $\boldsymbol{x}\in\Delta^D$, $\alpha_0=\sum_{i=1}^{D}\alpha_i$, $\sum_{i=1}^{D}p_i=1$ and $\tau>0$. In the above, there are two separate indices i and h due to the construction of the Flexible Dirichlet: the basis of independent Gamma distributed random variables is normalised, as in the Dirichlet case, and then the i^{th} element chosen at random is assigned one further independent Gamma variable. The above representation holds on the Lebesgue measure if we set the final component $x_D=1-x_1-\ldots-x_{D-1}$. If we set $\tau=1$ and $p_i=\alpha_i/\alpha_0$ the original Dirichlet probability density is retrieved. The variable τ can be seen to control the number of modes of the Flexible Dirichlet density, in fact, any number of modes up to dimensionality D can be achieved. Therein also lies a suitable choice for p and $\boldsymbol{\alpha}$. This distribution contains $2D+1$

parameters and the first moments of the Flexible Dirichlet are expressed as:

$$\mathbb{E}(X_i) = \frac{\alpha_i + p_i\tau}{\alpha_0 + \tau}; \tag{4.17}$$

$$\mathrm{Var}(X_i) = \frac{\mathbb{E}(X_i)(1 - \mathbb{E}(X_i))}{\alpha_0 + \tau + 1} + \frac{\tau^2 p_i(1 - p_i)}{(\alpha_0 + \tau)(\alpha_0 + \tau + 1)}; \tag{4.18}$$

$$\mathrm{Cov}(X_i, X_j) = -\frac{\mathbb{E}(X_i)\mathbb{E}(X_j)}{\alpha_0 + \tau + 1} - \frac{\tau^2 p_i p_j}{(\alpha_0 + \tau)(\alpha_0 + \tau + 1)}; i \neq j. \tag{4.19}$$

Ongaro and Migliorati (2013) highlight that the Flexible Dirichlet distribution has a more flexible dependence structure than the Dirichlet distribution, albeit the covariance is still negative due to the sum-one constraint. Similarly, this distribution allows for multimodality, which cannot be accommodated by the Dirichlet. Contrary to the shifted-scaled Dirichlet distribution, the moments and covariance structure of Flexible Dirichlet have closed form, so can prove useful in an expert elicitation exercise. The Flexible Dirichlet distribution similarly allows for inclusion of essential zeros in a compositional vector - this is achieved through setting one of $\alpha_i = 0$, and this implies that p_i now reflects the probability that the i^{th} component is strictly positive (Ongaro and Migliorati, 2013)

Ongaro and Migliorati follow to extend the Flexible Dirichlet distribution in Ongaro and Migliorati (2014). The Extended FD distribution (EFD) is derived by augmenting to the basis of FD, only now instead of the i^{th} variable being assigned a separate Gamma-distributed part, a Gamma-distributed random variable is added to each component of the basis. The obvious difference between FD and EFD is that for the former τ variable, expressed as a single positive real number, the EFD includes τ_i, for $i = 1, ..., D$ as an exponent for every x_i. Analytically, the probability density function for the EFD is seen below:

$$\pi(\boldsymbol{x}; \boldsymbol{\alpha}, \boldsymbol{p}, \boldsymbol{\tau}) = \frac{1}{\prod_{h=1}^{D}\Gamma(\alpha_h)}\left(\prod_{h=1}^{D} x_h^{\alpha_h - 1}\right)\sum_{i=1}^{D}\frac{\Gamma(\alpha_i)\Gamma(\alpha_0 + \tau_i)}{\Gamma(\alpha_i + \tau_i)}x_i^{\tau_i}p_i. \tag{4.20}$$

with h and α_0 as defined previously.

Similarly to FD, this distribution is conjugate with the multinomial likelihood and its moments can be expressed in closed form:

$$\mathbb{E}(X_i) = \alpha_i \sum_{h=1}^{D} \frac{p_h}{\alpha_0 + \tau_h} + \tau_i \frac{p_i}{\alpha_0 + \tau_i}; i = 1, ..., D. \tag{4.21}$$

The EFD is seen more favourable to the FD for reflecting clusters in compositional data sets, and when it is also desired that the size of the composition is considered. However, the authors recognise that more insight is required into the precise properties of marginal and conditional distributions, and the covariance structure (Ascari et al., 2017). Similarly, parameter estimation even via expectation–maximization (EM) algorithms is not stable due to the presence of multiple local maxima, thus careful consideration needs to be given to the starting values of the EM algorithm.

4.2.6 Shadow Dirichlet distribution

Thus far, generalisations of the Dirichlet distributions have addressed inclusion of additional parameters in order to increase flexibility of the probability density function. The domain under consideration has remained the simplex Δ^D. However, this domain may not always be appropriate, if there is prior knowledge from the application perspective. For instance, if the multinomial likelihood is known to lie in a restricted subset of the simplex. Frigyik et al. (2010) recognise examples in language processing where this is often the case. An interesting example given by the authors that can be extended to further fields of application is modelling probability of words from a dictionary that are related or are synonyms. Frigyik et al. (2010) give an example of the words `espresso` and `latte`. Given that one of these words is observed in a list of words from a dictionary, the probability that the second word is observed is close to the probability of the first, with some difference ϵ. This relationship between the two words would be included in a specific bounded variation model, and is expressed as an ϵ-bound

in the probability mass function (pmf) of the bounded variation model, and this imposes a restriction on the domain of the mass function. Therefore, the pmf is restricted on the domain of the prior probability space, Δ^D in this instance. Frigyik et al. (2010) construct an equivalent of the Dirichlet distribution on a restricted simplex domain, and call this the Shadow Dirichlet distribution. One simple approach to restrict Δ^D to a subset $\widetilde{\Delta}^D$. An illustration of this could be

$\widetilde{\Delta}^3 = \{\boldsymbol{x} = (x_1 x_2, x_3) \in \mathrm{R}^3$ subject to the constraints $x_1 \geq 0.5, x_2 \leq 0.5, x_3 \leq 0.5$ and $\sum_1^3 x_j = 1$, j=1,2,3$\}$. Then, one could define the usual Dirichlet distribution over $\widetilde{\Delta}^D$ and re-normalise with respect to $\widetilde{\Delta}^D$. A disadvantage of this naive approach, as recognised by Nallapati, Minka and Robertson (2007) is that the re-normalisation term is not analytically tractable and requires numerical integration, which can become cumbersome with increasing D. Similarly, this re-normalised form of the Dirichlet does not have closed-form moments or covariance structure. Therefore, it is problematic outside of theoretical consideration of the restricted simplex support. To accommodate a subset of the simplex Frigyik et al. (2010) instead decompose the Dirichlet distribution into 2 parts: its generating Dirichlet distribution $\widetilde{\pi}(\boldsymbol{x})$ and a matrix M which is a continuous mapping from Δ^D to $\widetilde{\Delta}^D$. The matrix M is left stochastic, full rank and invertible. The probability density function for the Shadow Dirichlet distribution is as follows:

$$\pi_{sh}(\boldsymbol{x}; \boldsymbol{\alpha}, M) = M\widetilde{\pi}(\boldsymbol{x}; \boldsymbol{\alpha}). \tag{4.22}$$

hence

$$\pi_{sh}(\boldsymbol{x}; \boldsymbol{\alpha}, M) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^D \Gamma(\alpha_i) \det|M|} \prod_{i=1}^D (M^{-1}\boldsymbol{x})_i^{\alpha_i - 1}. \tag{4.23}$$

where $\alpha_0 = \sum_{i=1}^D \alpha_i$ are the usual Dirichlet precision terms and $(M^{-1}\boldsymbol{x})_i$ is the i^{th} element of the product of inversed matrix M and $\boldsymbol{x}$.

The Shadow Dirichlet is conjugate to the multinomial likelihood and its first moments and covariance structure have closed form:

$$\mathbb{E}(X_i) = \frac{M\alpha_i}{\alpha_0}, \tag{4.24}$$

$$\text{mode}(\mathbf{X}) = \frac{M(\alpha_i - 1)}{\alpha_0 - D}, \alpha_i > 1, \tag{4.25}$$

$$\text{Cov}(X_i, X_j) = M\Big(\frac{-\alpha_i\alpha_j}{\alpha_0^2(1+\alpha_0)}\Big)M^T. \tag{4.26}$$

where M^T is the transpose of matrix M and $\alpha_0 = \sum_{i=1}^{D} \alpha_i$.

The matrix M can, in theory, be estimated directly from the data, but is recognised by the authors as a non-convex problem (Frigyik et al., 2010). Two approximating solutions are offered - first, to use the standard Uniform distribution as the support in approximating M, placing an upper bound on the size of the convex hull of likelihood mass functions in the simplex. The second approximation suggested is to use the empirical mean of the likelihood mass functions and specify one column of M as a convex combination of the same vertex on the standard simplex Δ^D and the empirical mean pmf. On the other hand, substantial prior knowledge, if available for a specific application of the Shadow Dirichlet distribution, can be included in the specification of M. The authors provide some examples from machine learning that incorporate specific dependence structures and these are useful in defining M, for example, if points on the boundary of the full simplex Δ^D are of interest to the restricted simplex $\widetilde{\Delta}^D$. Frigyik et al. (2010) further discuss what can be defined as a restricted simplex, and the effect of any such constraints on M. In general they point out that M is required to be injective. Separate consideration is given to mapping the vertex points of Δ^D to $\widetilde{\Delta}^D$ using M, and also an explicit $\pi_{sh}(\boldsymbol{x})$ is defined if there needs to be a projection from Δ^D to $\widetilde{\Delta}^D$, which results in $\dim|\widetilde{\Delta}^D| < \dim|\Delta^D|$. This could prove useful when one wishes to express uncertainty about a subset of a compositional vector $\boldsymbol{x}$ on a restricted domain.

4.2.7 Inverted Dirichlet distribution and Dirichlet Type II distribution

In the univariate case if we wish to model the odds of success, as opposed to the probability of success, we can turn towards an extension of the Beta distribution. Due to Johnson and Kotz (1970) the following transformation holds:

$$\text{If } A \sim \text{Beta}(\alpha, \beta), \text{then } \frac{A}{1-A} \sim \text{Beta}^*(\alpha, \beta). \tag{4.27}$$

Where $\text{Beta}^*(\alpha, \beta)$ is known as the Beta prime distribution with probability density function

$$f(x) = \frac{x^{\alpha-1}(1+x)^{-\alpha-\beta}}{B(\alpha, \beta)}; \tag{4.28}$$

for $x \in [0, \infty); \alpha, \beta \in (0, \infty)$; and $B(\alpha, \beta)$ is the Beta function.

Tiao and Cuttman (1965) introduced a similar transformation for the multivariate case, and this concerns the original Dirichlet distribution. If $\boldsymbol{X} \sim \text{Dirichlet}(\alpha_1, ..., \alpha_D)$ then through the transformation $Y_i = \frac{X_i}{X_D}$ for $i = 1, ..., D-1$, then

$$\boldsymbol{Y} \sim \text{Dirichlet}_{ID}(\alpha_1, ..., \alpha_D). \tag{4.29}$$

Tiao and Cuttman (1965) state that it is also possible to define the Dirichlet Type II distribution as a ratio of chi-squared distributed random variables in a similar set-up as above. The probability density function of the Dirichlet Type II distribution is given by

$$\pi_{ID}(\boldsymbol{y}) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{D} \Gamma(\alpha_i)} \Big(\prod_{i=1}^{D-1} y_i^{\alpha_i - 1} \Big) (1 + y_1 + ..., + y_{D-1})^{-\alpha_0}. \tag{4.30}$$

where $y_i > 0$; y_i is a ratio of compositional terms as defined above and $\alpha_0 = \sum_{i=1}^{D} \alpha_i$. This distribution is not conjugate to the multinomial likelihood, but

instead to the negative multinomial distribution. The moments and covariance structure of the Inverted Dirichlet distribution are not trivial to compute and involve derivatives of products of gamma functions of dependent terms. The joint moment generating function for the Inveted Dirichlet distribution is provided in Tiao and Cuttman's original work. Although this distribution is a popular choice for Dirichlet mixture models (Bdiri and Bouguila, 2011) and applications in clustering algorithms (Bdiri et al., 2014) the Inverted Dirichlet distribution does not have straightforward implementation for the purposes of serving as a prior distribution in an expert elicitation procedure.

4.2.8 Other generalisations of the Dirichlet distribution

In this section we have explored generalisations of the Dirichlet distribution and applications to compositional data analysis since Aitchison's contributions from the 1980s onwards. Some other contributions to increase flexibility of the Dirichlet have aimed towards modelling observations that could be categorised as any of the possible compositional components - one example of this are missing responses or non-responses. For these purposes the Grouped Dirichlet (Ng et al., 2008) and Nested Dirichlet distribution (Ng et al., 2009) have been developed. Although advantageous over previous generalisations of the Dirichlet, especially when modelling heirarchical structures, these advancements do not have moments or covariance structure expressed in closed form, and any parameter estimation relies on costly E-M algorithms. One final interesting consideration given here to modelling uncertainty about compositional vectors is through the work of Tu (2016). The author challenges the constraint that the concentration parameters of the Dirichlet distribution must be strictly positive $\boldsymbol{\alpha} > 0$, and adjusts the probability density function in order to accommodate negative concentration terms and avoid a divergent normalisation factor, which usually occurs when $\boldsymbol{\alpha} \leq 0$. Tu (2016) instead introduces a lower bound for each component of $\boldsymbol{x}$, so that the probability density function now takes the form

$$\pi_{\mathrm{mD}}(\boldsymbol{x};\boldsymbol{\alpha},\epsilon) = \begin{cases} 0 & \text{if } \exists\epsilon, x_i < \epsilon; \\ \frac{\prod_{i=1}^{D} x_i^{\alpha_i - 1}}{Z(\boldsymbol{\alpha},\epsilon)} & \text{otherwise.} \end{cases} \tag{4.31}$$

where $Z(\boldsymbol{\alpha},\epsilon)$ is the normalisation factor, and an additional requirement imposed is that $\frac{1}{D} \geq \epsilon > 0$. The motivation for extension of the Dirichlet distribution to accommodate negative concentration parameters arises from the need to model sparsity about each variable in a composition. This can be achieved with the usual Dirichlet distribution up to an extent when $0 < \boldsymbol{\alpha} < 1$. As shown in Figure 4.1 for $0 < \boldsymbol{\alpha} < 1$ the mass of the distribution lies at the vertices of the 3-simplex, however, the author deems this to be insufficient in expressing strong sparsity such that in a Bayesian analysis the posterior is not easily dominated by the likelihood when the prior is expressed as in the figures above. Tu (2016) sees this as an especially important aspect when model size is considerably smaller than the training data set, and the modeller wishes to portray a balance between prior information in contrast to a large body of evidence. As seen previously in this section, generalisations of the Dirichlet distribution are often subject to moments and covariance structures that cannot be expressed in closed form, and this modified sparse variant of the Dirichlet is no exception. Although π_{mD} is conjugate to the multinomial likelihood model. Still, the modified sparse Dirichlet distribution is an interesting theoretical contrast to the other Dirichlet variants presented in this section, as an attempt to challenge underlying constraints imposed by construction the distribution function.

4.3 Gaussian distribution

In this section we explore applicability of the Normal family of distributions to compositional data sets. The classical Gaussian distribution is not wholly appropriate for compositional parts, as the parts cannot take values over the entire $\mathbb{R}$, however, the distribution may be applied post log-ratio transformation on compositional data, as proposed by Aitchison. A remark made by Kieschnick and McCullough (2003) was that the Normal distribution was used by researchers when addressing the conditional distribution of a two-part composition, provided that this model included a set of predictor variables. Outside of this specific case, the use of the Gaussian distribution for compositional data is limited.

4.3.1 Truncated Gausssian distribution

In the setting of compositional data analysis, the Normal distribution can be truncated at the endpoints $[0, 1]$ (Dobigeon and Tourneret, 2007) in an attempt to restrict the probability density function to a support smaller than $[-\infty, \infty]$. The probability density function for the multivariate truncated Gaussian distribution in the general interval $[\boldsymbol{a}, \boldsymbol{b}]$ is expressed as follows:

$$\pi_{TN}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{a}, \boldsymbol{b}) = \frac{\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}}{\int_{\boldsymbol{a}}^{\boldsymbol{b}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}d\boldsymbol{x}}. \tag{4.32}$$

where $\boldsymbol{a}, \boldsymbol{x}, \boldsymbol{b} \in \mathbb{R}$ and $\boldsymbol{a} < \boldsymbol{x} < \boldsymbol{b}$. The above probability density function portrays a double truncation on the Normal distribution, as specified by the inequality relationship $\boldsymbol{a} < \boldsymbol{x} < \boldsymbol{b}$. The finite integral in the denominator term is the D-dimensional normalisation constant, where $D = |\boldsymbol{a}| = |\boldsymbol{x}| = |\boldsymbol{b}|$ the lengths of the respective vectors.

There also exist one-sided (single) truncations, as developed by Tallis (1961) and are defined by specifying an upper bound or a lower bound on the domain of $\boldsymbol{x}$, for example, $\boldsymbol{x} < \boldsymbol{b}$. Further works by Tallis (1963, 1965) give extensions to

truncations through linear combinations of constraints and through the use of planes. For the purposes of compositional data we retain a two-sided truncation on the interval $[0, 1]$. As with the usual multivariate Normal distribution, parameter estimation needs to be carried out for $\boldsymbol{\mu}$ and Σ. When the truncation points are known, frequentist methods to give estimates of $\boldsymbol{\mu}$ and Σ rely on Maximum-Likelihood or the method of Instrumental variables (Lee, 1979; Amemiya, 1973). A highlight of an application of the Truncated Normal distribution in modelling uncertainty about a set of proportions is through the work of Ezbakhe and Pérez Foguet (2019). Modelling strategy consisted of utilising the generalised additive model structure to include addition of non-linear smooth terms. Albeit the application is univariate, as each part of the composition is considered in separation from the others, it was found that the Truncated Normal distribution provides superior modelling of uncertainty, as contrasted with the extended Beta distribution.

Some other general properties of the truncated Normal concern the form of the marginal distributions, which are themselves not truncated Normal. This is contrasted with the unbounded multivariate Normal case, where marginal distributions are indeed also Normal. Cartinhour (1990) gives an analytic form of the the marginal distribution:

$$\pi_d(x_d) = \frac{\exp(\frac{-(x_d-\mu_d)^2}{2\sigma_{dd}})}{p\sqrt{2\pi\sigma_{dd}}} \int_{b_{d-1}}^{a_{d-1}} ... \int_{b_1}^{a_1} \frac{\exp(\frac{-(x_1-m(x_d))^T\Sigma_1^{-1}(x_1-m(x_d))}{2})}{\sqrt{(2\pi)^{d-1}|\Sigma_1|}} dx_1...dx_{d-1}; \tag{4.33}$$

for $b_d \leq x_d \leq a_d$, $d = 1, ..., D$ and the density function is zero otherwise.

In the above, p is the normalisation term, $m(x_d) = \boldsymbol{\mu} + \frac{c(x_d - \mu_d)}{\sigma_{dd}}$, and the covariance matrix Σ can be partitioned as follows:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_1 & c \\ c^T & \sigma_{dd} \end{pmatrix}.$$

Cartinhour (1990) expresses that the joint marginal distribution function for the truncated Gaussian distribution is a product of a truncated Gaussian and a so-called skewness function that adjusts the shape of $N(\mu_d, \sigma_{dd})$.

Example

An illustration of the Truncated Gaussian distribution on the simplex could be the following scenario: the lower limits are set to 0 and upper limits to 1; the mean vector $\boldsymbol{\mu} = (1/2, 1/2, 1/2)$ and the covariance matrix generated from realisations of the uniform distribution, such that

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.89 & -0.73 & -0.48 \\ -0.73 & 1.97 & 1.41 \\ -0.48 & 1.41 & 1.02 \end{pmatrix}$$

This defines a 3-dimensional truncated Gaussian variable, and we can represent realisations from it on the simplex, as below.

Due to the choice of $\boldsymbol{\mu}$ the density is concentrated at the centre of the simplex, however, contrasted with earlier Dirichlet realisations we can see that this density has less regular shape and is stretched out towards the edges of X_2 and X_3. We can also see some local central contours closer to the edges, where the mass becomes more concentrated.

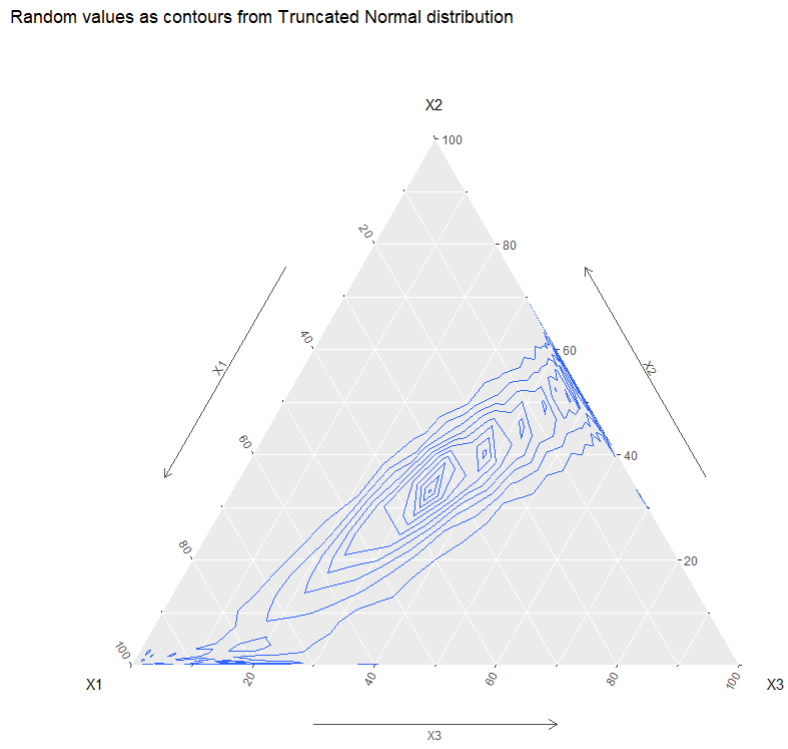


Figure 4.2: Ternary plot of contours of the Truncated Gaussian distribution

4.3.2 Logistic skew-Normal distribution

The lognormal distribution is the distribution of a positive-valued random variable whose log-transformation follows the Gaussian distribution. This result is first due to McAlister (1879) and popularity of this distribution increased in the 20$^{\text{th}}$ century with developments in the area of analysis of variance. Still, in a Bayesian context the lognormal distribution was used as a conjugate to the multinomial likelihood as early as Lindley (1964). The link to compositional data analysis was established by Aitchison and Shen (1980) through the log-ratio transformation approaches as explored in Chapter 2. The additive logistic skew-Normal distribution acts with respect to additive log-ratio (alr) transformation, as defined in Chapter 2, assuming that the alr-transformed components can follow a $(D-1)$ variate Normal distribution. This is referred to as the logistic Normal model for logratio-transformed compositional parts. This is a suitable model in many instances, but may fall short when transformed compositional data displays skewness. To account for this, the logistic skew-Normal distribution relies on the

existing multivariate skew-Normal distribution (Azzalini and Valle, 1996) for a $D-$variate vector of values. Additional to the classical Gaussian distribution, the skew-Normal contains an extra parameter ζ to control the shape of the distribution and determine its maximum skewness. Mateu-Figueras et al. (2005) then augment the skew-Normal distribution to act on the set of alr-transformed compositional variables. Similar adjustments are proposed with the isometric logratio transformation, and it is noted that centred logratio parameterisation will only yield a constant (degenerate) distribution due to its dimensionality. However, in applying the alr and ilr transformations, the key idea is that the skew-Normal density is transformed using the inverse of each of the logratio transformations. In the alr case, the new distribution has $(D+4)(D-1)/2$ parameters and the moments of the additive logistic skew-Normal distribution are not analytically tractable. The use of these distributions rely heavily on the assumption of skew-Normality, and known tests such as Anderson-Darling or Cramer-von-Mises can be used to assess the differences between the empirical and hypothesised distribution functions. While this distribution is attractive in providing alternative modelling techniques for compositional data that is significantly skewed post-transformation, from the perspective of Bayesian prior elicitation the facilitator may find difficulties in formulating questions about these types of distributions in a way accessible to the experts. Similarly, there are $(D+4)(D-1)/2$ parameters required for the specification of the additive logistic skew-Normal distribution, and this may prove a lengthy and challenging exercise.

4.4 Liouville family of distributions

As we explored in Section 4.2 of this chapter, the Dirichlet family of distributions can have restrictive dependence structure. It would be particularly challenging to model compositional parts that display positive correlation. Aitchison attempted to rid the statistical community of these restraints through the use of the log-ratio

transformations and the Normal family. This brings about a reduction in dimensionality of the problem, from D to $D-1$. A disadvantage to this approach is that now modelling independence between compositional parts Aitchison (1986) is a challenge also. Rayens and Srinivasan (1994) aim to resolve this dilemma by remaining in the simplex space and modelling nontrivial dependence structures through the use of the Liouville family of distributions. The generalised Liouville distribution contains the Dirichlet class of distributions, and finds practical use in copula distributions, which are defined later in this chapter. The set-up for the Liouville distribution on the simplex Δ^D and compositional vector $\boldsymbol{x}$ are as follows:

The vector $\boldsymbol{x}_{D-1}$ is defined to lie in an irregular right-angle simplex

$\Delta^* = \{(x_1, ..., x_{D-1}) \in \mathbb{R}^{D-1} : \sum_{i=1}^{D-1} x_i \leq 1, x_i \geq 0 \ \ \forall i\}$. Then, a kernel function $u : {\mathbb{R}_+}^{D-1} \to {\mathbb{R}_+}^1$ such that

$u(x_1, ..., x_{D-1}) = f\{(x_1/q_1)^{\beta_1} + ... + (x_{D-1}/q_{D-1})^{\beta_{D-1}})\}$ and $f(\cdot)$ is a continuous function $f : {\mathbb{R}_+}^1 \to {\mathbb{R}_+}^1$.

Also $\beta_i, q_i > 0, \forall i$ values that need to be specified.

Then, the generalised family of Liouville distributions on Δ^* is defined if the probability density of $\boldsymbol{x}_{D-1}$ is

$\pi_{LV}(\boldsymbol{x}; \boldsymbol{\alpha}) = A \cdot u(x_1, ..., x_{D-1}) \cdot x_1^{\alpha_1 - 1} \cdot ... \cdot x_{D-1}^{\alpha_{D-1} - 1}$ for $(x_1, ..., x_{D-1}) \in \Delta^*$ and 0 otherwise. A is a normalisation term.

The usual Dirichlet class is given in the above if we set $q_i = \beta_i = 1, \forall i$; and the function $f(\cdot)$ is taken as $f(\xi) = (1-\xi)^{\alpha_D - 1}$.

Rayens and Srinivasan (1994) go further to derive the covariance structure of the generalised Liouville distributions, with references to the Dirichlet case. The authors find that for the cases $q_i = \beta_i = 1$ expected values and covariance structures can be expressed in relatively compact form of a one-dimensional integral of $f(\xi)$. In the case, however, where $q_i = \beta_i \neq 1$ the authors recognise a need

for Monte Carlo integration. The basis for generating random Liouville vectors is essential for the Monte Carlo integration, and the starting point is generation of Dirichlet random variables - this once again highlights that the Liouville class is a generalisation of the Dirichlet. Again, this can be a computationally-heavy task if is no problem-specific or natural choice for $f(\cdot)$, and so the procedure of specifying the richer covariance structure is weighed up against choosing the appropriate $f(\cdot)$ - a task that can be paralleled with model selection in a statistical analysis.

Applied techniques modelling compositional data with the Liouville functions have been sparse, understandably so due to the necessary selection of $f(\cdot)$ and similar calibration of q_i and β_i values. In the remaining sections of this chapter we will meet copula functions, some of which do rely on the described Liouville approach to specify a more exotic dependence structure than ones readily offered by variants of Dirichlet or through Aitchison's log-ratio transformation.

4.5 Distributions on a sphere

The square-root transformation of compositional D-part vector $\boldsymbol{x}$ presented in Chapter 2 gives rise to data that can be modelled on a multi-dimensional sphere of dimension $D-1$: S^{D-1} (Scealy and Welsh, 2011). When $\sqrt{\boldsymbol{x}}$ lies on the positive orthant of S^{D-1} and is not on the boundary of the orthant, it can be modelled using the Kent distribution (Kent, 1982). Equally known as the Fisher–Bingham distribution, it is a five-parameter distribution function.

The distribution function can be described as follows: for $y_i = \sqrt{x_i}$ a point on the unit sphere S^{D-1} the density function of the D-dimensional Kent distribution is

$$\pi_{\text{Kent}}(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \kappa) \propto \exp\{\kappa \gamma_1^T \cdot \boldsymbol{y} + \sum_{i=2}^{D} \beta_i (\boldsymbol{\gamma_i}^T \cdot \boldsymbol{y})^{\mathbf{2}}\}. \tag{4.34}$$

where $\sum_{i=2}^{D} \beta_i = 0$ and $0 \leq 2|\beta_i| < \kappa$ and $\boldsymbol{\gamma_i}$ orthonormal vector for $i = 1, ..., D$.

In higher dimensions $D - 1 > 3$ the normalisation constant of the Kent distribution on the hypersphere is non-trivial to compute, so finds limited use in compositional data analysis. Scealy and Welsh (2011) uses the Kent distribution in the task of compositional regression, as it is deemed suitable for handling essential zero values in a compositional data set. Parameters of the Kent distribution find similar interpretation to those in a Dirichlet distribution: $\kappa > 0$ is a parameter responsible for concentration of the distribution, like $\alpha > 0$ in the Dirichlet family. β is a parameter that governs ellipticity of the contours on the surface of the hypersphere, and $\boldsymbol{\gamma_i}$ are the orthogonal direction vectors. Scealy and Welsh (2011) exploit the idea of the direction vectors in a regression framework through a mapping to linear functions of the predictor variables. Even though the Kent distribution has easier interpretability of parameters than the classical Gaussian, for example, Scealy and Welsh (2011) find that square root-transformed compositional data often lie on the boundary of the orthant of S^{D-1}. A resolution to this is to adjust the underlying spherical distribution to allow for folding (Scealy and Welsh, 2014), however, these approaches to modelling compositional data have not been popular in more applied areas, most probably due to the existence of methods with easier implementation.

4.6 Uninformative distributions

In preceding sections we explored distribution families about a set of proportions $\boldsymbol{x}$ summing to unity. Each distribution considered as a possible choice to model uncertainty about $\boldsymbol{x}$ implies making that choice in favour of one over the others, and deeming it appropriate to the scientific question. In this section we outline probability distributions on the simplex space that would reflect a maximal lack of subjective knowledge about the distribution for the probability parameter in a multinomial likelihood. We therefore seek to explore uninformative prior distributions on the simplex, and start with the Jeffreys prior, which was the

first formal and reproducible way of constructing a diffuse (uninformative) prior distribution.

Remaining in the same set-up, such that the $\boldsymbol{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D$ subject to the constraints $x_j \geq 0$ and $\sum_1^D x_j = 1$ with each x_j compositional component, and in the instance of the multinomial distribution $\boldsymbol{x}$ takes on the role of the event probabilities. Recall, the probability mass function for the multinomial model is

$$f(\boldsymbol{y}|\boldsymbol{x}) = \frac{N!}{y_1!...y_D!} x_1^{y_1}...x_D^{y_D}.$$

with $\sum_{i=1}^D y_i = N; y_i \in \mathbb{Z}_+^D$ and $i = 1, ..., D$.

Then the Jeffreys prior for the multinomial distribution with respect to the parameter $\boldsymbol{x}$ is

$$\pi_J(\boldsymbol{x}) \propto \sqrt{I(\boldsymbol{x})} = \sqrt{\mathbb{E}\Big[\Big(\frac{d}{d\boldsymbol{x}} \log f(\boldsymbol{y}|\boldsymbol{x})\Big)^2\Big]}. \tag{4.35}$$

From the above, the information matrix $I(\boldsymbol{x})$ is diagonal with entry values being $\frac{\mathbb{E}(Y_i)}{x_i^2} = n/x_i$ for $i = 1, ..., D$. Thus, the Jeffreys prior for $\boldsymbol{x}$ follows the Dirichlet distribution with all concentration parameters equal to $1/2$, and this is a proper prior distribution.

Zellner (1977, 1996); Zellner and Min (1992) built upon the idea of a structured way to produce a diffuse prior distribution. The Maximal Data Information Prior Density (MDIPD) due to Zellner are derived through maximising the difference between the average information in the data and the information in the prior density (Zellner, 1996), and the MDIP prior emphasises the information in the likelihood function, in this case the multinomial. For the multinomial likelihood with the same set-up as above, derivation of the MDIPD involves working with the following negative entropy of the multinomial pmf:

$$I_D(x_1, ..., x_{D-1}) = x_1 \ln(x_1) + x_2 \ln(x_2) + ... + (1 - \sum_{i=1}^{D-1} x_i) \ln(1 - \sum_{i=1}^{D-1} x_i). \tag{4.36}$$

Then it is necessary to maximise

$$\int ... \int I_D(x_1, ..., x_{D-1})\pi(x_1, ..., x_{D-1})dx_1...dx_{D-1} - \int ... \int \pi(x_1, ..., x_{D-1})\ln\pi(x_1, ..., x_{D-1})dx_1...dx_{D-1}. \tag{4.37}$$

The resulting MDIPD is a proper distribution function taking the form

$$\pi_{MDIP}(x_1, ..., x_{D-1}) \propto x_1^{x_1} x_2^{x_2} ... x_{D-1}^{x_{D-1}} \Big(1 - \sum_{i=1}^{D-1} x_i\Big)^{1-\sum_{i=1}^{D-1} x_i}. \tag{4.38}$$

4.7 Uniformity over the simplex

In this short section we explore how uniform prior distributions can be represented on the simplex, and any change that occurs from transitions between the Euclidean space.

For ease of representation on a ternary plot let us consider the case where dimension $D = 3$ and generate a random uniform sequence $U[0, 1]$ and then normalise it by dividing each element by the row sum. The ternary plot in Figure 4.3 depicts clear lack of uniformity when the points are projected onto the simplex.

To overcome the centre-clustering of points, we can follow Rubin (1981) and in the first step draw $D - 1$ points from the standard uniform distribution, then order them by size and further add to the list values of 0 and 1 to obtain $\{0, u_1, u_2, ..., u_D, 1\}$. Then take differences between consecutive numbers in the list. This procedure drives the results in Figure 4.4, and we can be more convinced that uniformity over the simplex is satisfied in this instance, at least graphically. In both cases 500 data points were generated.

We can see that mere normalisation of uniformly distributed variables does not imply that the spread of points on the simplex space is also even. The ordered difference method allowed us to correct central clustering of points, as seen in

Figures 4.3 and 4.4.

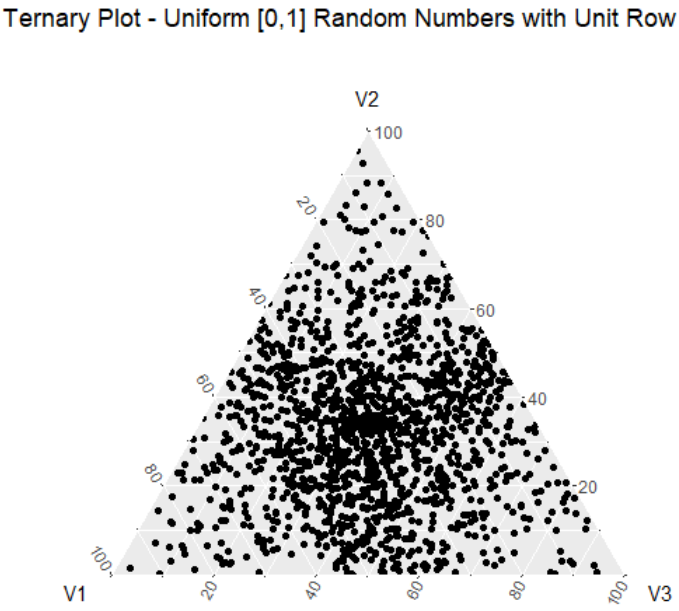


Figure 4.3: Uniformly generated elements, normalised.

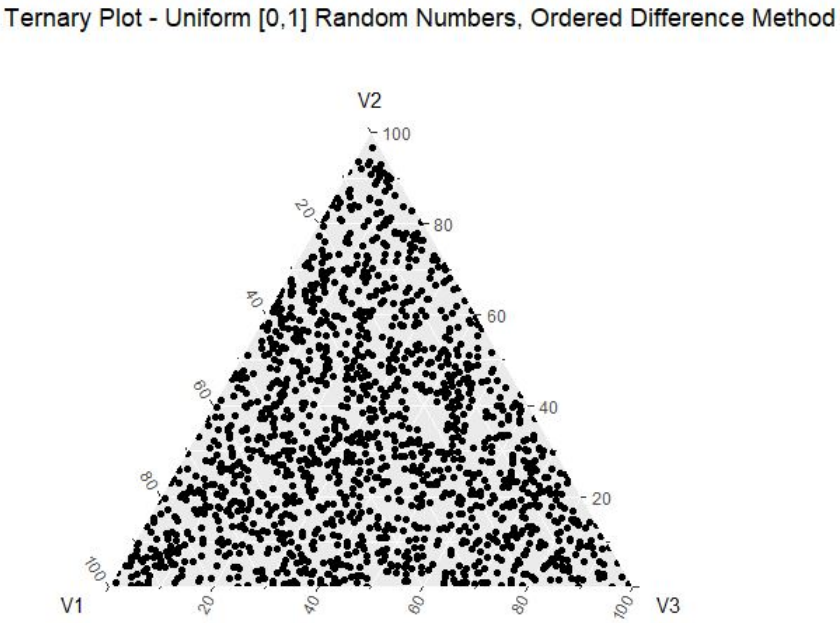


Figure 4.4: Uniformly generated elements, ordered difference method.

4.8 Other considerations

In this section, we give consideration to other mathematical constructions that can express uncertainty about compositional data, but fall outside the topic realms mentioned previously in this chapter. In describing multivariate distributions on the simplex we have settled on accepting the covariance structure as governed by the distributional form. For example, for the Dirichlet in Section 4.2.1 we saw that all parts of a composition have a common variance parameter (since this depends on the overall concentration parameter α_0), yet each variable has its own expected value. Similarly, the variables are deemed mutually independent, given that the sum-to-unity constrained is considered which drives a negative correlation (Mosimann, 1962). Hence, the covariance between at least two distinct compositional parts is negative, should it be assumed they come from a Dirichlet distribution. In practical applications this may not always be representative of any natural processes that influence the "true" data-generating process.

With the Connor Mosimann distribution and other flavours of the Dirichlet we explored how these can accommodate for positive correlation between compositional parts, and even the scenario where more than one mode can occur in compositional vectors. Unfortunately, many of these distributions are costly to sample from, and do not have analytically-expressed moment functions.

4.8.1 Dirichlet-tree distribution

Dennis III (1991) and Minka (1999) provide a different outlook on the above problems by introducing a tree-structure approach to modelling dependence between compositional parts through the Dirichlet-tree distribution. The difference with previous methods lies in the parametrisation approach adopted by Dennis III (1991) - the vector $\boldsymbol{p}$ of event probabilities in the multinomial sample is now regarded as a finite stochastic process, as illustrated in Figure 4.5.

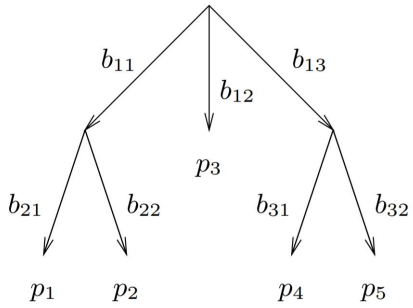


Figure 4.5: General tree structure for finite stochastic process (Dennis III, 1991; Minka, 1999).

Under this parametrisation, the tree above consists of nodes indexed by i, branches indexed by j, branch probabilities represented by b_{ij} and leaf probabilities p_d, for $i = \{1, 2, 3\}, j = \{1, 2, 3\}$ and $d = \{1, 2, 3, 4, 5\}$ for this example. The probability of a leaf is the product of probabilities of the branches that make a path to that leaf. For instance, $p_1 = b_{11}b_{21}$ and $p_3 = b_{12}$.

Dennis III (1991) then uses the tree structure T as a basis for the Dirichlet-tree distribution, with an explicit probability density function given by:

$$\pi(\boldsymbol{p}; \alpha, T) = \prod_d p_d^{\alpha_{\text{parent}_d}-1} \prod_i \frac{\Gamma(\sum_j \alpha_{ij})}{\prod_j \Gamma(\alpha_{ij})} \Big(\sum_{dj} \delta_{ij}(d) p_d \Big)^{\beta_i}. \tag{4.39}$$

where α_{parent_i} is the usual Dirichlet concentration parameter α for the branch that leads up to node i. $\beta_i = \alpha_{\text{parent}_i} - \sum_j \alpha_{ij}$ if i is not the root node, while $\beta_i = 0$ if i is the root node. Finally, $\delta_{ij}(d) = 1$ if branch ij leads to d and $\delta_{ij}(d) = 0$ otherwise.

Through this parametrisation we can see that at each node the Dirichlet-tree distribution is assigned a separate concentration parameter α_{ij} and this means that the variance at each node can be different across the nodes. Hence, for each p_d the variance is independent of variances of the remaining leaf probabilities. The Dirichlet tree structure also allows us to model dependent subcompositions, since by the above set-up the leaves in a sub-tree are correlated, because they depend on the shared branches of that sub-tree. The Dirichlet-tree distribution

can be reduced to the classical Dirichlet distribution (tree of depth 1) by setting $b_i = 0, \forall i$. Minka (1999) also notes that the same effect can be achieved through a specific choice of α values, if one doesn't want to disrupt an existing tree structure. The Dirichlet-tree distribution also holds the convenient conjugate property with the multinomial likelihood and has moments that are expressed analytically (Dennis III, 1991). Indeed, this reparametrisation has found uses in compositional regression and clustering problems, for example Mao and Ma (2020). Another interesting application has been by Liu et al. (2014) through the use of the Dirichlet tree as a basis for random forest classification in the task of facial recognition. The tree-structure approach has not been yet reported in the exercise of expert elicitation, although similar structures have been explored, as we see through the use of copula vines in due course.

4.8.2 Copulae functions

In similar regard, the notion of copulae is deemed useful for describing dependence between random variables. A copula function (copula) is a special type of a general multivariate distribution on the hypercube, which can be constructed to accommodate different desired relationships between random variables. A copula represents a multivariate cumulative density function (CDF) and this CDF can be decomposed into one-dimensional marginal CDFs, which can be written separately from a dependence structure for the multivariate CDF.

Definition 4.3. *(Copula) A multivariate copula is the joint distribution of several random variables X_1,...,X_n, with each X_i following some marginal distribution. A common choice for such can be the Uniform(0,1).*

A fundamental idea in the theory of copulae is Sklar's Theorem (Sklar, 1973), which allows us to describe the joint distribution of $X_1, ..., X_n$ by their marginal distributions and a copula, C:

Theorem 4.1. *(Sklar's Theorem) For random variables $X_1, ..., X_n$ with cumu-*

lative distribution functions (CDF) $F(x_1, ..., x_n) = \mathbb{P}(X_1 \leq x_1, ..., X_n \leq x_n)$ *and marginal CDF* $F_i(x) = \mathbb{P}(X_i \leq x)$ *for* $i = 1, ..., n$*, there exists a copula* $C : [0, 1]^n \to [0, 1]$*, such that* $F(x_1, ..., x_n) = C[F_1(x_1), ..., F_n(x_n)]$*.* C *is unique if each* $F_i(x)$ *is continuous.*

There exist several classes of copulae, for modelling dependence under different conditions and aims sought. Nelsen (2007) gives an introduction to the topic of copulae. However, for the purposes of modelling compositional data the Gaussian copula is a popular choice for $F()$ as defined above. Similarly, if one wishes to model uncertainty about extreme events, one is interested in the tails of a distribution function. An appropriate copula for this instance would be the Gumbel or the Fréchet. Generally, if two random variables Y and Z have well-defined marginal distribution functions, such that $Y \sim f_1$ and $Z \sim f_2$ for probability densities f_1 and f_2 and cumulative densities F_1 and F_2, then a copula $C(Y, Z)$ is the distribution of $(F_1(Y), F_2(Z))$.

For $Y \perp\!\!\!\perp Z$ we have $C(Y, Z) = \prod yz = yz$ and the density function

$c(y, z) = \frac{\delta}{\delta y}\frac{\delta}{\delta z}C(Y, Z) = \frac{\delta}{\delta y}\frac{\delta}{\delta z}yz = 1$ on $[0, 1]^2$, for two random variables Y and Z.

Similarly, for a vector $\boldsymbol{X} = (X_1, ..., X_D)$ a Gaussian copula defined at a point $(x_1, ..., x_D)$ is $C[F_1(x_1), ..., F_D(x_D)] = \Phi_{D,\boldsymbol{R}}\Big\{\Phi^{-1}(F_1(x_1)), ..., \Phi^{-1}(F_D(x_D))\Big\}$, where $\Phi_{D,\boldsymbol{R}}$ is the cumulative density function of a D-variate Gaussian distribution with correlation structure $\boldsymbol{R}$. $\Phi_{D,\boldsymbol{R}}$ also has zero-mean and unit variance for each $(F_1(x_1), ..., F_D(x_D))$. $\boldsymbol{R}$ reflects a chosen dependence structure. Φ^{-1} is the inverse of the (marginal) univariate standard Gaussian CDF, and as before, $F_i(x_i), i = \{1, ..., D\}$ are chosen marginal cumulative density functions. Clemen and Reilly (1999) go on to differentiate

$C[F_1(x_1), ..., F_D(x_D)]$, $\frac{\delta}{\delta x_1}...\frac{\delta}{\delta x_D}C[F_1(x_1), ..., F_D(x_D)] = c[x_1, ..., x_D]$ as above, to obtain:

$$
\begin{aligned}
c[x_1, ..., x_D] &= f(x_1, ..., x_D; \boldsymbol{R}) \\
&= \frac{\prod_{j=1}^{D} f_j(x_j)}{|\boldsymbol{R}|^{1/2}} \exp\Big\{ -\frac{1}{2}[\Phi^{-1}(F_1(x_1)), ..., \Phi^{-1}(F_D(x_D))]^T (\boldsymbol{R^{-1}} - \mathrm{Id}_D) \\
&[\Phi^{-1}(F_1(x_1)), ..., \Phi^{-1}(F_D(x_D))]\Big\}.
\end{aligned}
\tag{4.40}
$$

For Id_D identity matrix of order D, $f_j(\cdot)$ being the density function of $F_i(\cdot)$ and $i = \{1, ..., D\}$.

For the multinomial likelihood, the copula function could serve as a prior distribution, and a starting point could be the specification of $f_i(x_i)$ which could be the Beta distribution for $x_1, ..., x_D > 0$ and $\sum_{i=1}^{D} x_i = 1$. However, to satisfy compositional constraints, namely that the means of the marginal distributions sum to unity, and that there are also no redundant variables that can lead to a singular Gaussian distribution (Elfadaly and Garthwaite, 2017), it is necessary to reparameterise x_i. The following parametrisation is suggested:

$Z_1 = x_1$, for the first component. The last component is $Z_D = 1$ and the remaining components in-between are $Z_i = \frac{x_i}{1 - \sum_{j=1}^{i-1} x_j}, i = 2, ..., D-1$.

Then Z_i is assigned a marginal Beta distribution, which now restrict Z_i to $[0, 1]$. A Gaussian copula density for $(Z_1, ..., Z_D)$ dependence structure is defined as in Equation 4.41. Connor and Mosimann (1969) reduced the copula to the Generalised Dirichlet distribution in the instance of $Z_1, ..., Z_{D-1}$ being independent.

Vine copulae

Copulae can be combined into a tree-like structure, which builds on the idea of the Dirichlet-tree distribution. In the example above $C(Y, Z)$ is a copula of two random variables Y and Z - such copulae are known as bivariate. A vine copula rests on the idea of decomposing a multivariate distribution function into a (finite) set of bivariate copulae (Kurowicka and Joe, 2010). A tree-like structure is then obtained and the decomposition of the multivariate distribution can be formed from -

1) The marginal (prior) distributions, such as $f_i(x_i)$ above;

2) Unconditional copulae $c_{i,i+1}$;

3) Conditional copulae $c_{i,i+j|(i+1,\ldots,i+j-1)}$.

As an illustration, for three random variables X_1, X_2, X_3, the following decomposition into bivariate copulae is possible:

$$\begin{aligned} f(x_1, x_2, x_3) = f(x_1) \cdot f(x_2) \cdot f(x_3) \cdot c_{13}(F_1(x_1), F_3(x_3)) \\ \cdot c_{23}(F_2(x_2), F_3(x_3)) \cdot c_{12|3}(F_{1|3}(x_1|x_3), F_{2|3}(x_2|x_3)); \end{aligned} \tag{4.41}$$

with $f_i(x_i), F_i(x_i)$ and $c(\cdot)$ defined as above.

Bedford and Cooke (2002) recognised the connection between vines and graph structure. In the example above, the vine can be represented as:

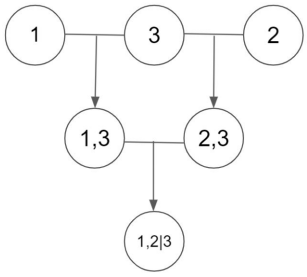


Figure 4.6: Vine copula structure for three random variables.

The tree structure in Figure 4.6 has three layers, the top level governing independent marginal distributions, the second layer - the unconditional copulae, and the final bottom layer - the conditional copula function. Including more variables would drive a tree structure with more layers. However, each layer must adhere to conditions that the first layer has D nodes with $D-1$ edges (branches) and all nodes must be connected; and secondly, every edge contributes to the joint density in the next layer down. Hence, each layer in a vine copula adheres to a certain structure. Generally, every vine copula has a Regular Vine (R-vine) structure, as in example above. The R-vine can be decomposed into two further classes: when a layer contains a central component with a star-like structure with

edges to the other nodes in the layer, this is known as Canonical Vine (C-vine). Similarly, another class of R-vines is the Drawable Vine (D-vine) and here every node is connected to at most two edge in a layer, i.e. each tree has a path.

There is a great body of literature about copula functions in general, and the specific vine structures as briefly outlined in this section. Applications of copulae functions are similarly abundant, from financial modelling (Rodriguez, 2007) to risk assessment (Bedford et al., 2016). Ortego and Egozcue (2013) discuss applications of copulae to model dependence relationships within a compositional data set. They also contrast association measures such as Spearman's correlation ρ and Kendall's τ rank correlation with the spurious Pearson correlation. It was found that subcompositional coherence criterion is not met by the Spearman's, Kendall's and the copula approach, and that the dependence structure modelled through copulae (seven common types were selected for the exercise) is also spurious. For the purposes of this thesis, copula methods have found use in expert elicitation of multivariate probability densities over the simplex space, serving to challenge existing approaches that rely on the restrictive Dirichlet distribution. These modern approaches are explored in more detail in Chapter 7.

Chapter 5

Uncertainty modelling of Markov chains

In this chapter, we explore parallels between compositional data and transition matrices stemming from a discrete-time Markov chain. After introducing discrete-time Markov models, we look towards modelling uncertainty about the row elements of the transition matrix. Later, we consider long-term behaviour of the chain's row-wise uncertainty through the stationary distribution. We also present current insights into doubly-constrained matrices and confusion matrices with examples.

5.1 Discrete-time Markov chains

A stochastic process $\{X_t, t \in T\}$ is a collection of random variables X_t, indexed by a set T. This process is defined on the sample space Ω and equipped with a σ-algebra F and a base probability measure P. When T is a set of times, the stochastic process is known as a temporal stochastic process, which is our interest here. Conversely, for example if T is the set of spatial coordinates, then the process is called a spatial process. Let us consider the first scenario where a point $t \in T$ is a single time point. Hence, X_t is a random variable that depicts the

value observed at time t. This can be equally represented by $\{X_t, t = 0, 1, 2, 3...\}$. The discrete-time stochastic process takes values in a measurable state space S that is also equipped with some appropriate σ-algebra. The values taken by X_t are known as states of the process $x_t \in S$, and they may be discrete values or continuous values. An essential idea in the construction of Markov chains is the Markov Property:

Definition 5.1. *Let $X(t)$ be a stochastic process in discrete time T and discrete state space S. Then X_t has the Markov Property if $\forall t \in T$ and states $x_0, x_1, ..., x_t, x_{t+1} \in S$ the following holds: $\mathbb{P}(X_{t+1} = x_{t+1}|X_t = x_t, X_{t-1} = x_{t-1}, ..., X_0 = x_0) = \mathbb{P}(X_{t+1} = x_{t+1}|X_t = x_t)$*

The Markov Property leads us to consider the transition probability in a discrete-time Markov process. The transition probability conditions on the current state of the process and gives the probability of the random variable at the next time point:

$p_{ij} = \mathbb{P}(X_{t+1} = j|X_t = i)$ where $i, j \in S$. In the situation where these transition probabilities stay constant as time progresses, this type of discrete-time Markov process is called time-homogeneous. Such a time-homogeneous Markov process is known as a Markov chain. Formally, a Markov chain is defined by its transition probabilities and the initial distribution of its states:

Definition 5.2. *The initial distribution of a Markov chain is a probability distribution (π_0) on a sample space S with $\pi_0(i) = \mathbb{P}(X_0 = i)$, such that the Markov chain starts in state i, $\forall i \in S$. Also, $\sum_{i \in S} \pi_0(i) = 1$.*

Generally, we can denote the distribution of the chain at time t by $\pi_t(i) = \mathbb{P}(X_t = i)$. For a finite state space S, we can organise the transition probabilities into a square $N \times N$ transition matrix $\mathbf{P}$ where each $(i, j)^{\text{th}}$ element of this matrix is given by $p_{i,j}$. Since $\sum_j p_{ij} = 1$ the sum of each row of the square matrix P now sums to unity and such matrices are known as right-stochastic. In the situation where P is symmetric, its columns will also sum to unity, allowing us to call it

it left-stochastic. A matrix that has each row and each column adding to unity is known as doubly stochastic. An elementary example of a doubly stochastic matrix is a matrix with strict diagonal elements equal to one.

Definition 5.3. *A square matrix* $\mathbf{P} = (p_{ij})$ *is doubly stochastic if* $\sum_i p_{ij} = \sum_j p_{ij} = 1$, *where* $p_{ij} \geq 0$.

A less trivial definition of a non-square doubly stochastic matrix is provided by Caron et al. (1996):

Definition 5.4. *An* $N \times M$ *matrix* $\mathbf{P} = (p_{ij})$ *is doubly stochastic with uniform marginals of size* $N \times M$ *if* $\sum_{j=1}^{M} p_{ij} = M$ *for* $i = 1, 2, ..., N$ *and* $\sum_{i=1}^{N} p_{ij} = M$ *for* $j = 1, 2, ..., M$, *where* $p_{ij} \geq 0$ *for* $i = 1, 2, ..., N$ *and* $j = 1, 2, ..., M$.

Applications of doubly stochastic matrices are sparse in statistics, let alone in the realm of Bayesian analysis. Some consideration has been given by Huang et al. (2014) in the analysis of cancer transition rates. Nevertheless, these types of matrices make for interesting objects in the study of graphs (Wang et al., 2016) and simplex geometry (Fiedler, 2011).

Returning to the previously considered right-stochastic transition matrices $\mathbf{P}$ for a discrete Markov chain and finite state space $S = \{0, 1, 2, ..., N\}$, we can construct the distribution of the chain at time $t + 1$:

$$\pi_{t+1}(j) = \mathbb{P}(X_{t+1} = j) = \sum_{i=1}^{N} \mathbb{P}(X_t = i)\mathbb{P}(X_{t+1} = j | X_t = i) = \sum_{i=1}^{N} \pi_t(i) p_{ij}. \quad (5.1)$$

The above equation can be written in matrix form simply as $\pi_{t+1} = \pi_t \mathbf{P}$ and describes the distribution of the chain at step $t + 1$. By induction, we can arrive at the distribution of the chain at step t using the initial distribution and the transition matrix $\mathbf{P}$, and this is given by $\pi_t = \pi_0 \mathbf{P}^t$.

Definition 5.5. *Let* i *and* j *be two states of a Markov chain.* j *is accessible from* i *if, having started in state* i *the chain can visit state* j *with non-zero probability. States* i *and* j *are defined to communicate if* i *is accessible from* j *and*

j is accessible from i. A Markov chain is irreducible if all pairs of its states communicate.

Definition 5.6. *For a Markov chain* $\{X_0, X_1, ...\}$*, the period of a state i is the greatest common divisor* $d_i = gcd\{n : \mathbf{P}^t(i,i) > 0\}$*. An irreducible Markov chain is known to be aperiodic if its period is 1.*

From a statistical viewpoint we may be interested in estimating the transition probabilities p_{ij} for a given physical phenomenon. Let now $S = 1, 2, ..., D$ be the discrete state space, and observe c successive transitions of the Markov chain and the last transition is defined by $X_c = x_c$. The likelihood function for a transition matrix $\mathbf{P}$ is given by

$$l(\mathbf{P}|\mathbf{x}) = \prod_{i=1}^{D}\prod_{j=1}^{D} p_{ij}^{n_{ij}}. \tag{5.2}$$

where n_{ij} is the number of observed transitions from state i to state j, and the total number of transitions is c. We can estimate $\mathbf{P}$ by its maximum likelihood estimate $\hat{\mathbf{P}}$ where each element of the matrix is given by

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_{j=1}^{D} n_{ij}}. \tag{5.3}$$

This frequentist approach considers only the aleatory uncertainty - the uncertainty due to the probabilistic variability of observed moves of the Markov chain. If we were interested further in uncertainty surrounding the transition probabilities themselves, the epistemic uncertainty, we would need to adopt a Bayesian approach to the problem. In the frequentist framework, this could be akin to modelling random transition matrices Takahashi (1969). Due to the Markov property, the transition matrix $\mathbf{P}$ can be decomposed into its individual rows with non-negative entries and unit sum. This may remind us of the compositional framework explored in earlier chapters, and we explore this topic later in the chapter.

5.2 Stationary distribution of a discrete-time Markov chain

If it is the case that the distribution of the chain at every next time step is identical to its distribution at the previous time step, then π is called the stationary distribution of the Markov chain. Formally, it is π such that the following relationship is satisfied:

$$\pi(j) = \sum_{i \in S} \pi(i) p_{ij}, \forall i \in S. \tag{5.4}$$

The concept of the stationary distribution is essential in the following Basic Limit Theorem:

Theorem 5.1. *Suppose $X_0, X_1, X_2, ...$ is an irreducible and aperiodic Markov Chain with a stationary distribution $\boldsymbol{\pi}(\cdot)$ with arbitrary initial distribution π_0. Then $\lim_{t \to +\infty} \pi_t(i) = \pi(i), \forall i \in S$*

If such a stationary distribution $\boldsymbol{\pi}$ exists for a specified Markov chain, then it is the unique solution to

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}. \tag{5.5}$$

For example, for a 2-state discrete-time Markov chain with transition matrix defined by

$$\mathbf{P} = \begin{pmatrix} p_{11} & 1 - p_{11} \\ 1 - p_{22} & p_{22} \end{pmatrix}$$

the stationary distribution $\boldsymbol{\pi}$ of the chain remaining in each state is given by the balance equations (5.5):

$$\pi_1 = \frac{1 - p_{22}}{2 - p_{11} - p_{22}}, \quad \pi_2 = \frac{1 - p_{11}}{2 - p_{22} - p_{11}}, \quad \text{since } \pi_1 + \pi_2 = 1.$$

In the setting of row-stochastic matrices $\mathbf{P}$, we would like to analyse how incorporating uncertainty into the matrix elements by the use of uniformly generated random numbers (other distributions will also be considered) affects the stationary distribution $\boldsymbol{\pi}$. Such a stationary distribution satisfies Equation 5.5. In higher D dimensions -

$$\mathbf{P} = \begin{pmatrix} \mathrm{p}_{11} & \mathrm{p}_{12} & \dots & 1-\sum_{j=1}^{D-1} p_{1j} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{p}_{D1} & \mathrm{p}_{D2} & \dots & 1-\sum_{j=1}^{D-1} p_{Dj} \end{pmatrix}.$$

In order to gauge stationary behaviour of a D-dimensional transition matrix we may again solve Equation 5.5. However, as D increases computational approaches may be sought. A naive way to understand long-term behaviour of the chain, provided all elements p_{ij} are known, we can examine the limiting distribution $\boldsymbol{\pi} = \lim_{N\to\infty} \mathbf{P}^N$ and if the result satisfies Equation 5.5 then $\boldsymbol{\pi}$ is the stationary distribution. Alternatively, eigendecomposition of the transition matrix $\mathbf{P}$ can be carried out as a way to identify the stationary distribution.

Returning to the 2-dimensional example, to find the eigenvalues of $\boldsymbol{P}$, we evaluate for λ the following determinant

$$\det\left[\begin{pmatrix} p_{11} & 1-p_{11} \\ 1-p_{22} & p_{22} \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right] \stackrel{!}{=} 0. \tag{5.6}$$

and the solutions to the characteristic equations stand at $\lambda_1 = 1, \lambda_2 = p_{11} - p_{22}$.

Next, we seek the eigenvectors $\boldsymbol{v} = (v_1, ..., v_D)$ associated with the found values for λ, in the usual sense:

$$\boldsymbol{P}\mathbf{v} = \lambda\mathbf{v}. \tag{5.7}$$

However, our interest lies in the stationary distribution $\boldsymbol{\pi}$, which can be derived from the normalised eigenvector associated with λ_1 above. This can be

done either analytically or computationally. As already discussed, we can also investigate the limiting distribution by raising $\boldsymbol{P}$ to a high power, and observing whether any convergence occurs.

We can go further and diagonalise our transition matrix to take the following form through eigendecomposition

$$\boldsymbol{P} = \boldsymbol{M}\boldsymbol{D}\boldsymbol{M}^{-\mathbf{1}}, \tag{5.8}$$

with $\boldsymbol{M}$ being the invertible matrix of its eigenvectors, and $\boldsymbol{D}$ is a diagonal matrix with elements being the eigenvalues of $\boldsymbol{P}$. Then, if we were to raise $\boldsymbol{P}$ to a high power, it holds that

$$\boldsymbol{P}^N = \boldsymbol{M}\boldsymbol{D}^N\boldsymbol{M}^{-1}. \tag{5.9}$$

In theory, this eases any calculations since $\boldsymbol{D}^N$ is simply its diagonal elements raised to the needed power.

5.2.1 Example

Maximum wind speed measurements have been taken at the Leeds Bradford Airport weather station during the month of February 2021. 28 wind speeds have been recorded during the month (1/02/2021 until 28/02/2021 inclusive) in miles per hour (m.p.h). Wind speeds have been classified into three states on each given day: Class 1 denotes wind speeds less than 10 m.p.h; Class 2 denotes speeds strictly between 10 and 20 m.p.h; finally Class 3 denotes speeds greater than 20 m.p.h on a given day.

A sequence of wind speed classes are given in the following:

1 1 1 2 1 2 3 2 2 2 2 3 2 3 2 2 2 2 3 2 2 2 3 3 2 2 1 1

The above can be interpreted as follows, if we consider the initial state 1 - over

the month of February 2021, on three day-pairs the maximum wind speed did not exceed 10 m.p.h, then on two occurrences the wind speed changed class from 1 to 2, and finally on no day the maximum wind speed increased from Class 1 to Class 3. Similar reasoning for the remaining classes can lead to estimate a transition matrix:

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{pmatrix} = \begin{pmatrix} 3/5 & 2/5 & 0 \\ 1/8 & 9/16 & 5/16 \\ 0 & 5/6 & 1/6 \end{pmatrix}$$

In all instances, the stationary distribution is found to be $\boldsymbol{\pi_W} = (0.185, 0.593, 0.222)$ to 3 d.p. which means that if the transition matrix probabilities were to be a true reflection of wind speeds beyond February 2021, the speeds would in the class 10 m.p.h to 20 m.p.h almost 60% of the time, in the range 20 m.p.h or greater 22% of the time, and remained under 10 m.p.h 18.5% of the time.

So far, we have assumed the elements of $\boldsymbol{P}$ are deterministic. However, we may also wish to incorporate some uncertainty in the matrix elements. This can also aid in estimating the transition matrix $\boldsymbol{P}$ from a Bayesian perspective. We would look towards the most probable matrices $\boldsymbol{P}$ - those transition matrices that, in the context of Bayesian analysis, have the highest posterior probability. For this, a prior distribution needs to be chosen, and a popular choice is a conjugate family, such as the Dirichlet. Before any movements of the Markov chain are observed, the statistician may wish to reflect ignorance in the prior distribution, as to reduce the bias brought in, especially if there is no reason to make the prior subjective. This choice drives the posterior probability (and hence, estimation of the transition matrix $\boldsymbol{P}$) to be driven by the observed movements of the Markov chain. This Bayesian approach has the advantages of introducing greater numerical stability to estimating $\boldsymbol{P}$, as well as adhering to any physical constraints of the problem. Similarly, the Bayesian approach provides an estimate of uncertainty

about $\boldsymbol{P}$ when the number of movements between states is small.

Let $\mathbf{P}_i$ denote the ith row of the transition matrix $\mathbf{P}$. Then, a multivariate distribution of a Dirichlet type can be assigned to each $\mathbf{P}_i$. For the moment, let us assume the simplest conjugate Dirichlet distribution:

$$\mathbf{P}_i \sim \text{Dir}(\boldsymbol{\alpha}_i); \boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iD}),\ i = 1, 2, ..., D.$$

The multinomial likelihood in this scenario would be the observed transitions $\mathbf{X}$ assumed to be independently distributed, and by conjugate analysis, the posterior uncertainty in each row of transitions of the Markov chain can be expressed by:

$$\mathbf{P}_i|\mathbf{x} \sim \text{Dir}(\boldsymbol{\alpha}_i + \mathbf{n}_i) \tag{5.10}$$

.

where $\mathbf{n}_i$ are the counts of the observed transitions in state i.

For such right-stochastic transition matrices, we can use the families of distributions described in Chapter 4 in order to accommodate for uncertainty in a Markov chain's transition between states. If we have little prior information about the possible moves of the Markov chain, we may wish to adopt a Jeffreys' prior, such that $\mathbf{P}_i \sim \text{Dir}(0.5, 0.5, ..., 0.5)$ or a uniformly flat prior on the simplex $\mathbf{P}_i \sim \text{Dir}(1, 1, ..., 1)$. The resulting posterior distribution is also conjugately Dirichlet, however driven by the observed transitions $\boldsymbol{X}$.

Another improper prior for this case could be when the concentration parameter $\boldsymbol{\alpha}$ of the Dirichlet distribution approaches zero, which yields:

$$\mathbb{P}(\mathbf{P}_i) \propto \prod_{j=1}^{D} \frac{1}{p_{ij}}; \forall i, j = 1, 2, ..., D.$$

The posterior distribution is thus $\mathbf{P}_i|\mathbf{x} \sim \text{Dir}(n_{i1}, ..., n_{iD})$. This is an interesting

case, as the posterior mean here is exactly the maximum likelihood estimate: $\mathbb{E}(p_{ij}|\mathbf{x}) = \frac{n_{ij}}{n_i} = \hat{p}$.

There may be situations of scarce yet important a-priori information, for example that transitions between certain states are not possible. For example, as considered for a discrete time birth-death Markov chain. Alternatively, it can be that the probability of remaining in the same state is zero, hence $p_{ii} = 0, \forall i = 1, ..., D$. Thus, the marginal prior distribution should be restricted solely to those states where transitions are possible, and the other transition probabilities be made zero. For example, in the instance of the Dirichlet:

$$\mathbb{P}(p_{ij}) = \begin{cases} \frac{p_{ij}^{\alpha_i - 1}(1-p_{ij})^{\beta_i - 1}}{\mathrm{B}(\alpha,\beta)}, & \text{if i} \neq \text{j, } \alpha, \beta > 0; \\ 0, & \text{if i=j.} \end{cases} \tag{5.11}$$

As an alternative, it is possible to set up a hierarchical prior structure on the transition probabilities for a discrete-time Markov chain, if it is known that specific chain transitions are not possible due to physical constraints of a particular problem. In the example of the Dirichlet, we may impose a distribution on the concentration parameters $\boldsymbol{\alpha}$ to lie in a certain range, such as $\alpha_{ij} \sim \text{Uniform}[1, 5]$. This choice excludes distributional mass lying close to the edges of the simplex.

Next, if we were to express each element of the transition matrix as a probability distribution, we may equally construct the problem row-wise or element-wise. Each row $\boldsymbol{P_i}$ of transition matrix $\boldsymbol{P}$ may follow some multivariate distribution, such as the familiar Dirichlet($\boldsymbol{\alpha_i}$). The transition matrix may then be written as:

$$\mathbf{P} = \begin{pmatrix} \text{Dir}(\boldsymbol{\alpha_i}) \\ ... \\ \text{Dir}(\boldsymbol{\alpha_D}) \end{pmatrix}$$

Equally, element-wise we can express marginal uncertainties:

$$\mathbf{P} = \begin{pmatrix} \text{Beta}(\alpha_{11}, \sum_{j=2}^{D} \alpha_{1j}) & ... & \text{Beta}(\alpha_{1D}, \sum\limits_{j \neq D} \alpha_{1j})) \\ ... & ... & ... \\ \text{Beta}(\alpha_{D1}, \sum_{j=2}^{D} \alpha_{Dj}) & ... & \text{Beta}(\alpha_{DD}, \sum\limits_{j \neq D} \alpha_{Dj})) \end{pmatrix}$$

.

provided that marginal distributions exist. We may be interested in stationary long-term behaviour of a Markov chain defined by the above transition matrix. When transition elements are deterministic, or based on estimates through observed movements of the Markov chain, the stationary distribution is yielded through solving the detailed balance equation, or by carrying out eigendecomposition. Similarly, it may occur that the limiting distribution of the chain is also the stationary, as was described earlier in the section. Now when elements of $\boldsymbol{P}$ are expressed as probability distributions, it may be difficult to obtain an analytic closed-form for the stationary distribution. Consider a simple example when $\boldsymbol{P}$ represents a 2×2 matrix, such that the Markov chain moves solely between two states. In the above set-up expressing elements of $\boldsymbol{P}$ as marginal Beta distributions it holds that:

$$\mathbf{P} = \begin{pmatrix} \text{Beta}(\alpha_{11}, \alpha_{12}) & \text{Beta}(\alpha_{12}, \alpha_{11}) \\ \text{Beta}(\alpha_{21}, \alpha_{22}) & \text{Beta}(\alpha_{22}, \alpha_{21}) \end{pmatrix}$$

Carrying out eigendecomposition or solving the detailed balance equations for a reversible $\boldsymbol{P}$ from Equation 5.5 would involve (in the simplest case) division of one probability function by another, and convolution of probability distributions in order to find the determinant. This process may not even be analytically feasible for $D > 2$ and is made even more complicated by the fact that the elements in each row are not independent of each other.

A more practical way of understanding the stationary distribution is through con-

structing random matrices by drawing from the Dirichlet distribution row-wise, then observing limiting behaviour of the chain through matrix multiplication. Attempt at analytic solution would again involve convolutions of dependent random variables, if $\boldsymbol{P}$ were taken as defined immediately above.

5.2.2 Example

The following cases represent three-state Markov chains with row-wise uncertainty following the Dirichlet distribution with varying concentration parameters $\boldsymbol{\alpha}$. Simulations are taken from each transition matrix and limiting behaviour is observed through iterative matrix multiplication. In total 100,000 transition matrices are simulated for each scenario. Contour plots on the simplex represent the row-wise distribution of the stationary vector, and are used instead of scatter plots on the simplex for illustration of the distribution. The colour of each contour plot represents the density of the stationary distribution in that region, and the vertices of the simplex $P(j)$ denote the corresponding element in the stationary distribution, $j = (1, 2, 3)$.

Suppose a 3-state Markov chain has the following transition matrix

$$\boldsymbol{P_1} = \begin{pmatrix} \text{Dir}(0.5, 0.5, 0.5) \\ \text{Dir}(0.5, 0.5, 0.5) \\ \text{Dir}(0.5, 0.5, 0.5) \end{pmatrix}$$

In a Bayesian analysis this case can represent the uninformative Jeffreys' prior distribution over each row of $\boldsymbol{P_1}$, and the stationary behaviour of $\boldsymbol{P_1}$ can be observed in Figure 5.1.

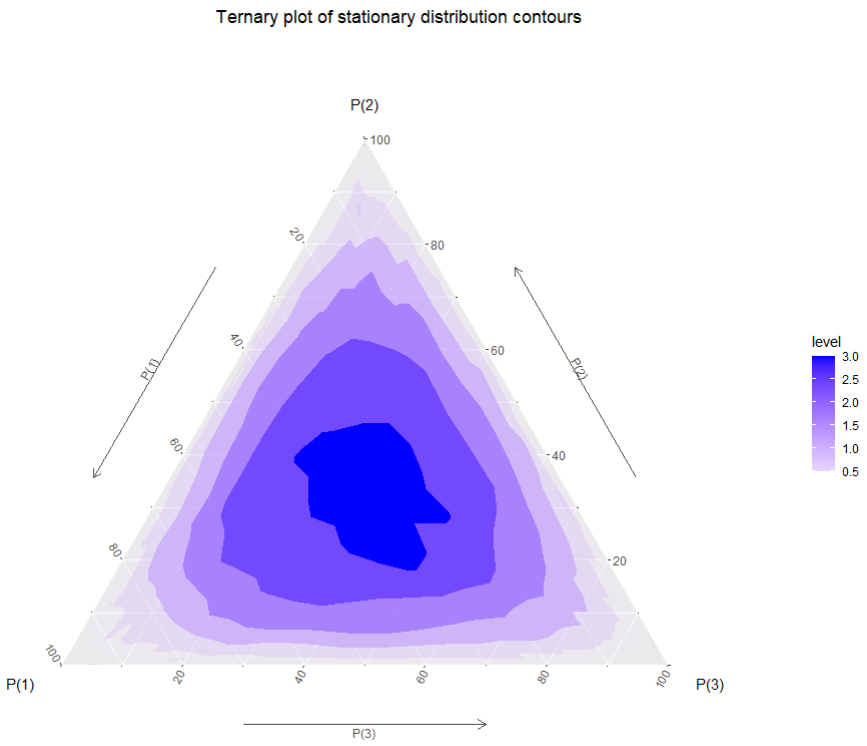


Figure 5.1: Stationary distribution over row-wise uncertainty of $\boldsymbol{P_1}$.

We can see that the long-term behaviour of the chain with the Jeffreys' transition matrix shows moderately equal spread across the entire simplex space with a central mass. Similar behaviour is also the case for a deterministic transition matrix with equal probabilities of movements to other states.

As alluded to in Chapter 2, if we express uncertainty using the Uniform distribution but on the simplex, this is the case of Dirichlet concentration parameters $\alpha_{ij} = 1$. It is interesting to contrast this scenario, depicted in transition matrix $\boldsymbol{P_2}$, with the previous Jeffreys' transition matrix.

$$\boldsymbol{P_2} = \begin{pmatrix} \text{Dir}(1,1,1) \\ \text{Dir}(1,1,1) \\ \text{Dir}(1,1,1) \end{pmatrix}$$

In Figure 5.2 we can see that in the central region of the simplex there is considerably more mass than in Figure 5.1, as seen by the higher density curve levels

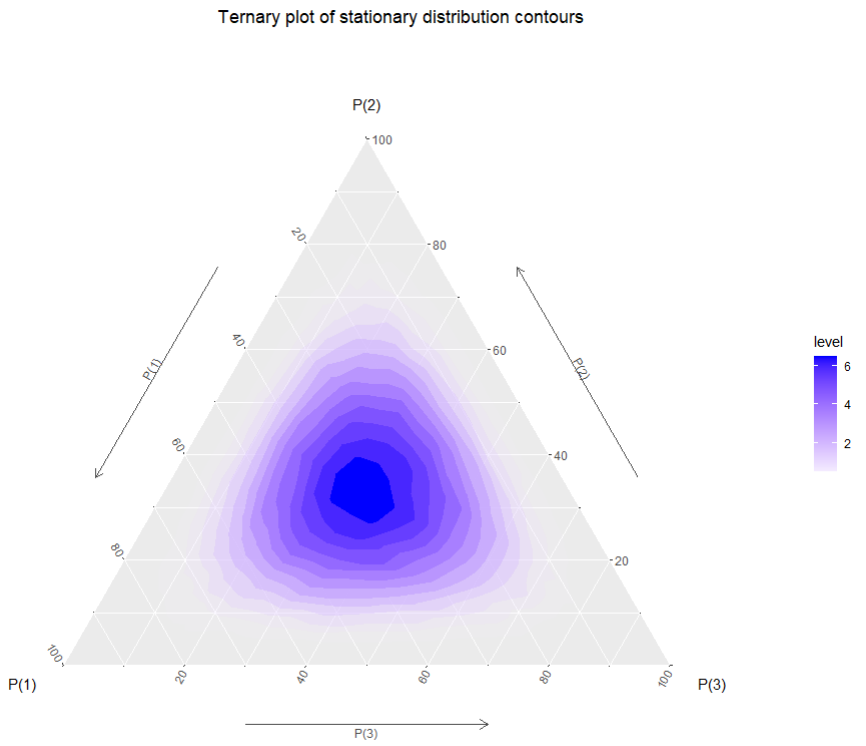


Figure 5.2: Stationary distribution over row-wise uncertainty of $\boldsymbol{P_2}$.

and also the spread of the density, which is lacking at the edges of the simplex in Figure 5.2.

Now, let us move towards examining the stationary behaviour of $\boldsymbol{P_3}$, in which uncertainty in the second row is expressed by driving the mass of the Dirichlet distribution away from the centre and towards the edges of the simplex. This is expressed by $\boldsymbol{\alpha} = (0.1, 0.1, 0.1)$. The stationary distribution of this set-up can be seen in Figure 5.3.

$$\boldsymbol{P_3} = \begin{pmatrix} \text{Dir}(1,1,1) \\ \text{Dir}(0.1,0.1,0.1) \\ \text{Dir}(1,1,1) \end{pmatrix}$$

We can see that the behaviour of the stationary vector is rather unusual - there is a high distributional mass towards the vertex $P(3)$, as well as other regions of mass in the simplex in Figure 5.3 being as concentrated as those in Figure 5.4

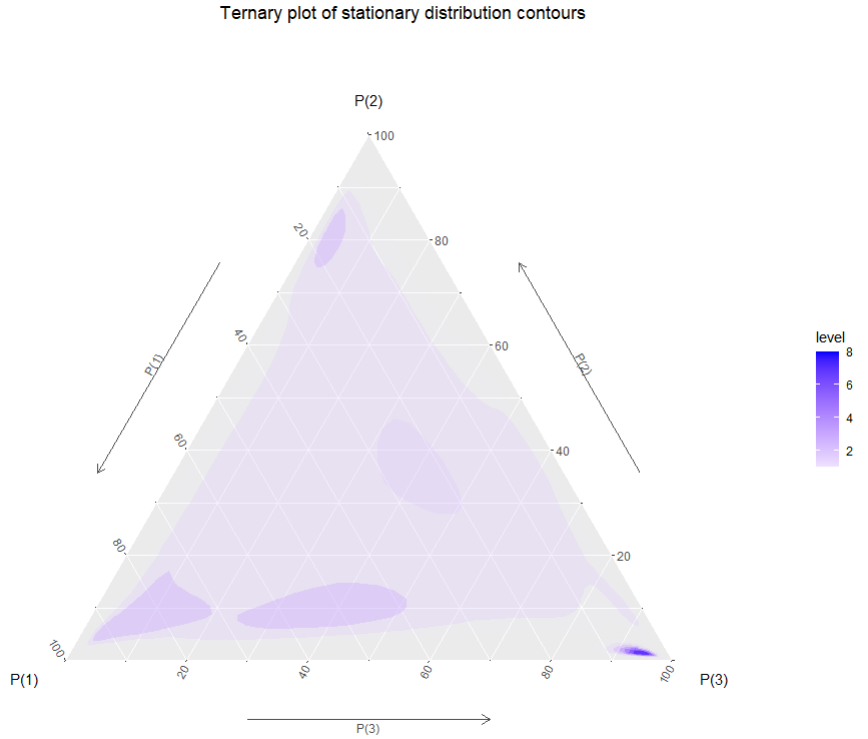


Figure 5.3: Stationary distribution over row-wise uncertainty of $\boldsymbol{P_3}$.

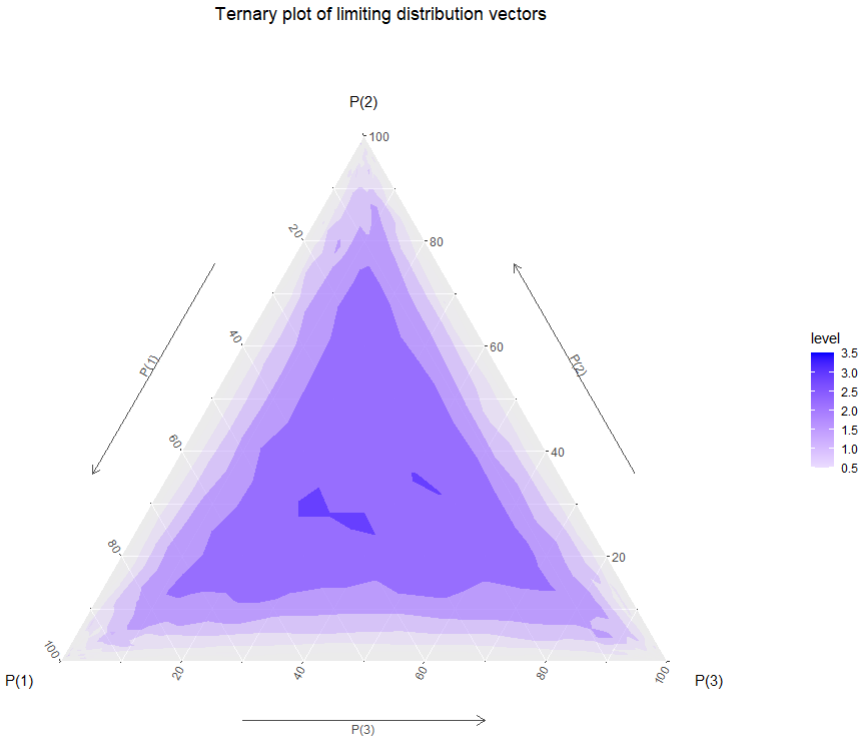


Figure 5.4: Stationary distribution over row-wise uncertainty of $\boldsymbol{P_3}$.

(note different scales of distributional mass). We may suspect that the stationary distribution is not stable, and this is indeed the case if we consider a different starting point (seed) for the matrix simulations. Figure 5.4 shows another instance of limiting behaviour of $\boldsymbol{P_3}$. We can now see a very different depiction to Figure 5.3, and it finds more similarity to the stationary distribution of the Jeffreys' transition matrix above. Further investigation allows us to conclude that the stationary distribution of $\boldsymbol{P_3}$ is highly sensitive towards the matrices generated from row-wise Dirichlet distributions, and the stationary distribution varies between the two scenarios depicted in Figure 5.3 and Figure 5.4. To see this, we can consider marginal distributions of Dirichlet(0.1,0.1,0.1), which would be Beta(0.1,0.2), and compare them to marginal distribution of the last row of $\boldsymbol{P_3}$, being Beta(1,2), as seen in Figure 5.5. We can see that for each draw from Dirichlet(0.1,0.1,0.1) and Dirichlet(1,1,1) the probability of obtaining a random transition close to the value of 1 is quite different. Thus, over time it is possible to see clusters of distributional mass at one of the corners or close to the edges of the simplex, where parameter $\boldsymbol{\alpha}$ values are less than 1. Moreover, if there is a high probability that an element of $\boldsymbol{P_3}$ is in the interval $[0, 0.2]$, this may result in imprecision from a computational standpoint after repeated matrix multiplication. Note that this sensitivity is not reflected in $\boldsymbol{P_1}$, $\boldsymbol{P_2}$ and neither in the cases that follow, and presents one of the issues when parameter values of the Dirichlet distribution are drawn close to the edges of the simplex ($\boldsymbol{\alpha} < 1$).

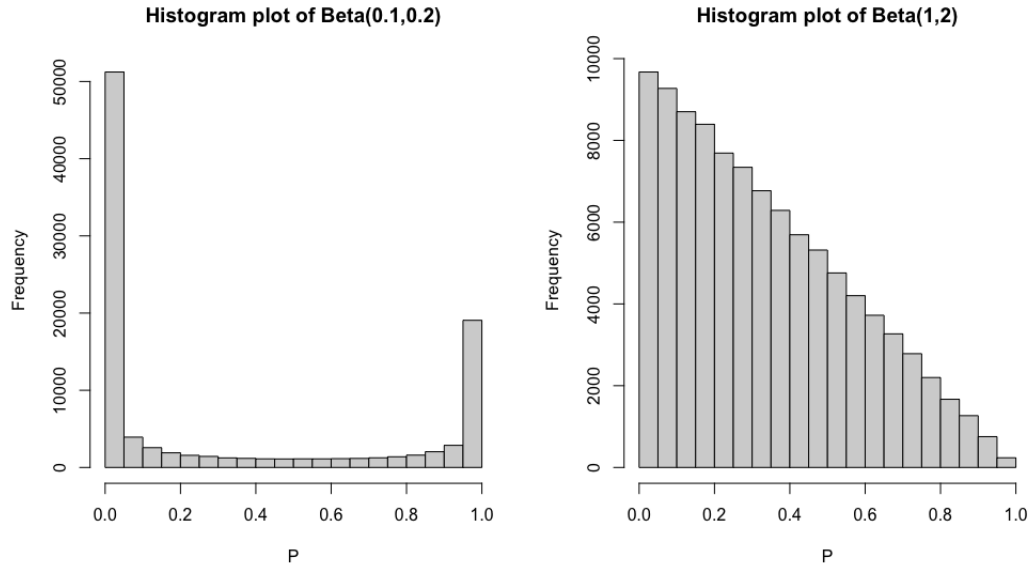


Figure 5.5: Frequency of random draws for each row of $\boldsymbol{P_3}$

Next, let us consider $\boldsymbol{P_4}$ where we express a high preference to vertex 3 in the first row of the transition matrix, which is a situation found often in applications of compositional data. The stationary behaviour of $\boldsymbol{P_4}$ can be seen in Figure 5.6.

$$\boldsymbol{P_4} = \begin{pmatrix} \text{Dir}(1,1,10) \\ \text{Dir}(1,1,1) \\ \text{Dir}(1,1,1) \end{pmatrix}$$

If we only consider the random variable that follows $\text{Dir}(1,1,10)$, the ternary plot of this would show a mass concentrated at the third vertex $P(3)$. However, in Figure 5.6 the stationary mass has been pulled away from $P(3)$ and some distribution mass is found in the centre of the simplex, while the peak of the distribution is on the tertile boundary, where $P(3)$ is around 0.6, while $P(1)$ and $P(2)$ are close to 0.3. This shows that the long-term uncertainty is projected to average between the row-wise uncertainties in $\boldsymbol{P_4}$, and this behaviour is stable unlike what was seen with $\boldsymbol{P_3}$.

To follow from this, let us portray a scenario where significant probability mass is assigned to the first vertex in the last row of the transition matrix $\boldsymbol{P_5}$

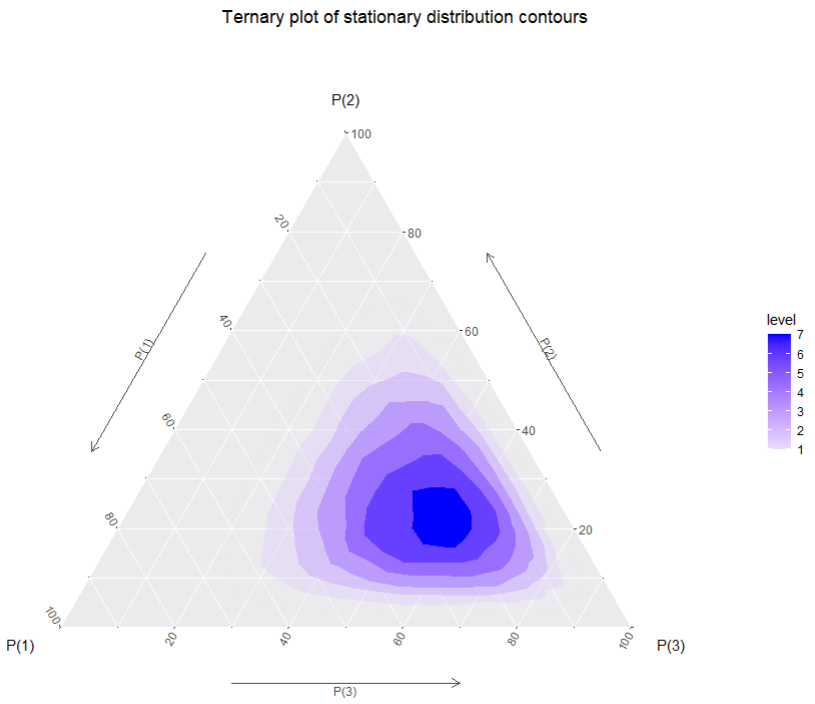


Figure 5.6: Stationary distribution over row-wise uncertainty of $\boldsymbol{P_4}$.

$$\boldsymbol{P_5} = \begin{pmatrix} \text{Dir}(1,1,10) \\ \text{Dir}(1,1,1) \\ \text{Dir}(10,1,1) \end{pmatrix}$$

Stationary behaviour can be observed in Figure 5.7. Again, we see that the stationary distribution is highly concentrated (regardless of the simulation starting seed). The stationary mass is also found halfway between vertices $P(1)$ and $P(3)$ on the simplex, and this follows from how $\boldsymbol{P_5}$ was defined.

Finally, let us investigate the Wind Speed example from earlier in this chapter. Suppose that instead of the transition matrix $\boldsymbol{W}$ defined previously from daily wind speed changes, we define an uncertain transition matrix expressed as $\boldsymbol{P_6}$

$$\boldsymbol{P_6} = \begin{pmatrix} \text{Dir}(3,2,0.01) \\ \text{Dir}(2,9,5) \\ \text{Dir}(0.01,5,1) \end{pmatrix}$$

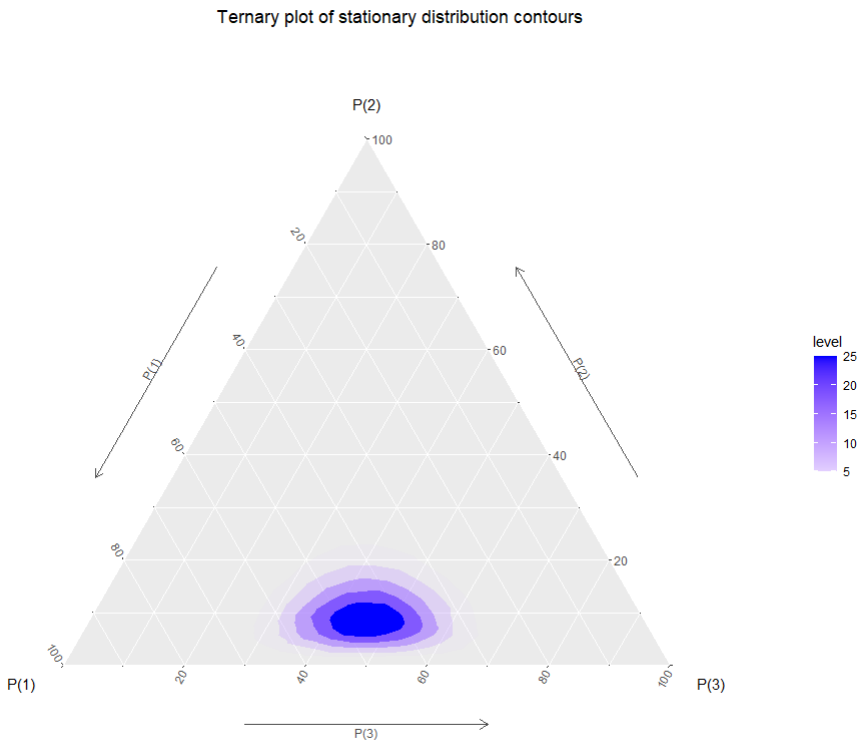


Figure 5.7: Stationary distribution over row-wise uncertainty of $\boldsymbol{P_5}$.

Estimates for concentration parameters $\boldsymbol{\alpha}$ of the Dirichlet distribution are indeed taken from the data, but realisations of these Dirichlet vectors will follow those exact values expressed in $\boldsymbol{W}$ with very small probability. Again, if we simulate 100,000 transition matrices from $\boldsymbol{P_6}$, the long-term behaviour can be seen in Figure 5.8. Very small α value was assigned to transitions from class 3 to class 1, and vice-versa. This is to reflect the estimated $\boldsymbol{W}$, but also to include a small chance of those transitions taking place.

Resulting stationary behaviour is in line with $\boldsymbol{\pi_W} = (0.185, 0.593, 0.222)$ to 3 d.p, as seen when elements of $\boldsymbol{W}$ are deterministic and estimated from the observed wind speeds in February 2021. The mass of the density in Figure 5.8 reaches its peak when $P(1)$ is around the value 0.2, and for $P(2)$ and $P(3)$ it is around 0.7 and 0.2 respectively. This is coherent with π_W as above, yet in this case we are able to demonstrate sensitivity of stationary behaviour with respect to uncertainty around the elements of the transition matrix $\boldsymbol{W}$. In this applied example we have conditioned on the fact that on February 1 2021 the maximum

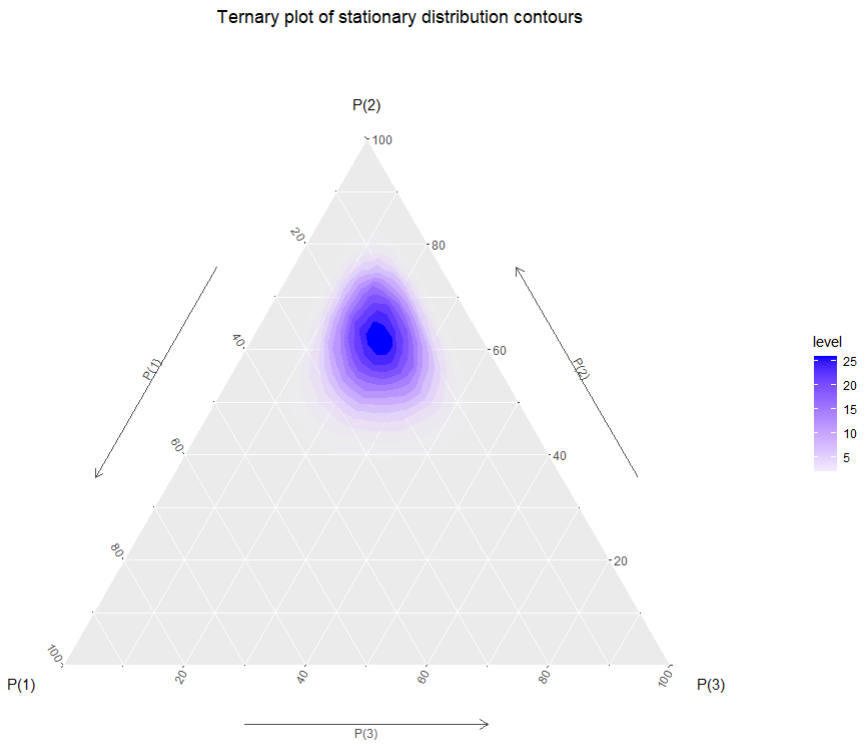


Figure 5.8: Stationary distribution over row-wise uncertainty of $\boldsymbol{P_6}$.

wind speed was in class 1, and have not included any uncertainty about the initial state of the Markov chain. The same assumption was made about $\boldsymbol{P_1}$ through to $\boldsymbol{P_6}$. The initial distribution is not seen to have an effect on the limiting behaviour due to the Markov property, and so its influence is negligible in the long term. However, should we wish to explicitly define a distribution for the initial state $\boldsymbol{P}_0$, it could be the multinomial distribution, such that $\mathbb{P}(\boldsymbol{P}_0 = p_0) = \theta_{P_0}$ for $\theta_d \in (0, 1)$ and $\sum_{d=1}^{D} \theta_d = 1$. Then, for a conjugate analysis θ_d can again follow the Dirichlet distribution, and Bayesian inference for $\boldsymbol{P}$ can be conducted as previously with observed transitions of the Markov chain.

In all the above, we have assumed that transitions are possible between all states of the Markov chain that we have defined, even with a small probability. The set-up of the simulation exercise has also assumed that no other states are possible, and the observed states are all that exist. However, with increasing dimensionality of the transition matrix, or increasing complexity of the relationships between states, the considerations above can become of critical importance. Especially if

there is suspicion that more states occur than can be currently observed, a way to approach this is through Hidden Markov Models (HMMs). These can be thought as an extension to a two-layer Markov model - the top layer is unobserved and is represented as a Markov chain (discrete and continuous-time extensions are possible), and the bottom observed layer is dependent on the states of the layer above. This thesis does not consider treatment of Hidden Markov models, which is a separate area of study. Introduction to HMMs can be found in the work of Dymarski (2011), and they are exceedingly popular in modelling financial processes (Mamon and Elliott, 2007) and animal movement models (Zucchini and MacDonald, 2009), for example.

5.3 Other considerations

In this chapter, we have solely considered right-stochastic matrices, whose rows adhere to the sum-one constraint. Column-wise, we have assumed independence, making parallels with compositional data analysis explored in earlier chapters of this thesis. In this section, we deviate from the row-sum constraints and illustrate applications of Bayesian modelling to other types of constrained matrices. The first such example deals with confusion matrices. A confusion matrix $\boldsymbol{C}$ is a square matrix that is used to examine the accuracy of a model or an algorithm in comparison with some ground truth. Rows of the confusion matrix usually represent true states (classes) in an experiment, and the columns represent predicted states (classes) as driven by the model or the algorithm. The smallest confusion matrix $\boldsymbol{C}_1$ is simply an accuracy score, or the outcome of a statistical test. The next more complex confusion matrix has size (2×2) and can occur for a binary (classification) problem:

$$\boldsymbol{C}_2 = \begin{pmatrix} c_{TP} & c_{FN} \\ c_{FP} & c_{TN} \end{pmatrix}$$

where subscript TP stands for the number of true positives; FN represents false negative, FP is the number of false positives, and TN is true negative. Caelen (2017) examines the (2×2) confusion matrix from a Bayesian standpoint, treating the elements in $\boldsymbol{C}_2$ as realisations of the familiar multinomial distribution. Caelen then assigns a conjugate Dirichlet prior distribution to the probabilities that drive each element of $\boldsymbol{C}_2$ above, and posterior distribution as described in Equation 5.10 thus follows. This conjugate analysis is used for comparison between two competing models (and thus, two realisations of $\boldsymbol{C}_2$) to the author's interest. The key motivation in Caelen's work is to investigate a scenario where no training or testing data set is available for a scientist to conduct a bootstrap-like approach to compare between two competing models. Instead, the only information the scientist has access to is the confusion matrix $\boldsymbol{C}$. Caelen finds that the Bayesian approach with the conjugate Dirichlet distribution yields the same summary statistics as the bootstrap method. However, the former also includes uncertainty about the unknown probability vector in the multinomial distribution, which is expressed as counts in the confusion matrix. As with the numerical investigation of stationary behaviour of a Markov chain in this chapter, Caelen notes that injecting prior knowledge into the Dirichlet prior in the form of specified concentration parameters, lowers the variance of the posterior distribution. A distinct difference between Caelen's modelling approach to what has been considered in this chapter, is that Caelen does not treat the rows of the confusion matrix as independent. Instead of the previously-defined

$$\mathbf{P} = \begin{pmatrix} \text{Dir}(\alpha_{11}, \alpha_{12}) \\ \text{Dir}(\alpha_{21}, \alpha_{22}) \end{pmatrix}$$

for a (2×2) matrix of uncertain probabilities, Caelen defines

$$\mathbf{P}^* \sim \text{Dir}(\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}),$$

because there is no constraint on each row of $\boldsymbol{C}_2$. This relaxation of constraints allows for dependence modelling between columns of $\boldsymbol{C}_2$, as well as the rows.

In a completely opposite scenario, we may wish to constrain the columns of our transition matrix $\boldsymbol{P}$ as well as the rows. Such matrices have been defined earlier in this chapter as doubly stochastic (bistochastic). Let us refer to a bistochastic matrix as $\boldsymbol{P}_{\mathcal{B}}$ and its elements $p_{\mathcal{B}ij}$ for $i, j = 1, ..., D$. The set of doubly stochastic matrices is known as a Birkhoff polytope (Fiedler, 2011) and has dimension $(D-1)^2$ and can be represented as a convex polyhedron in $\mathbb{R}^{(D-1)^2}$. Doubly stochastic matrices find use in modelling physical processes (Louck, 1997) and in machine learning applications, for example, in spectral clustering problems and affinity matrices, where reducing a matrix to be doubly stochastic can make a data set more easily used with clustering algorithms (Zass and Shashua, 2006).

The most simple bistochastic matrix has value 1 for its diagonal elements, and zeros elsewhere. Other examples are given by

$$\mathbf{P}_{\mathcal{B}} = \begin{pmatrix} 1 - p_{\mathcal{B}} & p_{\mathcal{B}} \\ p_{\mathcal{B}} & 1 - p_{\mathcal{B}} \end{pmatrix},$$

for some $p_{\mathcal{B}} \in [0, 1]$;

$$\mathbf{P}_{\mathcal{B}} = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix},$$

or from a different viewpoint

$$\boldsymbol{P}_{\mathcal{B}} = \begin{cases} 1/p_{\mathcal{B}ij} & \text{if i} \neq \text{j, i=j=1,...,D;} \\ 0, & \text{otherwise.} \end{cases} \tag{5.12}$$

The general structure above can be extended to form larger doubly-stochastic

matrices of several blocks of $\boldsymbol{P}_{\mathcal{B}}$ on the diagonal, and zero values elsewhere. A doubly stochastic matrix can be generated from a $D \times D$ matrix through iterative normalisation of rows and columns of the original matrix. Several works have also considered uncertainty about the elements in bistochastic matrices - Guillotte and Perron (2012) has shown the existence of a Jeffreys' prior distribution over $\boldsymbol{P}_{\mathcal{B}}$, and some results have been obtained that reflect probability distributions over the Birkhoff polytope, although applications are sparse. Cappellini et al. (2009) give rise to a method to define a probability distribution over the matrix $\boldsymbol{P}_{\mathcal{B}}$ based on assuming a Dirichlet distribution over the columns of $\boldsymbol{P}_{\mathcal{B}}$, then a sophisticated normalisation algorithm is employed. This algorithm is based on Sinkhorn's (Sinkhorn, 1964) iterative procedure that normalises the rows and columns of a square matrix until it is bistochastic up to a certain accuracy. Cappellini et al. (2009) note that this approach can be extended to base distributions other than the Dirichlet. For example, a Gaussian-type distribution and a Uniform-type distribution, however, throughout the work there is explicit statement of either necessary independence between the columns of $\boldsymbol{P}_{\mathcal{B}}$ or necessary correlation between columns of $\boldsymbol{P}_{\mathcal{B}}$ for construction of the matrix distributions.

Chapter 6

Expert elicitation for Bayesian prior specification

6.1 Introduction to expert elicitation

Expert elicitation is an information-gathering exercise carried out to construct an informative prior probability distribution, which plays a key role in Bayesian analysis. Expert elicitation has found ever-increasing uses where data is sparse or the scientific question is not well researched. For instance, we can account for statistical likelihoods of extreme events in civil engineering (Lamb et al., 2017) or risk assessment in financial investments (Katsis et al., 2003). Approaching the problem from a frequentist point of view or using an uninformative flat prior would make Bayesian posterior distributions heavily data-driven. In cases where data is obtained through destruction of an object or where repeat experiments are not possible without significant losses to the client or ethical conflicts, expert elicitation brings about a sensible alternative to formalise and quantify current scientific understanding.

Expert elicitation has been adhered to standardisation and structuring over the last half-century, and the outcomes of the exercise often drive stakeholder deci-

sions. Expert elicitation is part of the wider expert knowledge elicitation (EKE) framework (O'Hagan et al., 2006). The aim of general knowledge elicitation is to extract information about a question of interest to the experts and the stakeholders through an interview or a discussion. Depending on the subject domain and the overall aim of the knowledge elicitation exercise the interview can be free of structure, or be a focused discussion using pre-designed questions.

The first scientific inquiry into expert elicitation was performed by the U.S Nuclear Regulatory Commission in 1975, who exposed substantial differences in the experts' judgements (Frye Jr, 2012). Since then, the statistical community has seen development of elicitation exercises focused on participation of only one expert (Morris et al., 2014) and similarly, elicitation that relies on consensus decision of a group of experts (O'Hagan et al., 2006). With the increase in computing power Bayesian analysis has become a popular approach to drive decision-making processes, and expert elicitation has proven especially popular in describing uncertainty about parameters of prior distributions in public policy decision-making (Morgan, 2014). The use of expert elicitation may similarly be preferred in instances when evidence for more than one statistical model is conflicting, or to consolidate multiple sources of data stemming from different specialisations in a field of science. The development in elicitation techniques and its popularity implies that the exercise should be treated as rigorously as the process of acquiring empirical data.

On the contrary, one may seek other tactics to expert elicitation when there is a high degree of belief that empirical evidence is sufficient to quantify uncertainty about a given quantity. Also, if there is a shortage of financial resources or there exist time constraints to conduct a thorough interview to explore the experts' insights, or upon discovery that the experts' knowledge is not entirely relevant to quantify beliefs. More importantly, if one or more of the experts has significant difficulty in quantitatively expressing their beliefs during the training stage of the elicitation exercise, the exercise should not go forward(O'Hagan et al., 2006).

In this chapter, we investigate past and current approaches to eliciting parametric distributions, and then in Chapter 7 progress to elicitation in the context where uncertainty is modelled about a set of proportions.

6.2 Elicitation of individual judgements

Before proceeding to specific ways to construct probability distributions given a set of expert judgements about a quantity of interest θ, we note the differences between elicitation exercises conducted on an individual and group level.

Since the focus of this thesis lies with parametric probability distributions, let θ be the parameter in focus of the elicitation exercise. Denote θ to be a scalar quantity as the parameter value in a univariate probability distribution and $\boldsymbol{\theta}$ is a vector in a multivariate probability distribution. Also, assume that the appropriate panel of experts is available, has been selected in a fair manner and possesses substantial knowledge about the parameter in question for the construction of a probability distribution. See O'Hagan et al. (2006); Bolger (2018) for discussion on selection of experts.

In the scenario where one expert is available, the elicitation exercise usually commences with a brief to the scientific scenario and may follow with a training exercise about quantities unrelated to the main question. For instance, a popular training question focuses on expressing uncertainty of distance between two cities or populations of countries. After this, the expert is asked carefully constructed questions about the parameter of interest θ. The facilitator must avoid asking directly about summary statistics, for instance, the mean and variance of the probability distribution to build up a plausible quantitative description of θ. Cognitively, quantities like these are very hard to express without the expert succumbing to biases and heuristics (Baddeley et al., 2004).

By setting hard constraints on θ the facilitator may ask the expert for instance, "What is your probability that θ is larger or equal to 5?" or "What is your

probability that θ is between 4 and 5?". If θ can be any value on the real line, in this instance the facilitator has introduced anchors by stating the values "4" and "5", which can influence the expert's judgement on θ. However, if previously the expert has been explicitly asked to provide an interval of values where θ is likely to lie, the above questions would not be unreasonable for one of the intervals in the given range that the expert expressed. Elicitation of quantile judgements is a very popular method of constructing univariate and multivariate distributions. Though, for the multivariate case elicitation of quantile judgements is used to express uncertainty about marginal distributions and more information is later sought about the covariance structure.

Popular approaches to construct a probability distribution around θ involve eliciting at least three judgements from the expert, which can help determine a unique distribution or a small set of plausible distributions that reasonably reflect the expert's knowledge about the uncertainty of θ. Typically, interest lies in the most likely value of θ and a measure of spread of the distribution.

In the quantile approach a popular measure of central tendency of the univariate distribution is the median M. To determine the median, the facilitator could pose the following question - "For which value M would the true value of θ have equal chance of being greater or less than M?". To understand the spread of the distribution about θ upper and lower quartiles are used. For example, for the lower quartile L a question posed by the facilitator could be - "For which value L does the true value of θ have the same probability being below L and between L and M, given that θ is less than M?". This framework is often extended to pose questions about the tertiles of a distribution or percentiles to gain insight into the tails of a distribution (O'Hagan, 2019*b*), alongside a judgement for the mode as the measure of central tendency.

Given the set-up that $\theta \sim f(\cdot)$, the probability distribution $f(\cdot)$ would itself be dependent on some hyperparameters, for example $f(\cdot)$ could be $\mathrm{N}(\mu, \sigma^2)$ or

Beta(α, β). In this case $f(\cdot)$ depends on two parameters, a minimum of two judgements are sought from the expert (O'Hagan et al., 2006) and this is enough to determine sets of parameter values of $f(\cdot)$ that can equally well fit the expert judgements. For instance, in the above setting and given that the target distribution is $\theta \sim$ Beta(α, β) the expert could be questioned about the median and the lower quartile uncertainty for θ as their probabilistic judgements.

Once the expert gives probabilistic judgements about θ the facilitator can ask whether they seem a reasonable representation of the expert's knowledge about θ. If so, the facilitator follows with fitting a probability distribution $\theta \sim f(\cdot)$ to the given judgements, having previously determined one or more probability density functions that may be appropriate for the scientific investigation. Then follows a stage of feedback, where the expert is presented the fitted distribution, often in the form of a graphical plot of the density and summary statistics relevant to the scientific question. The facilitator may prompt the expert to think about whether the peak of the density function reflects the expert's view on the location for the most probable value for θ. Similarly for the tails of the distribution, the expert may be asked to think whether those captured probabilities reflect the expert's knowledge about less likely values of θ. If the expert disagrees with the fitted distribution, the facilitator needs to determine which aspect was not captured well by previous questioning, refine the expert judgements, fit the distribution and reflect on any changes with the expert. Following the fitting procedure is the feedback loop, the expert may be presented the distribution and asked to judge whether the fit reflects their initial beliefs well. Winkler (1967) discusses existence of such a "satisficing" prior distribution that reflects the expert's judgement during an elicitation exercise. If a satisfactory fit is not achieved, the expert would return to the previous stage and provide judgements about a slightly differing set of quantiles. Similarly, the facilitator may not show the fitted distribution, and instead ask for other quantile assessments as a form of a validation of the judgements. There are also advantages to eliciting more

judgements than the minimum number required to fit a distribution - practice known as overfitting. (O'Hagan et al., 2006) supports that overfitting allows for expression of uncertainty about the expert judgements without the facilitator's input on the feedback discussion, which may introduce anchoring effect. However, attempting to fit a probability distribution to many elicited judgements may be more challenging, since this also rests on the choice of distance metric (O'Hagan et al., 2006). Garthwaite et al. (2005) discuss the advantages of questioning the experts about equal odds, rather than eliciting judgements about probability percentiles. Similarly, using this method avoids any anchoring of particular length values.

If there are any inconsistencies of the judgements and the state space of θ, they may too be addressed. For example, as we see in Chapter 7, this could be the case where resultant expected values of the Dirichlet variables do not adhere to the sum-unity constraint. An alternative outcome of the elicitation exercise is that the expert gives reason that the underlying distribution function (or the set of plausible functions) is unsuitable for expressing their uncertainty about θ.

There exist practised alternatives to the bisection method. A graphical interface is often a helpful tool in an elicitation exercise, as it may help alleviate experts' decision fatigue and accelerate the feedback procedure (Goldstein and Rothschild, 2014). One such example of a graphical procedure is the trial roulette method. The sample space for θ is divided into b equally-spaced bins and the expert is provided with n chips which need to be allocated in a way to reflect the expert's judgement about θ. For example, placing two chips in a certain sub-interval reflects twice as much confidence as placing one chip in the same interval. As a result, the expert is able to see an image resembling a histogram, from which the parameter θ is estimated by software such as the MATCH Uncertainty Elicitation Tool (Morris et al., 2014). A particular distribution may be specified by the facilitator, or a best fit from several candidate distributions is selected by the software. In instances where it is not as obvious to structure questions to

express uncertainty about θ that would have reasonable meaning to an expert, the roulette method may be of more use.

In the scenario where a more flexible distribution is needed, the facilitator may look into eliciting judgements about a distribution with more than one mode. This could occur if there are two or more strongly influencing factors in the scientific question, for example, if event A occurs then θ is expected to increase, but if event B occurs then the trend is reversed. The facilitator could pose questions about conditional statements for both A and B and yield a fitted mixture distribution based on these statements. Alternatively, a graphical method such as MATCH's roulette could be used, but for good practice this should be reinforced with the above conditional judgements.

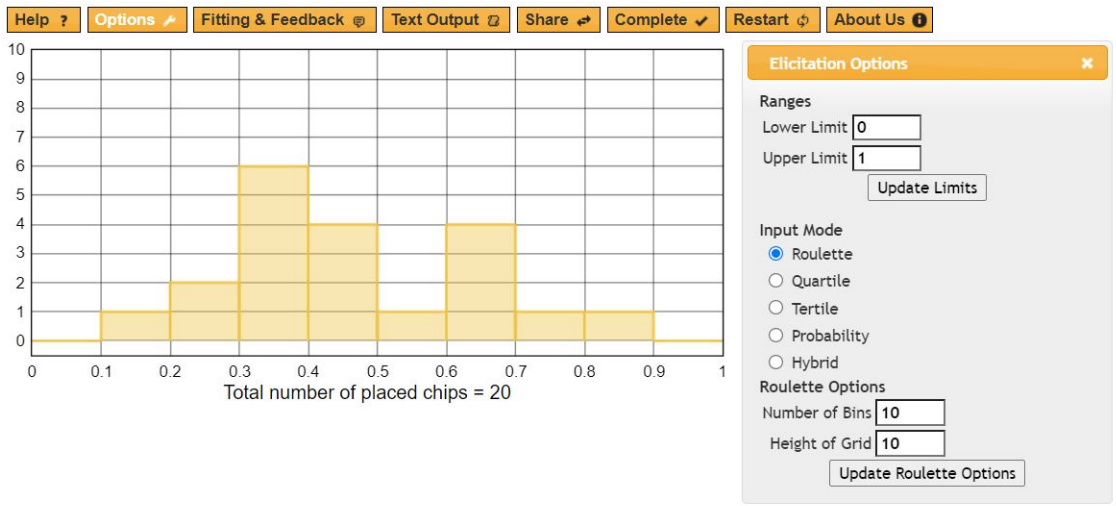


Figure 6.1: MATCH elicitation tool: roulette method with 20 chips, bimodal distribution (Morris et al., 2014).

Individual elicitation may be carried out as part of a group exercise where it is impossible to unify the group of experts for discussion due to inability to travel or scheduling difficulties. The scenario where only one expert is available for questioning is most susceptible to biases and inconsistencies, especially if the expert shows overconfidence or if an inexperienced facilitator is not familiar with the nuances of the scientific question. Overconfidence can lead to distributions with a lower variance, compared to those given by a conservative expert or through a

group consensus. Another frequently occurring difficulty can be the availability heuristic or hindsight bias. This is the situation where the expert draws upon the most memorable event relating to θ, which, in reality, may be the most extreme and unlikely. In the case of hindsight bias, it is drawing upon the most recent events that would drive the expert's judgement on θ (Evans, 1988).

6.3 Elicitation of group judgements

When the outcome of an elicitation exercise bears significant weight upon a decision or when multiple experts are competent and willing to take part in the exercise, a group elicitation is preferred. Moreover, if the facilitator would like to gain insight on uncertainty of the entire elicitation procedure, asking for multiple distinct judgements has fewer opportunities for bias than asking one single expert about differing judgements, or repeating the exercise at a different point in time. O'Hagan et al. (2006) provides a review where multiple experts are consulted. This can be summarised from two directions: consensus-seeking and mathematical aggregation. In the first instance, the experts may be gathered together over several hours or days and the facilitator would aim to reach a group consensus about uncertainty of a parameter through open discussion and feedback loops. In the second case, the experts sustain their beliefs and an aggregate distribution function is derived through appropriate weights being allocated to each judgement.

6.3.1 Mathematical aggregation

In a work by French (1983) we can observe a difficulty of distinguishing between the experts and the task of who should assign the weights and in which manner. Also, it would make sense to question at which point in the elicitation exercise the weights are assigned, and whether this would have impact upon the experts' confidence, hence reflecting on the judgements given (or refined). An approach

deemed most democratic to all members of the expert panel was outlined by Stone (1961). Given the scenario of a linear pool for N experts, each is assigned a weight $w_i, i \in [1, N], \sum_1^N w_i = 1$. Then a consensus distribution $f(\theta)$ for θ is obtained through a weighted average of individual distributions as provided by the experts:

$$f(\theta) = \sum_{i=1}^{N} w_i f(\theta_i).$$

Alternatively, a logarithmic weighting can be used, such that for some normalising constant C:

$$f(\theta) \propto C \prod_{i=1}^{N} f(\theta_i)^{w_i}.$$

For example, suppose we have obtained the following judgements from three experts:

$$\begin{aligned} &\text{Expert } 1 : \theta \sim N(0.6, 1), \\ &\text{Expert } 2 : \theta \sim N(5, 2), \\ &\text{Expert } 3 : \theta \sim N(3, 4). \end{aligned} \tag{6.1}$$

Also, the experts have been assigned respective weights: $w_1 = 0.5, w_2 = 0.2, w_3 = 0.3$. A consensus distribution $f(\theta)$ through a linear pooling approach is depicted in Figure 6.2 .

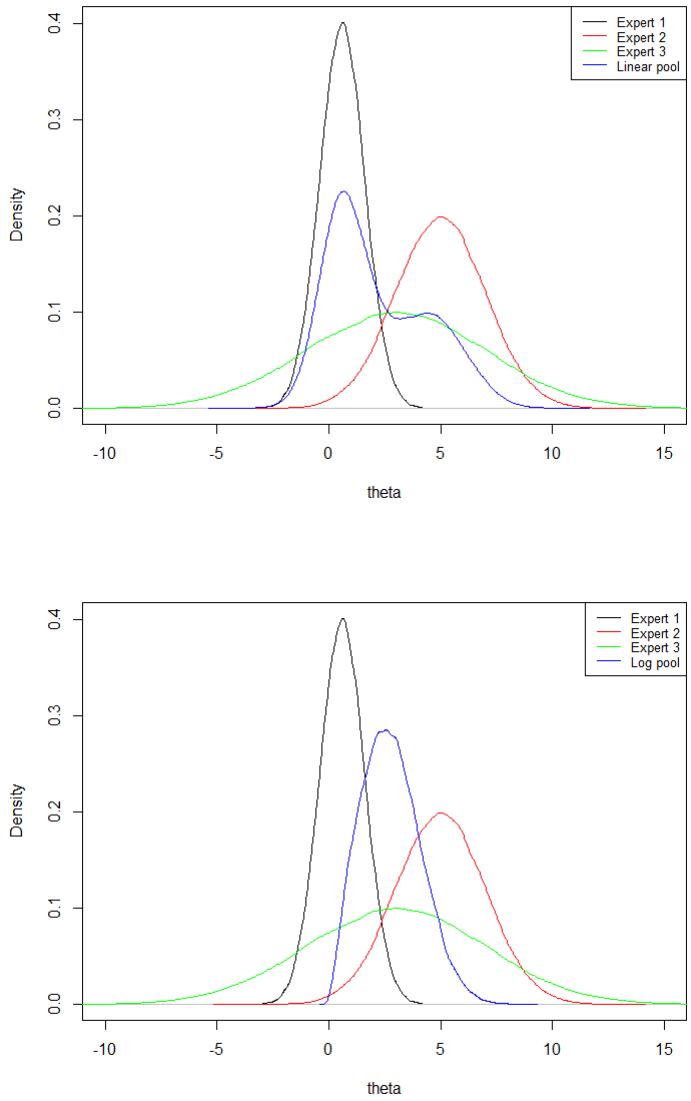


Figure 6.2: Three prior distributions $f_1 = N(0.6, 1), f_2 = N(5, 2), f_3 = N(3, 4)$ with respective weights $w_1 = 0.5, w_2 = 0.2, w_3 = 0.3$. Blue lines in plots show consensus distribution $f(\theta)$ from a linear pool and log-weighted pool for three experts respectively.

As recognised by Dallow et al. (2018), the linear pooling method includes a range of differing opinions, whereas the logarithmic pool seeks to illustrate a compromise distribution. Moreover, assignment of w_i may prove a challenging task. The facilitator could consider expertise of each participant and assign weights through a ranking. Unfortunately, as seen from CoDA, an increase from 0.25 to 0.5 is not the same as an increase of 0.3 to 0.6 as a reflection one person being "twice as experienced" as the other. Therein lies even more cognitive biases, but now from the side of the facilitator. A more rigorous way to ease this task could be through the first training stage of elicitation, and ranking the experts given how well they are able to express their judgements about arbitrary quantities such as distance between cities. However, these seed questions often have little relation to the scientific problem, and the entire pooled analysis could be judged unfair by the decision maker. Finally, expert knowledge can be tied together, as suggested by Lindley (1985) through a multivariate Gaussian distribution that would incorporate experts' differing judgements.

6.3.2 Delphi method

An elicitation procedure that combines individual and group judgement is the Delphi Framework (Brown, 1968). In this protocol, in order to avoid group confrontations and provide equal opportunity for group members to voice their opinions, irrespective of their status and experience, the facilitator takes on the role of a mediator. The members of the expert panel are granted anonymity and express their judgements through the facilitator, who also has an active role in the feedback loop as they communicate judgements and respective reasons between the experts. Each expert is given an opportunity to review their judgements based on the reasoning given by other experts. The aim of Delphi is to maximise the sharing of knowledge between members of the expert panel and to reduce the chance of more conservative experts being stifled by overconfident ones. A distribution is fitted to aggregated results, with equal weightings given to all

participant experts after the final feedback iteration. In an attempt to regularise group discussion to avoid psychological biases, any fruitful ideas arising from initial counter-arguments may be lost, which is an evident drawback of Delphi. However, given that the suggested number of experts per panel has ranged from eight people (Hodgetts, 1977) to hundreds (Hejblum et al., 2008) in a web-based setting, summarising of discussion points by the facilitator, let alone reaching a consensus appears to be a very difficult task to carry out successfully.

6.3.3 Sheffield Elicitation Framework

Sheffield Elicitation Framework (SHELF) developed by Oakley and O'Hagan (2010) is a procedure to reach a judgement about the quantity of interest at group consensus level, alongside a structured feedback and consolidation routine. This consensus-seeking approach is likewise referred to as behavioural aggregation. The procedure is described in Figure 6.3. Alike the process of individual elicitation described earlier, SHELF commences with specification of the task, expert selection and exposing the experts to probabilistic thinking about uncertain quantities. In phase 4 the experts discuss any evidence relevant to the scientific investigation from their expert knowledge and practice. Then the experts provide individual judgements about the unknown quantity or parameter θ in a similar way as discussed previously for one expert, and individual probability distributions are fitted to these judgements. Phase 7 follows on with comparison of the fitted distributions and justification with the use of expert insights and evidence. The fitted probability distribution is then revised in order to reflect this augmented body of evidence, and a feedback loop is carried out until a consensus is reached within the expert cohort.

SHELF has proven to be very popular and reasonably robust given an experienced facilitator. The shortfalls here could be loss of anonymity of the experts, the potential of oppressive group dynamics leading to biased judgements in favour of the more extroverted expert in more senior position outside of the elicitation

setting. In some instances, the Chatham House Rule of confidentiality can be used during the elicitation exercise, in order to facilitate openness in the discussion. Similarly, the exercise is conducted at the pace of the facilitator, who is responsible for accommodating the comfort of several people at once.

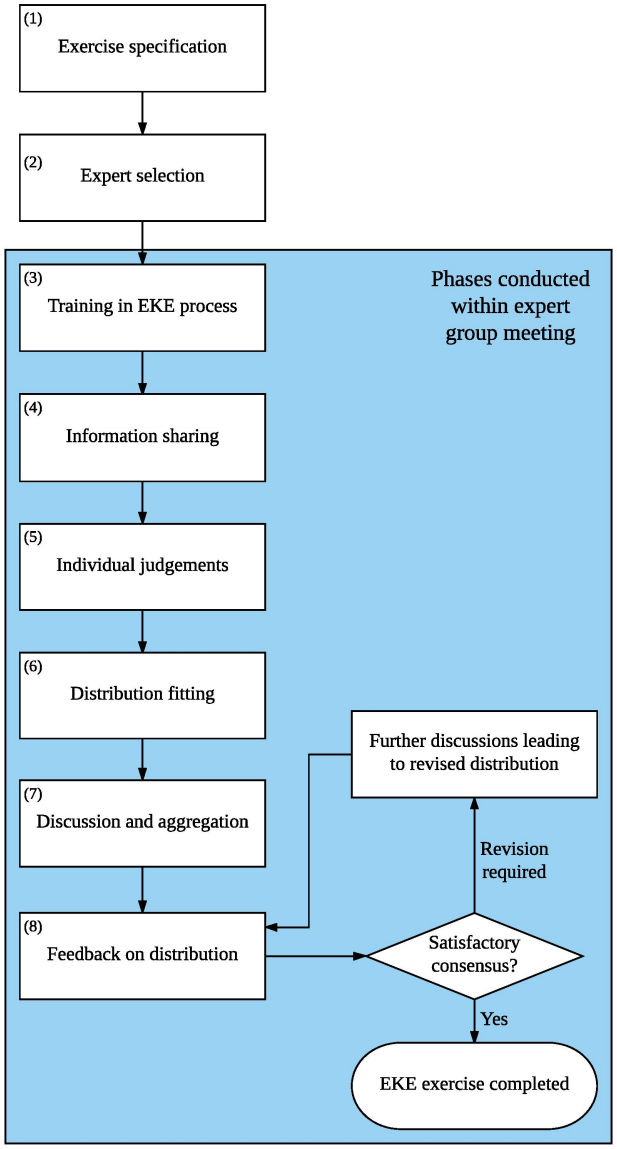


Figure 6.3: SHELF Framework (Oakley and O'Hagan, 2010).

6.3.4 Elicitation of multivariate distributions

When the elicitation exercise seeks insight into about uncertainty of several variables $\boldsymbol{\theta} = (\theta_1, ..., \theta_D)$ is now a vector or matrix. This implies that the questions asked by the facilitator must be more complex and numerous in order to construct a joint (or multivariate) distribution that can express uncertainty about $\boldsymbol{\theta}$. A common technique to address this task is to dissect a multivariate distribution into its marginal and conditional components O'Hagan et al. (2006). If there is evidence to judge variables $\boldsymbol{\theta}$ independent, the joint distribution is expressed as a product of marginal distributions $\theta_i \sim f_i(\cdot)$ for $i = 1, ..., D$. Discussion about independent and dependent variables with the expert is done through consideration of the scientific question of the elicitation exercise; whether or not gaining new information about θ_i would change the expert's beliefs about $\theta_j, i \neq j$. If independence between variables has been established, the facilitator would progress to establish a univariate distribution capturing uncertainty about each θ_i, using techniques previously described.

From a practical viewpoint, Garthwaite et al. (2005) noted that judgements about univariate distributions are conveninent for this application, since the concept of independence is straightforward to understand. This is confirmed by Bar-Hillel (1973), who discovered that the probability of conjunctive events - occurrences that must happen in conjunction with one another - is overestimated, and the opposite trend is seen for disjunctive events, due to the phenomenon of anchoring.

On the other hand, if the expert has reason to judge that there is a dependence relationship between θ_i and $\theta_j, i \neq j$, depending on the scientific question the facilitator may attempt to change the structure of the variables and express the problem through conditional independence (O'Hagan et al., 2006). This may be an attractive direction to pursue when then one of the variables corresponds to some baseline or placebo quantity in a scientific experiment. O'Hagan (1998) provides further details on this approach.

Still, if the dependence structure cannot be overcome in the above manner, it remains for the facilitator to gain insight about the multivariate probability distribution that encompasses the expert's uncertainty in $\boldsymbol{\theta}$. As well as questioning the expert about the marginal quantities $\theta_i, i = 1, ..., D$, there is a need to determine a dependence structure between the variables.

The concept of dependence in an expert elicitation exercise is first established in the training process (Figure 6.3), and then coherently implemented in the stages that follow Information Sharing. The idea of dependence can be communicated to the experts as the association between two quantities such that, when informing uncertainty about one quantity, light is shed on uncertainty about another quantity (or several quantities). As an example, if the distribution of expert beliefs about a quantity X does not change after judgement is given about the distribution of another quantity Y, then these X and Y can be regarded as independent. In the case where this statement cannot hold, we say that there is dependence between X and Y. Lad (1996) notes that this concept of dependence may or need not be shared across the experts, for it to be held valid within one expert's judgement contribution to the elicitation exercise. This is because experts may possess different levels of knowledge, and all this needs to be discussed and validated in the consensus-seeking elicitation framework such as SHELF.

Dependence-modelling can be achieved through an association measure like the Pearson correlation coefficient between the variables, which is an easier cognitive judgement than covariance between random variables, for instance Kadane and Wolfson (1998). Effective elicitation of correlation could be achieved by using graphical aids, such as asking the expert to plot a linear relationship between two variables.

O'Hagan et al. (2006) highlights that for the exercise of multivariate elicitation it is the expert's beliefs in strength of association between pairs of variables that are sought. Similarly, some correlation measures only have interpretation for

repeatable variables (O'Hagan et al., 2006), thus may not be appropriate when gaining insight upon epistemic uncertainty.

6.3.5 Example: Gaussian distribution

In this example we highlight differences in elicitation questioning between elicitation for parameters of the univariate Gaussian distribution and its multivariate analogue.

Let $X \sim N(\mu, \sigma)$ be a univariate random variable following the Gaussian distribution, which serves as a prior for some underlying data. For an uninformative (or flat) prior distribution, it is common to set the parameters $\mu = 0$ and $\sigma = 10^6$ (or of similar order). This yields the Normal distribution centred around zero with (relatively) little concentration around the mean. For the construction of a subjective prior distribution, we require expert judgements about the Gaussian parameters of interest μ and σ. The structure of the univariate Gaussian distribution facilitates the difficulty of asking experts about judgements about the mean (or average), since in this case its median, mode and mean are all equal. However, we are adhering to the strong assumption that the Gaussian distribution is indeed the correct and appropriate distribution for the scientific question under exploration, which, in this example, is abstract. To give some structure to the scientific scenario, let us suppose that we are interested in the body length L (metres) of dogs at an animal rescue centre in Leeds. To elicit judgements around the average body length, we are able to question the experts equally about the mode (the most frequently occurring body length) or the median body length of the animals. In this case, $L \sim N(\mu_L, \sigma_L)$ For the former mode we may pose questions such as

Consider a very large sample of dogs at the shelter. What would be your estimate for the most common length of a dog's body?

Alternatively, questions about the interval of most common animal lengths could

be elicited, and later calibrated with other experts' judgements. To gain insight about the median body length, questions such as

Can you determine a value such that the body length of a typical dog at the shelter is equally likely to be greater than and less than this point?.

To follow, to gain insight upon the spread of the data around μ, we may reflect on the previously elicited location parameters and pose questions like

Suppose that the actual body length is below your stated median - can you determine a new value, such that the actual body length is equally likely to be less than and greater than this point?

this would provide a judgement about the lower quartile of the distribution. Similarly for the upper quartile -

Suppose that the actual body length is above your stated median - can you determine a new value, such that the actual body length is equally likely to be less than and greater than this point?.

When expert opinion is required about two or more unknown variables, the questions to gain insight about marginal and joint distributions become more numerous and complex. We have previously discussed the important notion of independence in this case, and from now on, let us assume that some dependence relationship exists between the two variables. Let us extend the existing scenario to eliciting judgements about two variables: weight W and body length L of dogs at an animal rescue centre in Leeds. Now, a new random variable is introduced $W \sim N(\mu_W, \sigma_W)$. We may gain insight on the animal weight through similar questioning as was done for the body lengths. On the other hand, we may attempt to model the two variables jointly, introducing a dependence structure through $\text{cov}_{W,L}$. Through conditioning on a realisation from L, we ask the expert on an assessment of W, for example -

Given that the length of the dog is more than 1 metre, what is your assessment of the lowest weight it can be?

More advanced questioning methodology, such as direct assessment of correlation between L and W (Garthwaite et al., 2005), can be used to inform the covariance structure. One other approach, where pairwise comparisons are applicable, such as this example, it is feasible to represent association between variables in the form of a regression, which is related to correlation. Garthwaite et al. (2005) suggest eliciting the function $m(l) = E(W|L = l)$ for this scenario, assuming the expert accepts a linear form of association between weight and body length. Then it would suffice to elicit two points, or more than two points for an overfit to assess accuracy of the straight line fit.

Chapter 7

Elicitation methods about a set of proportions

This chapter concerns advancements in the practice of expert elicitation to construct parametric multivariate probability distributions about a set of proportions. Building on general considerations of expert elicitation in Chapter 6, this chapter specialises in exploring situations where several uncertainties with an inherent dependence structure are modelled simultaneously. As in the case of eliciting judgements for a univariate probability distribution, where the experts may be indirectly questioned about the covariance structure in order to justify suitability of previous judgements, expert elicitation in a multivariate domain is a challenging process from the point of view of biases and misunderstanding of the concept of conditional probability by the experts. We overview elicitation protocols and techniques used for this task, and consider them from a practical nuance of elicitation, such as expert fatigue. Another aim of this chapter is to explore elicitation methods to reduce the number of statements provided by the experts for the feasible mathematical construction of a multivariate distribution on the simplex. For this purpose the Dirichlet distribution serves as leading example, but we also provide suggestions for extension to the distributions considered in

Chapter 4, if such do not already exist in the literature. This stems from the challenges currently recognised with increasing number of proportions (the size of the vector $\boldsymbol{\alpha}$ in Dirichlet($\boldsymbol{\alpha}$)) for full specification of the prior distribution where at least $3D$ judgements are required.

7.1 Literature review

O'Hagan et al. (2006) introduces the task of expert elicitation about a set of proportions by first reviewing such an exercise for a single proportion (p) in a Binomial model. The conjugate prior distribution is the Beta, $p \sim \text{Beta}(\alpha, \beta)$, and the elicitation exercise focuses on gaining judgement to determine the parameters α, β. O'Hagan et al. (2006) reviews methods of assessing this prior distribution driven by indirect techniques - the hypothetical future sample method (HFS) and the equivalent prior sample (EPS) as studied by Winkler (1967). Both these methods directly question the expert about the value of p; then through the EPS the expert is asked for the sample size estimate n on which the expert's assessment of p is based. These judgements yield estimates for the Beta distribution parameters by exploiting the binomial moments, thus $\hat{\alpha} = np$ and $\hat{\beta} = n(1-p)$.

Similarly for the HFS the expert is asked to think about the proportion estimate in light of a possible sample from a population of interest, and to update their initial estimate of p. The HFS and the EPS methods do not expose the expert to any direct distributional assessments of the prior density, and instead focus on the sampling distribution.

Winkler (1967) also explores more direct ways of conducting the elicitation exercise by informing the expert of an existence of the (prior) probability density function and the cumulative density function, respectively called the PDF and the CDF methods. For the latter, similar assessments are made as described in Chapter 6 - namely, the expert is asked for an estimate of the median value of p and at least one of the quantiles necessary for unique determination of α, β of

the prior distribution. Given the quantile judgements, a curve is fitted - either the parametric Beta distribution, or a non-parametric depiction. After the fitting process follows the feedback loop, where the expert is presented other features of the prior distribution, such as quantiles not previously assessed.

Likewise, the PDF method by Winkler (1967) aims to build the probability density function, as opposed to being structured through quantile judgements about the distribution of p. For this, the expert is similarly asked about the location of the central probability mass, by means of the mode, for example; followed by a judgements about location of values smaller or greater than the central measure. After a parametric Beta distribution or a non-parametric curve is fitted to these judgements, a similar feedback procedure follows to allow the expert to reflect on adequacy of earlier judgements. Weiler (1965); Duran and Booker (1988); Hughes and Madden (2002) also consider the elicitation exercise from the viewpoint of obtaining some central measure of the distribution mass, followed by quantile or percentile assessments for p, or an interval containing p.

Indirect and direct questioning of the expert may be combined for judgement refinement or consolidation, however Winkler noted that the indirect methods may give probability densities with a higher precision (smaller variance) than direct methods. This could be explained by the expert being conservative or not confident with assessing probability statements. Alternatively, the hypothetical future sample method may give rise to the availability heuristic and mis-reflecting the given sample size to their current knowledge. This effect was also explored by Schaefer and Borcherding (1973) and throughout the years it was found that training the experts in probabilistic assessments before the elicitation exercise has a positive effect in reducing precision of the elicited distributions.

A variant on assessing the spread of p was provided by Pham-Gia and Turkkan (1992) and may be more challenging on the expert cognitively. It asks about expert judgement about the mean (or median) of the Beta distribution, followed

by the average absolute deviation about the mean (median). As with previous methods, only two judgements are required and O'Hagan et al. (2006) states that this number may not be sufficient for identification of unusual judgements or to ensure that the elicitation has been performed robustly. León et al. (2003) bypass assessment of spread of the prior distribution and instead focus on two location judgements - the mean and the mode. Again, as recognised by O'Hagan et al. (2006) this approach is sensitive to individual psychology and extent of probabilistic reasoning. Not only is the procedure extremely sensitive to communication between the facilitator and the expert about differences between the mode and the mean, if those two judgements are very close together the expert may draw upon more extreme cases to 'balance' the mean to some seemingly reasonable distance away from the mode. In estimation of prior parameters α and β such judgements too can pose sensitivity problems. These issues can be also extended to questioning the expert about other location measure of the distribution (median) alongside the mean or the mode, as the only judgements elicited. Above all else, it is important to recognise that each method may be appropriate for a specific application and the given financial or time constraints on the elicitation exercise. Since the work of O'Hagan et al. (2006) some subject-specific comparisons of elicitation techniques with the Beta distribution have been carried out, for example in Grigore et al. (2016), but focus rather on expert-weighing and aggregation approaches, rather than different questioning strategies. O'Hagan et al. (2006) also recognises lack of empirical study to compare aforementioned techniques for the study of elicitation of uncertainty about a single proportion.

In instances where several possible outcomes are possible, as described in Chapter 2 and Chapter 4, the likelihood is the multinomial distribution and the conjugate distribution to convey uncertainty about the parameter that describes multinomial event probabilities $\boldsymbol{p} = (p_1, ..., p_D)$ can be one from the family of Dirichlet distributions. Equally, distributions described in Chapter 4 can quantify uncertainty about compositional data or transition probabilities in a discrete-time

Markov chain. Elicitation in this scenario is again focused on establishing parameters of the prior (Dirichlet) distribution. Bunn (1978) remarked that fractile assessment as carried out for the Beta distribution may be cognitively straining on the experts, and direct elicitation of the Dirichlet parameters is made more complicated by the unit-sum constraint implying more conditional judgements. The HFS approach was deemed preferable by statisticians before the widespread use of computers to aid in the elicitation exercises for the tasks of interactive expert questioning (roulette method) and distribution fitting.

A similar scenario was later addressed by Chaloner and Duncan (1987). This work is a natural extension of the authors' earlier work to elicit uncertainty about a single proportion p, and rests on questioning about the mode of each proportion, assuming that the underlying prior distribution is Dirichlet($\boldsymbol{\alpha}$). Dickey (1983) build on Winkler's work with the hypothetical future sample approach. The expert is questioned about the probability of each event $i = 1, ..., D$ taking place, and from this a mean vector is formed $\hat{\boldsymbol{\mu}}$ to estimate the mean of the Dirichlet distribution. Additional judgement is required about a weight parameter n, which in the Dirichlet distribution is described by the overall concentration parameter $\alpha_0 = \sum_{i=1}^{D} \alpha_i$. The expert is asked for probabilities of all the outcomes of the multinomial distribution, conditional on a hypothetical future sample provided by the facilitator.

Let us illustrate this technique using the eye colour example from Chapter 2. Suppose a simplified scenario where there exist three categories for the eye colour: Blue, Brown and Other. In a HFS the expert may be told that data from five more individuals was collected: one person was recorded to have blue eyes, three people had brown eyes and three people fell in the last category. Conditional on this information, the expert would be asked to provide further statements about probabilities of each of the eye colours being observed in a population. Dickey (1983) then determine the weight parameter n through the Bayes' theorem given the above conditional judgements, and determine how much the updated expert

probabilities differ. However, as recognised by O'Hagan et al. (2006), Dickey (1983) do not provide a structure of selecting the HFS based on which updated expert judgements are given, and this remains an outstanding problem.

Over the last decade there has been increased interest in quantifying expert judgement about a set of proportions, as well as developments to the Dirichlet distribution outlined in Chapter 4 to allow for more flexible modelling. A popular application has been in the medical sciences, especially modelling disease progression - Wilson et al. (2018); Rossi et al. (2019). The latter study bypasses the typical SHELF consensus-seeking approach and uses elicitation on seventeen individual participants either face-to-face or via telephone. Several distributions are fitted using a random search algorithm and employ the Dirichlet, Connor-Mosimann and the modified Connor-Mosimann distributions as chosen distributions. Remote participation and lack of need for the group to come to a consensus allowed the authors to also employ several approaches to elicit judgement: roulette approach, quantile assessment and HFS. It is unclear whether all the experts provided judgements using all three methods, or whether subsets of experts were selected for each method, and then the results aggregated. It was found that the graphical roulette method faced the known limitation that the experts were more focused on the shape of the approximated histogram, and some experts lacked exposure to probabilistic reasoning to confidently state their judgements using the quantile method. In the absence of suitable experts, Gupta and Upadhyay (2019) suggest using past data to inform hyperparameters of the Dirichlet distribution.

Wilson (2017) explored elicitation the modified Connor-Mosimann (mCM) distribution, as defined in Chapter 4. The Connor-Mosimann distribution contains $2(D-1)$ parameters (Connor and Mosimann, 1969), $D-2$ more than the Dirichlet distribution, which allows for increased flexibility and the mCM distribution has $4(D-1)$ parameters. Elicitation for the CM distribution is carried out based on the assumption that the multivariate density can be expressed in terms of

independent Beta-distributed variables defined by Z_j in Chapter 4. The mCM distribution rests on the definition of the scaled Beta distribution rescaled with respect to lower and upper limits. In our case, these would be the $[0, 1]$ interval. Wilson argues against the use of the roulette method to elicit uncertainty about each event probability for a multinomial likelihood, because a separate diagram would be required for each p_i. Instead, a quantile approach is employed for each marginal distribution of p_i, and at least three points are elicited, deemed by O'Hagan et al. (2006) to convey overfitting. It is similarly possible to question the experts in terms of odds or bets, for example, the lower quartile would be the value at which the expert would place a 1:3 bet. Parameter estimates for the distributions are obtained by minimising the sum of squares with respect to the target theoretical equivalent. Wilson compares fit of the three distribution to a proportion of size three and finds that the mCM distribution has the smallest difference with the target distribution, and individual proportions show closest marginal fit to the target, but marginally, the CM distribution has the largest interquartile range.

Zapata-Vázquez et al. (2014) also explore ways to elicit uncertainty through the Dirichlet distribution by incorporating assessment of suitability of the Dirichlet distribution to the question of interest. Zapata-Vázquez et al. (2014) make use of overfitting to reflect expert imprecision in their given judgements, and also to allow for information retention in case the Dirichlet is found to be an unsuitable distribution. SHELF is used as the elicitation protocol of choice. After the experts have been selected and prepared, marginal Beta distributions are elicited for each p_i using existing SHELF software and the quantile approach. Following on from this, Beta distributions are fitted for $p_1, ..., p_D$ and adjusted according to constraints, for example, the means of the marginal distributions have to sum to 1. Then $p_i \sim \text{Beta}(\hat{\alpha}_i, \hat{\beta}_i)$ is assumed to be a reflection of expert judgement that corresponds to the mean constraint. With D Beta distributions obtained, the Dirichlet distribution can be constructed if $\hat{\alpha}_i, \hat{\beta}_i$ are equal for all i. This

scenario is possible but unlikely, and for practicality purposes compromise values of $\hat{\alpha}_i, \hat{\beta}_i$ are needed to be shared across all marginal Beta distributions that have been obtained through the elicitation exercise. Zapata-Vázquez et al. (2014) denote $\hat{\alpha}_i, \hat{\beta}_i = n_i$ which can be thought to contain partial information on the concentration parameter α_0 of the Dirichlet distribution. At most D different n_i values can be obtained for the random variables under question, so a compromise n^* is sought. This compromise value of n_i can lie between the minimum n_i obtained n_{min} and the maximum respectively n_{max}. For instance, n^* can be the midpoint defined as $n_{mid} = (n_{min} + n_{max})/2$, the mean or the median of the set of n_i values. Zapata-Vázquez et al. (2014) further develop an optimal value for n through minimising an objective function through consideration of variance of the Beta distribution. Finally, a conservative approach is suggested, where $n^* = n_{min}$ in the situation where no more knowledge is desired to be expressed about the distribution fit than was already elicited. In this case, the Dirichlet fit is deemed not as an accurate representation of expert judgements, but rather done for convenience and a judgement-refining step is not seemed to provide any more information. When the expert is presented the fitted Dirichlet distribution and if they remain confident in their beliefs even though the Dirichlet distribution does not provide quantile judgements sufficiently close to the expert judgements (this is judged by the expert during the Feedback stage), then the facilitator concludes that the Dirichlet distribution is not suitable. Otherwise, the fitted distribution is accepted, or original judgements are refined.

An alternative approach is proposed by Evans et al. (2017) for eliciting a Dirichlet distribution to convey uncertainty about a set of proportions. Similarly, marginal Beta priors are elicited, however, not through the means of questioning the experts about percentiles, but through determining the most likely value for each category and then constructing an interval, which contains the modal value with "virtual certainty". This expression of confidence can be quantified by a probability of the interval containing the modal value being close to 1. Thus, for

each composition three values are obtained from the expert: the mode ξ, and the interval $[l, u]$ containing ξ with probability $\gamma = 0.99$, as suggested by the authors. Then, estimates for parameters of each Beta distribution are obtained through an iterative bisection algorithm until a solution is below some specified tolerance level. The procedure is extended to the multivariate Dirichlet realm by repeating the above set-up for D compositional random variables, thus $3D$ judgements are obtained from the expert without consideration of γ and the tolerance level. Specifying the lower and upper bounds for each modal value leads the authors to implicitly consider dependence structure between the random variables, and to determine a subset of the simplex, where the resulting Dirichlet density will lie. One restriction placed on this method is that the resulting $\hat{\alpha}$ parameter estimates of the Dirichlet distribution are constrained to $\hat{\alpha}_i > 1$ for $i = 1, ..., D$ in order to avoid singularities near the edges of the simplex. However, in practice and as illustrated later in this chapter, this setting may prove problematic to reflect expert judgement about a very small composition, in which case $\hat{\alpha}_i < 1$ would be more suitable. Despite this method being simple to implement for the experts, the authors highlight that the Dirichlet distribution may still prove restrictive for many applications.

In work reviewed thus far, elicitation has been performed to gain information about marginal prior distributions, which are later combined through reparametrisation or iterative approaches to the multivariate Dirichlet family. For approaches to elicit general multivariate densities see Daneshkhah and Oakley (2010). At the end of Chapter 4 we considered multivariate copula and vine functions, which can allow for independent specification of marginal distribution family and any dependence structure between the random variables. In application to a set of proportions, the marginal distribution and the copula have to adhere to the same constraints on the simplex: the covariance structure must be on the simplex space and the multinomial probabilities are still subject to the sum-one constraint. Elfadaly and Garthwaite (2017) conduct examination of graphical

elicitation methods for a Gaussian copula function. The original compositional probabilities $p_1, ..., p_D$ are parametrised such that

$$\theta_1 = p_1, \theta_D = 1; \text{and } \theta_i = \frac{p_i}{1 - \sum_{j=1}^{i-1} p_j} \text{ for } i = 2, ..., D-1. \tag{7.1}$$

A prior distribution is then assumed to hold over θ_i, and marginally $\theta_i \sim \text{Beta}(a_i, b_i)$. From Section 4.8.2 we have seen that for the construction of a copula we require $G(\cdot)$ and in this case $G(\cdot)$ is the cumulative density function of $\text{Beta}(a_i, b_i)$. Then, the copula is defined through its CDF $\Phi_{D-1,\boldsymbol{R}}(\phi^{-1}[G_1(\theta_1)], ..., \phi^{-1}[G_{D-1}(\theta_{D-1})])$ for a correlation matrix $\boldsymbol{R}$.

From a practical perspective, an elicitation exercise for the copula takes similar form as seen previously. Elfadaly and Garthwaite (2017) employ interactive bar charts for the expert to provide their judgement. For each one of D categories, a judgement is given about the median value. The second step of the exercise is to assume that the median for the first category is treated as an unconditional value. The expert is asked to assume that a hypothetical observation does not fall in the first category and is questioned about the median of the second category given this information. This procedure is repeated iteratively for the hypothetical observations not falling into previous accumulated categories. This step is performed to assess consistency of the expert (Elfadaly and Garthwaite, 2017). Finally, the expert gives lower and upper quartiles for the first (unconditional) category, and repeated iterative judgements are then given for the upper and lower quartiles of the remaining categories, conditional that a hypothetical observation does not fall in the previous $i-1$ categories. In total, for marginal specification of the prior distribution $4(D-1)$ judgements are elicited.

For full specification of the copula, judgements on the correlation matrix $\boldsymbol{R}$ need to be also provided. Elfadaly and Garthwaite (2017) illustrate their method with the Gaussian copula, thus $\boldsymbol{R}$ has to satisfy requirements of the multivariate

normal density, namely that $\boldsymbol{R}$ has to be positive-definite. To achieve this, the authors modify previous work by Kadane et al. (1980) to ensure that elicited values yield a positive-definite $\boldsymbol{R}$. Details of the elicitation structure can be found in the original paper, but the difference with earlier marginal elicited judgements now lies with questioning the expert about conditional assessments. Again, quartile values for category i are assessed conditional that earlier judgements that $p_1, ..., p_{i-1}$ are given by the corresponding assessed median values. Finally, for the category where the conditional median does not approximately equal the assessed median given earlier, the expert is asked to adjust their judgement. In the example given in the original paper for 4 categories it seems that discrepancy of larger than 0.05 warrants an adjustment. Another remark in this approach is that ordering of the categories is important, and ordering deemed most appropriate by the expert needs to be facilitated for easier assessment of conditional judgements. The multinomial quartiles can be then transformed to the Gaussian copula quartiles (Kadane et al., 1980; Elfadaly and Garthwaite, 2017) due to monotonic nature of the Gaussian copula function. The authors deem this transformation a simpler alternative transforming explicit correlation structure, such as the product-moment correlations for Gaussian random variables.

Vine copula functions also introduced in Chapter 4, and their judgement elicitation procedure closely follows that of the copula function. Wilson (2018) illustrates vine copula elicitation for a D-vine. The underlying approach is obtaining judgements to specify a marginal distribution for each compositional variable, an unconditional structure to represent unconditional relationships between the variables, as illustrated in Section 4.8.2, and, finally, conditional relationships between the variables. For specification of the marginal distributions the same method with parametrised θ_i as above (Elfadaly and Garthwaite, 2017) is used. Alternatively, Wilson (2018) calculates exact parameter values of the Beta distributions through eliciting judgements about quantiles. Expert confidence can also be expressed in this way, through specifying a weight w_i (the weights also adhere

to sum-unity) and adjusting the assessed mean and variance by the weights to yield summaries for each θ_i.

Iterative conditional statements are again utilised for the specification of the bivariate copula, assuming some ordering of categories $p_1, ..., p_{D-1}$. The bivariate copula represents the bivariate relationship $\theta_1, \theta_i | \theta_2, ..., \theta_{i-1}$. The expert is asked to suppose that the medians for categories $p_1, ..., p_{i-1}$ hold as those specified in the marginal elicitation step, then is asked to confirm that the median for p_i is still as assessed previously. Then, assessments are made about lower and upper quartiles of p_i, again conditional on previously specified (marginal) medians for $p_1, ..., p_{i-1}$. Given that there is structural dependence between $p_1, ..., p_{i-1}$ and p_i the last step is carried out to reduce the expert's uncertainty about the p_i. Wilson (2018) explores several candidate copula structures, and least-squares method is used to determine the best-fit to the elicited judgements. One possible drawback of the vine approach is that there is an assumption that the marginal distributions are invertible and continuous functions. While holding true with the Beta distribution, these assumptions may need further investigation if used with marginal distributions as specified in Chapter 4.

Werner et al. (2018) adopt non-parametric techniques for elicitation of conditional relationships in the process of constructing a copula function. The authors address ways to reduce cognitive strain on the experts through the use of networks to illustrate conditional relationships between variables. In cases where expert assessment are contradicting to previously given statements or mathematical constraints, Werner et al. (2018) suggest eliciting single conditioning sets of judgements and possible ranges for such judgements are provided by an algorithm. On the contrary, if there is a shortfall of information to uniquely specify a probability density, it is suggested that the final probability densities are modelled as minimally informative.

7.2 Study: elicitation using simplex dissection

In the previous review of elicitation methods to quantify expert uncertainty about a set of proportions, a common direction to conduct the elicitation exercise is to assess judgement about marginal distributions. This approach is systematic, as in most cases it allows for the expert to consider a similar (if not the same) set of judgements for each proportion in turn. Similarly, questioning about more than the minimally required number of points can be done (overfitting) to validate the expert assessments. However, with increasing number of compositions some of the elicitation methods require as many as $4D$ judgements, which can be time-consuming and cognitively challenging for the experts not only in the initial questioning and fitting steps of the elicitation exercise, but also in the feedback and discussion steps. Moreover, in considering marginal distribution assessments, the parameters of fitted (Beta) distributions are transformed in order to adhere to mathematical constraints of the multivariate Dirichlet distribution. This approach may compromise one or more of the original judgements given by the experts, in the extreme to an extent that the Dirichlet (or other target distribution) is deemed an unsuitable fit to describe expert uncertainty. For example, dependency of compositional variables implies that the mode of the Dirichlet distribution does not correspond to the marginal Beta modes, which the expert may feel very strongly about.

Some graphical approaches have been discussed for specification of upper and lower bounds on the modal probability of each compositional category. In order to fit a unique probability distribution, or to locate a set of plausible distributions that fit to a set of expert judgements, the number of judgements required is at least the number of unknown parameters of the prior distribution. As discussed previously, for the Beta distribution with two unknown shape and scale parameters, at least two judgements are sought. This is mathematically coherent, since at least two points are needed to determine a line fit. In this case, this is not

a straight line, but the assumed underlying beta density. To extend this, for specification of a plane in 3 dimensions, three points should be sufficient, given that they are not collinear. Similarly, 4 points are enough to determine a hyperplane, provided that the points do not all lie in the same 2-dimensional plane. Generally, the points require affine independence and this reflects with properties of the simplex Δ^D as explored in Chapter 2. There are further mathematical restrictions, such as the means of the marginal distributions have to sum to 1. Further complexity is seen through constraints on the quartiles: for example, for two categories the lower quartile of category 1 and the upper quartile of category 2 also have to sum to 1. Additional constraints on the quartiles is seen in higher dimensions as well.

In this section we propose and study suitability of a diagrammatic approach, which relies on dissection of the simplex to elicit a multivariate distribution about a set of proportions. This approach is motivated by aiming to reduce the number of judgements elicited from the experts and not rely on transformation of corresponding elicited marginal densities to drive insight on parameters of the Dirichlet distribution.

The partitioning (or simplex-dissection) approach relies on visual representation of the simplex and its decomposition into regions of certain area or volume. The experts could as well be asked to assign weights onto these regions, reconciling confidence in their earlier judgements when presented with marginal distributions at the feedback stage. Through the use of an `RShiny` app, the facilitator would guide the experts through the process of providing cumulative and comparative judgements about a set of proportions.

We aim to fit a three parameter Dirichlet distribution to some quantiles given by an elicitation procedure. These parameters $\boldsymbol{\alpha} \in \mathbb{R}^3_+$ each represent a concentration of the distribution in regions of the simplex for a random variable $\boldsymbol{X}$ with $x_i \in [0, 1]$ and $\sum_{i=1}^{D} x_i = 1$, $i \in [1, D]$. For three proportions the simplex is the

two-dimensional equilateral triangle.

For a scenario with three random variables X, Y, Z constituting a composition, the experts would give a judgements on the following three quantities:

$\mathbb{P}(X < X_1), \mathbb{P}(Y < Y_1)$, and $\mathbb{P}(Z < Z_1)$ for some quantile values

$X_1, Y_1, Z_1 \in (0, 1)$.

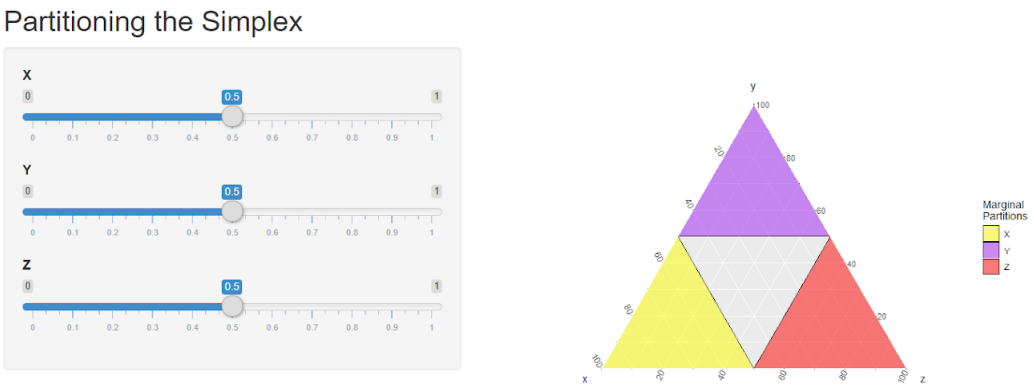


Figure 7.1: `R Shiny` snapshot for simplex dissection based elicitation exercise.

Additionally, two further comparative judgements $\mathbb{P}(X < Y)$ and $\mathbb{P}(Y < Z)$ would be made. Graphically, for the 3-dimensional case, this could be seen as triangle dissection through each vertex onto the middle point of the opposite edge.

We would like to fit these values simultaneously to theoretical Dirichlet quantiles, with a corresponding error. Hence, for three parameters of interest, we obtain 5 judgements. This generalises to $2D - 1$ judgements for D categories or proportions. With the previously-met quartile method there are at least $3D$ judgements needed for each X_i and for $D > 3$ this can become time-consuming and cognitively straining on the experts. From a practical perspective, it may be possible to aggregate categories into bigger classes in order to reduce elicitation exhaustion. However, decomposing the aggregated obtained parameter values into separate instances for the individual categories may not be possible for more

complicated distributions of the Dirichlet family.

The initial fitting procedure is composed of two parts - fitting marginal judgements to each compositional variable, and considering pairs of variables with fitting the Dirichlet using the comparative judgements. Thus, the approach is similar as for marginal quartile judgements, only now solely one judgement is reserved for each marginal distribution. However, instead of seeking separate estimates for parameters of $\text{Beta}(a_i, b_i)$ and re-scaling the parameters at the fitting stage of elicitation, we assume that each compositional random variable is distributed with $\text{Beta}(a_i, a_0 - a_i)$, where a_0 is akin to α_0 the precision parameter of the Dirichlet distribution.

The aim was to explore whether including comparative judgements provided enough information for a unique (or set of plausible) probability distribution on Δ^D. Through simple application of synthesised judgements about three proportions it was found that this specification of the problem lacks identifiability and many optimisation algorithms surveyed did not offer convergence. Surface plots of the loss function (sum of squared errors) can be seen in Appendix B (9.2), however the most distinct feature of these plots are numerous minimal troughs occurring when the estimated $\hat{\alpha}_i$ value is close to 0, which is not reflective of synthetic judgements. Imposing a constraint on $\hat{\alpha}_i$ did not provide improved fit in terms of sensitivity to starting values or the values of parameter constraints, and can also be viewed not suitable in certain practical situations, where an expert judges a proportion to have negligible probability.

Considering the task without comparative judgements and simultaneous fit of D marginal distributions where $X_i \sim \text{Beta}(a_i, a_0 - a_i)$, a_0 as above did not run into difficulties with determining a set of values for a_i. The Newton-Raphson algorithm for multiple roots or the L-BFGS-B algorithm with a lower bound on $a_i > 0$ have been found suitable for lower quartile judgements and in some cases - the median. It was found that the upper quartile judgements did not

reach convergence, and exploring surface plots suggested that this could be due to there being no crossing troughs in the error surface, which do appear for the lower percentile judgements. However, employing a stochastic global optimisation algorithm did yield better results. Accuracy of this fit compared with some existing methods for Dirichlet and Connor-Mosimann elicitation is explored in the final section of this chapter (Section 7.4) through an example.

7.3 Other considerations

In this section we explore further considerations to elicitation of distributions on the simplex as were defined in Chapter 4. Instead of asking the expert to give their uncertainty judgement about compositional proportions, the problem could be viewed from the perspective of asking the expert about ratios of proportions. This is motivated by Aitchison's approach to log-ratio transformations in CoDA. Elfadaly and Garthwaite (2016) consider the additive log-ratio (alr) transformation as the focus of the elicitation procedure, due to its interpretability. As seen in Chapters 2 and 3, the additive log-ratio transformation requires a reference category, which choice is often problem-driven - it can be the most common category, alternatively, a category that carries little importance to the expert (Other class, for example), and can be essentially overlooked in analysis. Nevertheless, since the alr is permutation invariant, the following distribution functions will hold for a different choice of reference category. For compositional parts $\boldsymbol{p} = (p_1, ..., p_D)$ Elfadaly and Garthwaite (2016) refer to p_1 as the reference category, selected for consistency and ease of notation. Then, through the alr transformation they define a variable $Y_i = log(p_i/p_1) = log(p_i/1 - p_2 - ... - p_D)$ for $i = 1, ..., D$. Then, $\boldsymbol{p}$ has the logistic Gaussian distribution, if the transformed variable Y has the multivariate Normal distribution: $Y_{D-1} \sim MVN(\boldsymbol{\mu}_{D-1}, \Sigma_{D-1})$. The distribution assumes that the covariance matrix Σ_{D-1} is non-singular (Elfadaly and Garthwaite, 2016). For the last transformed component Y_D the authors assume that

it follows a multivariate Gaussian distribution also, but the covariance matrix is singular in this case. This is driven by the sum-one contraint on the p variables. In order to bypass the singularity, the authors set that deleting the i^{th} row and i^{th} column of Σ_D for any i would yield a positive-definite covariance matrix. Otherwise, Elfadaly and Garthwaite (2016) rely on conditional properties of the singular Gaussian distribution by replacing the inverse of the covariance matrix by the generalised inverse, which is an alternative parametrisation of the Gaussian distribution (Embrechts, 1983).

Elfadaly and Garthwaite (2017) first proposed a way to elicit hyperparameters $\boldsymbol{\mu}_{D-1}$ through assessment of median values about each p_i. Through the monotonic alr transformation, insight can then be gained about medians of Y_D, and if needed, the judgements are normalised to sum-unity. In a more recent work Elfadaly and Garthwaite (2020) the expert is asked questions about the ratio of p_i/p_1 - relative parts of relative measures. Through similar reasoning and the log-ratio function, each element of the mean vector is given by $\mu_i = \mathbb{E}(Y_i) = \log(m_i^*)$, where m_i^* is the assessed median of ratio p_i/p_1. Hence, for assessing the mean vector for the MVN distribution, $D-1$ assessments are sought.

Elfadaly and Garthwaite's earlier work to elicit judgements about hyperparameters Σ_{D-1} relied again on assessment of upper and lower quartiles about each proportion p_i conditional on given median assessments for the categories that preceed p_i. A final step here is to assess conditional median judgements. The expert is asked to consider that a previous median judgement, say $m_{1,0}$ for p_1 has changed to a new value of $m_{1,1}$, and given this fact, the expert is asked to change their previous medians for $p_2, ..., p_D$ to new values $m_{2,1}, ..., m_{2,D}$. This is motivated by Gaussian quartiles being expressed through the interquartile range (Kadane et al., 1980) for the minimum number of judgements to elicit Σ_{D-1}. In Elfadaly and Garthwaite (2020) ratios of proportions are considered, similar lower and upper quartile judgements are sought instead about p_j/p_1 for $j = 2, ..., D$. After the distributions have been fitted, the feedback stage follows similarly with

presenting the expert the fitted unconditional median and quartile values, which are revised until the expert is satisfied that the fit reflects their judgements.

Relating to the log-ratio Normal distribution, Chapter 4 similarly explored the truncated Gaussian as a plausible distribution to describe uncertainty on the simplex. While no reported applications have been trialled with this distribution for compositional random variables, Albert et al. (2012) consider the truncated Gaussian as one of the candidates for a heirarchical model for interactions between several experts. Similarly, Donovan et al. (2016) consider the truncated normal on [0,1] support as one of the candidate models for the effects of permanent hearing threshold shift on aquatic mammals due to disturbances caused by renewable energy developments.

In Chapter 4 we have seen some interesting geometric properties of the Shadow Dirichlet distribution, which restricts the domain Δ^D, and no literature has thus far addressed this distribution from an elicitation standpoint. An advantage of the Shadow Dirichlet over simply re-normalising the classical Dirichlet over a sub-simplex support is that in the latter case the normalisation term is not analytically tractable, and numerical integration becomes challenging as D increases. Moreover, the Shadow Dirichlet does allow for conjugacy with the multinomial likelihood and for the moments to be expressed in closed form, unlike the re-normalised Dirichlet. A potential difficulty of estimating an additional term in the Shadow Dirichlet - the matrix M - is addressed in Chapter 4, and does not seem an immediate hindrance to an elicitation exercise.

In addressing the (Extended) Flexible Dirichlet distribution in Chapter 4 from the practical elicitation perspective, we need to consider existing methods for elicitation of mixed distributions. Dalal and Hall (1983) and Diaconis and Ylvisaker (1985) explored flexibility obtained through mixtures of general conjugate prior distributions, although no practical applications have been reported, apart from problems concerning variable selection (Garthwaite and Dickey, 1992, 1996). The

Flexible Dirichlet distribution could provide an alternative way to mathematical aggregation of expert judgements by means of a finite mixture of Dirichlet-distributed variables from each expert.

7.4 Application: misclassification of publication ratings

In this example we unite the ideas on stochastic matrices explored in Chapter 5 and expert elicitation. We consider misclassification of publication ratings, where an academic is asked to rate a paper they authored on the discrete scale of 1 to 4, 1 being the lowest rating and 4 the highest. This given rating is then contrasted with a similar rating given by a different academic. A misclassification occurs when the two judgements differ. Let us denote T to be a 4×4 matrix of misclassifications. We shall first carry out the task of eliciting (gaining expert opinion) on the distribution of the information about the chances of misclassification. Matrix T^{data} is the augmented (observed) counts of misclassifications of the ratings, and each row is to be assumed to follow the multinomial distribution. The elements on the diagonal of T^{data} correspond to an agreement between the two judgements.

$$T^{\text{data}} = \begin{pmatrix} 1 & 3 & 0 & 0 \\ 0 & 16 & 4 & 0 \\ 0 & 10 & 18 & 2 \\ 0 & 2 & 10 & 5 \end{pmatrix} \tag{7.2}$$

From the data we can see a central cluster of misclassifications with zero occurrence of "extreme" disagreements between the author and the external academic, for instance, where one rates a paper "4" while the other a "1". In terms of proportions, this can be rewritten as

$$T^{\text{data.prop}} = \begin{pmatrix} 0.25 & 0.75 & 0 & 0 \\ 0 & 0.80 & 0.20 & 0 \\ 0 & 0.33 & 0.60 & 0.066 \\ 0 & 0.012 & 0.59 & 0.29 \end{pmatrix}$$

In this example, we explore fits of the Dirichlet, the Connor-Mosimann and the modified Connor-Mosimann distributions, technical details of which can be found in Chapter 4.

7.4.1 Obtaining the prior

The form of the prior distribution is obtained through the process of expert elicitation. Due to limited resources available at the time of the elicitation, one expert (J. P. Gosling) took part in the study and the facilitator was A. Frantsuzova. The exercise was conducted in coherence with the Sheffield Elicitation Framework. The misclassification matrix consists of four categories and sixteen elements. For example, let us consider the first row of T^{data}: the first element informs that one paper was judged to fall into rating 1 by both the author and the external academic. This represents an agreement. Also, three papers were judged to fall into rating 1 by the authors, yet were rated 2 by the external academic. The second element of the first row can thus express some pessimism or under-confidence by the authors. On the contrary, if we look at the third row of T^{data}, these ratings express a good paper rating as judged by the authors. We see that no external academics judged the paper to fall into category 1; in 10 cases the author thought that their paper would be rated 3, whereas the external academic judged the papers to have a lower rating of 2. The third element of that row depicts that there were 18 cases of agreement between the authors and the external academic. Finally, 2 papers were judged less confidently by the authors themselves, and yet were given the top rating by the external academic. Hence, the final two rows of T^{data} can be seen to express authors' confidence in their

work.

For the task of expert elicitation, we assumed that the rows of T are independent of each other, so it is reasonable to treat each row as a 4-part composition. Since there are sixteen possible author-external ratings, we performed our analysis row-wise; that is we elicit four row-wise independent Dirichlet distributions - one per author rating.

For each row of T quartile judgements about each element were collected from the expert - the lower quartile, the median and the upper quartile. Let us denote T^e_{1r} as the first row of elicited judgements for T, that is, an author judges their paper to have rating 1 and the external academic judges the paper on rating r, $r = 1, 2, 3, 4$.

The following were the quartile judgements elicited during the exercise for row 1 of T:

$$T^{\mathrm{e}}_{11} = (0.12, 0.2, 0.3); \tag{7.3}$$

$$T^{\mathrm{e}}_{12} = (0.5, 0.6, 0.7); \tag{7.4}$$

$$T^{\mathrm{e}}_{13} = (0.15, 0.2, 0.25); \tag{7.5}$$

$$T^{\mathrm{e}}_{14} = (0, 0.0001, 0.005). \tag{7.6}$$

Similar judgements correspond to the remaining rows of T:

$$T^{\mathrm{e}}_{21} = (0.01, 0.02, 0.04); \tag{7.7}$$

$$T^{\mathrm{e}}_{22} = (0.57, 0.7, 0.8); \tag{7.8}$$

$$T^{\mathrm{e}}_{23} = (0.18, 0.24, 0.32); \tag{7.9}$$

$$T^{\mathrm{e}}_{24} = (0.02, 0.03, 0.06); \tag{7.10}$$

$$T^{\mathrm{e}}_{31} = (0, 0.01, 0.02); \tag{7.11}$$

$$T^{\mathrm{e}}_{32} = (0.08, 0.12, 0.18); \tag{7.12}$$

$$T^{\mathrm{e}}_{33} = (0.7, 0.8, 0.87); \tag{7.13}$$

$$T^{\mathrm{e}}_{34} = (0.05, 0.07, 0.1); \tag{7.14}$$

$$T^{\mathrm{e}}_{41} = (0, 0.0001, 0.005); \tag{7.15}$$

$$T^{\mathrm{e}}_{42} = (0.07, 0.1, 0.14); \tag{7.16}$$

$$T^{\mathrm{e}}_{43} = (0.25, 0.3, 0.4); \tag{7.17}$$

$$T^{\mathrm{e}}_{44} = (0.5, 0.6, 0.7). \tag{7.18}$$

7.4.2 Fitting a Dirichlet distribution

Using the SHELF package in R to fit a Dirichlet distribution to the above judgements, the following row-wise $\boldsymbol{\alpha}$ estimates are obtained (2 d.p):

$$\hat{\boldsymbol{\alpha}_1} = (2.78, 7.52, 2.59, 0.15); \tag{7.19}$$

$$\hat{\boldsymbol{\alpha}_2} = (0.35, 8.44, 3.14, 0.52); \tag{7.20}$$

$$\hat{\boldsymbol{\alpha}_3} = (0.23, 2.28, 13.20, 1.31); \tag{7.21}$$

$$\hat{\boldsymbol{\alpha}_4} = (0.18, 1.60, 4.71, 8.78). \tag{7.22}$$

From these values, we can deduce a similar matrix of proportions. Elements of T^{prior} represent expected proportions yielded under the Dirichlet prior distribution with parameter estimates as given in Equations 7.19 to 7.22:

$$T^{\mathrm{prior}} = \begin{pmatrix} 0.21 & 0.58 & 0.20 & 0.01 \\ 0.03 & 0.68 & 0.25 & 0.04 \\ 0.01 & 0.13 & 0.78 & 0.08 \\ 0.01 & 0.10 & 0.31 & 0.58 \end{pmatrix}$$

The standard deviations of a general proportion X_i are given by

$$\mathrm{sd}(X_i) = \sqrt{\frac{\hat{\alpha}_i(-\hat{\alpha}_i + \sum_{j=1}^{4} \hat{\alpha}_j)}{(\sum_{j=1}^{4} \hat{\alpha}_j)^2(1 + \sum_{j=1}^{4} \hat{\alpha}_j)}},$$

which, for us, yields

$$T^{\text{prior.sd}} = \begin{pmatrix} 0.11 & 0.13 & 0.11 & 0.03 \\ 0.04 & 0.13 & 0.12 & 0.05 \\ 0.03 & 0.08 & 0.1 & 0.06 \\ 0.03 & 0.08 & 0.11 & 0.12 \end{pmatrix}$$

The subjective judgement of the expert in this case gives a matrix of proportions quite similar to the observed values in T^{data} (the expert was not exposed to the data in the process of elicitation), with a similar central cluster and very low probability of occurrence of an extreme disagreement. When questioned separately about his overall uncertainty in any of the misclassifications, it was stated that judgements in row 3 have least of his confidence, partly due to the potential presence of bias when an author rates their own work. However, this was not reflected in the overall variance computed, which indicates row 1 to be the most uncertain. The biggest difference to be noted from these judgements is the last row corresponding to the highest rating of 4. Through the elicitation process we obtained a higher proportion of 4-to-4 ratings than shown by the data, and a ten-fold difference between the expected proportion of an author ranking their paper as 4 and the external academic giving a classification of 2. The expert also judged there to be a similar proportion between 1-to-1 and 1-to-3 ratings, with a greater uncertainty around the former. No dependency structure has been incorporated into eliciting the parameters, which is not wholly representative of the problem at hand. In this study, we are assuming that there is no dependence between the individual rows (authors' own ratings), and that the data-driving

process is homogeneous. The expert was informed that his median values have to sum up to 1.

Furthermore, we elicited quartile (lower quartile, median and upper quartile) beliefs about overall proportions of author and external ratings:

$$T^{\mathrm{e}}_{1,\mathrm{auth}} = (0.003, 0.004, 0.005); \tag{7.23}$$

$$T^{\mathrm{e}}_{2,\mathrm{auth}} = (0.23, 0.27, 0.31); \tag{7.24}$$

$$T^{\mathrm{e}}_{3,\mathrm{auth}} = (0.35, 0.40, 0.47); \tag{7.25}$$

$$T^{\mathrm{e}}_{4,\mathrm{auth}} = (0.21, 0.27, 0.31); \tag{7.26}$$

$$T^{\mathrm{e}}_{1,\mathrm{ext}} = (0, 0.0001, 0.0002); \tag{7.27}$$

$$T^{\mathrm{e}}_{2,\mathrm{ext}} = (0.25, 0.31, 0.36); \tag{7.28}$$

$$T^{\mathrm{e}}_{3,\mathrm{ext}} = (0.45, 0.49, 0.56); \tag{7.29}$$

$$T^{\mathrm{e}}_{4,\mathrm{ext}} = (0.19, 0.24, 0.28). \tag{7.30}$$

The above judgements yielded the following $\boldsymbol{\alpha}$ estimates for the Dirichlet distribution for the overall proportions of author and external ratings:

$$\boldsymbol{\alpha}^{\mathrm{auth}} = (1.77, 11.2, 16.9, 11.1) \tag{7.31}$$

$$P^{\mathrm{auth}} = (0.04, 0.27, 0.41, 0.27) \tag{7.32}$$

$$\boldsymbol{\alpha}^{ext} = (0.005, 10.7, 17.4, 8.33) \tag{7.33}$$

$$P^{ext} = (0.0001, 0.29, 0.48, 0.23) \tag{7.34}$$

This part of elicitation was carried out before extracting judgements between pairs of ratings and was done to rectify any inconsistencies. This can be verified by observing that

$$D = P^{\mathrm{ext}} - P^{\mathrm{auth}} T^{\mathrm{prior}}, \tag{7.35}$$

and this stands at $D = (-0.026, -0.00023, -0.0032, 0.029)$ to 2 significant figures.

	Z_1	Z_2	Z_3
$\hat{a}_1^{\text{CM}}$	1.75	17.19	10.18
$\hat{b}_1^{\text{CM}}$	6.43	5.79	0
$\hat{a}_2^{\text{CM}}$	0.62	5.94	20
$\hat{b}_2^{\text{CM}}$	19.97	2.54	3.17
$\hat{a}_3^{\text{CM}}$	0.26	1.59	20
$\hat{b}_3^{\text{CM}}$	18.96	10.10	1.93
$\hat{a}_4^{\text{CM}}$	0	2.42	4.06
$\hat{b}_4^{\text{CM}}$	20	20	7.66

Table 7.1: Estimated parameters of the Connor-Mosimann distribution.

7.4.3 Fitting a Connor-Mosimann distribution

In this application we also explored the more flexible Connor-Mosimann distribution and the related modified Connor Mosimann distribution. Technical details on these distributions can be found in Chapter 4. The fitting procedure followed Wilson (2017) for both distributions using the modcmfitr package in R. Marginal distributions of the modified Connor-Mosimann distribution follow the Scaled Beta distribution with four parameters: shape and scale parameters a and b and two parameters A and B that re-scale the standard Beta distribution. Using judgements as given in judgements (7.3) to (7.18), we again employed row-wise fitting of the paper ratings.

For the Connor-Mosimann distribution, the following parameter estimates for row 1 of T are illustrated by $\hat{a}_1^{\text{CM}}$ and $\hat{b}_1^{\text{CM}}$, and similar notation is taken for the other rows 2, 3 and 4 of T.

The above parameter estimates of the Connor-Mosimann and the modified Connor-Mosimann distributions are given to two decimal places, unless higher precision is required to illustrate that an upper bound is not a true zero, for example. Fitting was carried out with 100,000 iterations of the algorithm and 5,000 searches - the function default is 100, and Wilson suggests increasing this parameter. In considering the modified Connor Mosimann fits, it appeared that the parameter values have been restrained to take the maximum value of 10 as a possible way to regulate the algorithm, yet this may not always be a true reflection of

	Z_1	Z_2	Z_3
$\hat{a}_1^{\mathrm{mCM}}$	5.38	8.20	0.14
$\hat{b}_1^{\mathrm{mCM}}$	20	2.82	0
L	0.00001	00001	0.99
U	0.99	0.99	1
$\hat{a}_2^{\mathrm{mCM}}$	0.05	16.40	20
$\hat{b}_2^{\mathrm{mCM}}$	0.33	6.62	0
L	0.00001	0.00001	0
U	0.86	0.99	0.89
$\hat{a}_3^{\mathrm{mCM}}$	0	0.93	20
$\hat{b}_3^{\mathrm{mCM}}$	18.47	7.24	3.20
L	0.0031	0.039	0.34
U	0.41	1	0.99
$\hat{a}_4^{\mathrm{mCM}}$	0	0.0003	10.67
$\hat{b}_4^{\mathrm{mCM}}$	0.80	0.0073	19.98
L	0.00004	0.094	0.000093
U	0.094	0.48	0.99

Table 7.2: Estimated parameters of the modified Connor-Mosimann distribution.

the underlying distribution. Alternatively, the Connor-Mosimann likelihood surface may bear resemblance to the one met in investigating the simplex-partition elicitation method at the beginning of this chapter, and the algorithm finds satisfactory solutions close to, or at, its starting values. Adjustements were made by A.Frantsuzova to the default upper limit on a^{CM} and b^{CM} from 10, as originally set by Wilson, to 20. This change was carried out in order explore convergence near these boundaries and sensitivity to algorithm runs. Overall, increasing the upper bound on a^{CM} and b^{CM} gave fewer instances of convergence at exactly the upper limits.

In fitting the Connor-Mosimann distribution, we are able to estimate Beta distribution quantiles to compare with quantile judgements elicited from the expert. The elicited distributions yielded quantile values within 16% of those elicited from the expert in (7.3) to (7.18). In majority of instances, the estimated quantiles were identical to the expert judgement's precision level. For the modified Connor Mosimann distribution, elicited distributions yielded quantile values within 13% of those elicited from the expert in (7.3) to (7.18).

7.4.4 Fitting a Dirichlet distribution using simplex dissection

Using judgements obtained in (7.3) to (7.18) we can explore a Dirichlet distribution fit using the method as described in Section 7.2. Due to the smaller number of judgements required for this fitting procedure, we are faced with selecting a scenario - fitting using the lower quartiles (LQ) for each misclassification, the medians or the upper quartiles (UQ). Results for parameter estimates are shown in Table 7.3. The optimisation procedure used for finding parameter estimates was simulated annealing due to (Bélisle, 1992).

Parameter	LQ fit	Median fit	UQ fit
$\hat{\boldsymbol{\alpha}}_1$	(7.4, 0.02, 6.9, 8.4)	(3.9, 0.08, 0.01, 5.2)	(1.7, 0.5, 1.6, 20.3)
$\hat{\boldsymbol{\alpha}}_2$	(5.5, 0.06, 4.8, 4.7)	(3.1, 0.01, 0.2, 3.7)	(16.7, 0.3, 2.3, 0.2)
$\hat{\boldsymbol{\alpha}}_3$	(7.3, 6.5, 0.2, 6.4)	(4.9, 0.1, 0.009, 4.2)	(20.4, 1.3, 0.1, 0.7)
$\hat{\boldsymbol{\alpha}}_4$	(8.4, 7.9, 6.8, 0.4)	(2.1, 1.7, 0.2, 0.1)	(9.8, 0.07, 0.1, 0.8)

Table 7.3: Parameter estimates using simplex dissection fit.

Let us now compare the fit of distributions obtained using the simplex-partition method, contrasted with the SHELF approach. In the former cases the marginal distribution of the Dirichlet is the Beta distribution. Below plots depict uncertainty about elements of the first row of T - if we solely consider the scenario where an author considers their work to be of rating 1, and this is contrasted to the rating given by a different academic. Similar plots for the other ratings can be found in Appendix B (9.2).

For the simplex-partition method, we have considered results yielded by judgements given from both considerations of the lower quartile judgements, the median and the upper quartile judgements. From the plots it can be seen that the LQ partition approach exhibits similar variance to the results yielded from SHELF. In the last scenario where the expert gave very small probability judgements for T_{14}^{e}, the simplex partition method did not reflect this well, and the two methods give distinctly different marginal distributions. From similar plots

in Appendix B, we can observe that distinction between the two methods arises when one of the $\boldsymbol{\alpha}$ estimates is less than one. This again reflects on earlier nuances in this thesis, which are driven by Dirichlet parameter values close to the boundary of the simplex. This emphasises importance of the feedback loop in an elicitation protocol, in order for the expert to refine their judgements, if they feel that the resulting distribution is not suitable.

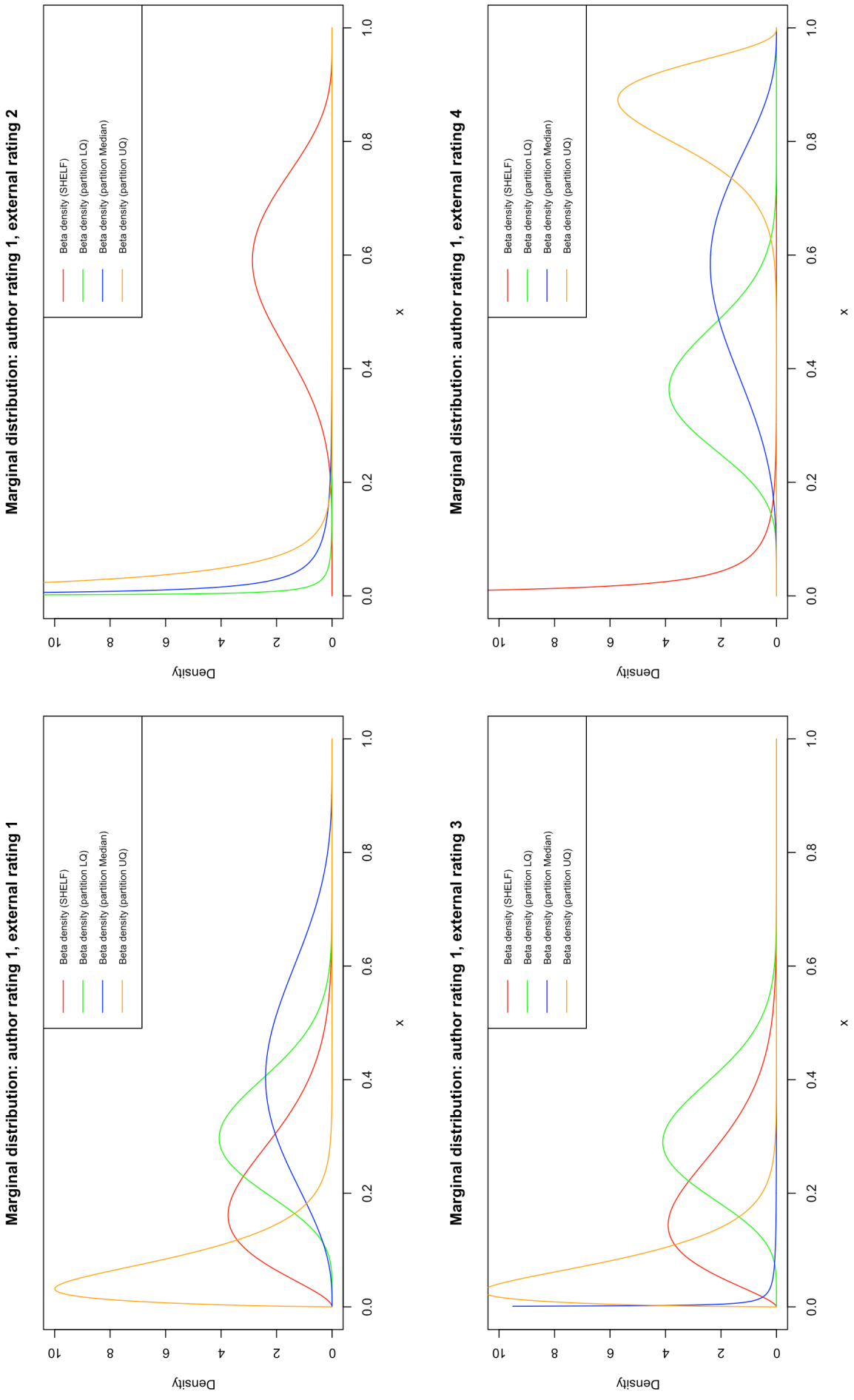


Figure 7.2: Marginal Beta distribution plots of ratings misclassifications.

7.4.5 Discussion

In this section we have explored several techniques to fit a distribution over the simplex to a set of judgements about misclassifications of ratings. For the SHELF approach, the expert was required to provide at least 48 judgements for a 4-dimensional distribution. The method of Evans et al. (2017) was also approached by providing a lower bound for each element of T, however, this was often not consistent with expert judgement on the upper bound and those suggested by the software.

All throughout structuring the problem and the elicitation exercise, we have assumed a somewhat simplified scenario with no inter-row dependence. Similarly, we have not considered measures of confidence (or modesty) of each author, which could have been elicited from the expert for further work. However, here arises a question whether this judgement should be elicited from the authors themselves. The expert gave an informal judgement that the probability that an author's own rating is lower than that given by a different academic - the expert believes there to be a 3 out of 4 chance that there is an agreement in rating between the two parties. This is an overestimate of any prior approximations (diagonal elements of T^{prior}) to this overall proportion.

Conjugacy of the Dirichlet distribution allows us to easily consider posterior uncertainty about ratings misclassifications after data are observed. We can also compare the elicited distributions with non-informative flat distributions on the simplex, as those described in Chapter 4. When united with a non-informative prior, posterior proportions are, as expected, driven almost wholly by the data and do not incorporate any knowledge of the practical setting. In both cases of the non-informative prior and the elicited distributions, row 1 had the greatest overall proportion variance, followed by row 2. Row 3 had the smallest such variance. Under these priors, a transition 1-to-1 or 1-to-4 ratings is equally likely a-priori, which may not have be a reasonable assumption. Under this scenario,

out of 100 ratings, we may expect nine to be judged 1 by the author and 4 by a different academic; whereas, under the subjective prior, we would expect fewer than one such disagreement on average. The uninformative priors, hence, inflate the extreme cases of over-confidence or modesty on the authors' behalf.

Chapter 8

Discussion

8.1 Conclusions

In this thesis, we have studied modern approaches to compositional data analysis and Bayesian prior elicitation. We then united the two fields and investigated approaches to elicitation of uncertainty for random variables lying on the simplex space.

In the first part of the thesis, we gave attention to compositional data analysis and concentrated on multivariate regression in the simplex. We illustrated parametric and non-parametric regression methods during sub-pixel modelling of tree species distributions in collaboration with Fera Science, UK. Regression was carried out from the angle of compositional data analysis with the use of log-ratio transformations, as well as the Box-Cox power transformation in a multivariate regression setting. Furthermore, we employed a non-parametric random forest fitting strategy to contrast predictive performance with existing parametric approaches. As well as demonstrating the importance for inclusion of spatial variables in the regression problem, where available, this work highlighted current and prevalent issues in compositional data analysis – that of modelling essential zeros in a compositional data set, and approaches to dimensionality reduction. In

relation with the latter, we also implemented the technique of principal variables by Cumming and Wooff (2007) on the set of explanatory spectral band values, and this drew out the variables which drove the most variability in the set of predictors. Using the reduced set of variables, prediction error remained comparable to that of the original set-up, while a reduction in computational time was achieved.

The thesis followed with a thorough review of parametric probability distributions used for modelling a set of proportions summing to unity. We considered recent developments of the classical Dirichlet distribution and any remaining inflexibilities of this distribution for modelling compositional data. The issues still prevalent are accommodation of essential zeros, as well as a restrictive covariance structure, although this is addressed by the more flexible Connor-Mosimann distribution, and other variants of the Dirichlet, such as the Extended Flexible Dirichlet distribution. There also exist distributions constructed for specific application purposes, for example, when compositional parts fall into natural groupings, or modelling hierarchical relationships, or only specific subsets of the simplex. In many instances, these distributions do not have probability density functions or moments expressed in closed form, and so rely on costly E-M algorithms for fitting.

Further distributions from the Gaussian and the Kent family were also explored, alongside their application to modelling compositional data either on the simplex, or after proportional data have been appropriately transformed. Additional consideration of modelling covariance structure between compositional parts was explored through the Dirichlet-tree distribution and copulae functions. The latter have found recent popularity in expert elicitation exercises to yield a more flexible covariance structure than that offered by the Dirichlet distribution. Following on from general compositional data, we focused on modelling uncertainty about a discrete-time Markov transition matrix, in particular stationary behaviour of a Markov chain. Illustrated using the classical Dirichlet distribution, long-term

behaviour of uncertainty following distributions explored in Chapter 4 is easily modelled, provided that a random sample from the distribution is obtainable. For the Dirichlet scenario, we observed that the form of the stationary distribution is highly unstable if row-wise uncertainty is expressed with parameters $\boldsymbol{\alpha} < 1$. This can be crucial in applications, where Bayesian posterior is driven strongly by the prior distribution in presence of little evidence. Stepping away from right-stochastic matrices, we gave further consideration of current modelling approaches used for contingency tables and found that application is still heavily reliant on the Dirichlet distribution, although the dependency structure is more flexible due to relaxation of sum-unity constraints on rows of contingency tables.

Next, we advanced to the task of expert elicitation for specification of a prior distribution over the simplex support. It was found that present methods are focused on fitting the Dirichlet distribution, the (modified) Connor-Mosimann and a copula distribution, when seen in light of describing uncertainty about a set of proportions. Methods are either based on eliciting marginal Beta distributions, followed by reparametrisation of elicited parameters to satisfy distributional constraints, or introducing the expert to hypothetical future samples. In the former approach, judgements about the location and spread of the distributions are made, often in the form of quartile judgements, or a modal value alongside tail percentiles. Efforts are made to ease cognitive fatigue using software, for example, interactive bar charts. Some implementations in the medical literature utilised a hybrid approach, a choice often driven by practical nuances of expert elicitation, such as availability of experts and their training in probabilistic reasoning.

We investigated an approach that bypasses eliciting judgements about individual compositional parts and aim to fit a joint distribution using one set of judgements. The motivation for this is that we wish to reduce the number of judgements required from the experts to avoid fitting distributions that contradict original statements, and to also address expert fatigue and duration of the elicitation exercise. This issue becomes especially prevalent with the more flexible

developments of the Dirichlet distribution, some of which can require at least $4D$ judgements. We proposed a fitting procedure where the simplex is partitioned, and the experts are asked to assign probabilities to these parts of the simplex. Similarly, comparative judgements between pairs of compositional variables could be elicited, in order to guide a covariance structure for a more complex distribution on the simplex. The application at the end of this thesis united ideas explored with right-stochastic matrices and expert elicitation. We conducted an elicitation exercise to convey uncertainty about elements of a 4×4 matrix of ratings misclassifications. Comparison was made between an existing SHELF procedure for the Dirichlet distribution, the (modified) Connor-Mosimann distributions and the simplex-partition method proposed previously. While the two Dirichlet approaches proved stable, the SHELF method exhibited closer judgement fit than the simplex-partition method, and the latter also did not reflect the expert judgement where probabilities were deemed very small.

8.2 Future directions

In the first chapters of this thesis, we met many currently-tackled problems of modelling compositional data – occurrence of essential zeros and a restrictive covariance structure implied by the classical Dirichlet distribution. In modelling tree species distribution, we have demonstrated that lower resolution satellite data can be effectively used for prediction of spectral bands in mixed pixels. Further work in this area could be aimed at extending the temporal and geographical scale of the application to confirm transferability of methods explored, or to highlight previously unseen relationships between spectral bands.

From a statistical viewpoint, it would be interesting to extend the method of principal variables to achieve dimensionality reduction in compositional data. This would be a contribution to existing approaches to principal component analysis on compositional parts (Aitchison, 1986). An outstanding question, however,

is the quantification of partial correlation between compositional parts – some very recent exploration has been conducted by Erb (2020) that relies on alr or clr-transformed compositions. Modelling inter-correlation between compositional parts also has attractive application in modelling uncertainty about Markov matrices, and may allow us to drop the row-independence assumption, which find direct application with doubly stochastic matrices currently sparsely considered in applied statistics.

From the standpoint of expert elicitation, the simplex-dissection approach for modelling a joint multivariate distribution requires further refinement. Even though solutions fitted to expert judgements by a stochastic global optimisation routine are stable, some parameter estimates appear may be far from reflecting expert judgement. These conclusions are based on comparison of marginal fits with the SHELF approach. Of course, one may ask whether comparing marginal fits is the only plausible way of judging suitability of an elicited distribution to expert judgement. To advance the front-end part of the simplex-dissection method, thought needs to be given to how the simplex is depicted in dimensions $D > 3$. When $D = 4$ the expert would be presented with a tetrahedron, and this is the last dimension that is visually plausible. Perhaps, pairwise plots of compositional parts with a third part being the augmented parts that remain (Other category, as in examples throughout this thesis), would allow us to maintain a simple graphical interface. However, which pairwise compositional variables should then be selected is a separate question, but this could be incorporated into earlier stages of the elicitation exercise.

A more general question in expert elicitation fitting and feedback stage can be considering what is more important – the scale of parameter estimates fitted, or whether the resulting distribution is deemed to reflect the expert judgements well. In the former case, if we take an example of the Dirichlet, a fitted distribution with very large estimates for $\boldsymbol{\alpha}$ may be deemed by the expert to fit as well as another Dirichlet distribution with smaller $\boldsymbol{\alpha}$ estimates. However, the latter

would have lower variance and any further Bayesian analysis would be heavily prior-driven. Perhaps, as discussed in Oakley and O'Hagan (2004); Gosling (2008), incorporating uncertainty in expert judgements themselves (where, if J is the expert judgement, questioning the expert if judgement $J \pm \epsilon$ is equally likely) can be a way forward to choose between potential candidate distributions.

Bibliography

Aitchison, J. (1982), 'The statistical analysis of compositional data', *Journal of the Royal Statistical Society: Series B (Methodological)* **44**(2), 139–160.

Aitchison, J. and Kay, J. W. (2003), 'Possible solution of some essential zero problems in compositional data analysis'.

Aitchison, J. and Shen, S. M. (1980), 'Logistic-normal distributions: Some properties and uses', *Biometrika* **67**(2), 261–272.

Aitchison, J. W. (1986), *The Statistical Analysis of Compositional Data.*, Springer Netherlands.

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K. and Rousseau, J. (2012), 'Combining expert opinions in prior elicitation', *Bayesian Analysis* **7**(3), 503–532.

Alsuroji, R. (2018), Multidimensional Proportional Data Clustering Using Shifted-Scaled Dirichlet Model, PhD thesis, Concordia University.

Amemiya, T. (1973), 'Regression analysis when the dependent variable is truncated normal', *Econometrica: Journal of the Econometric Society* pp. 997–1016.

Andersen, M. R., Winther, O. and Hansen, L. K. (2014), 'Bayesian inference for structured spike and slab priors', *Advances in Neural Information Processing Systems* **2**, 1745–1753.

Ascari, R., Migliorati, S. and Ongaro, A. (2017), 'The extended flexible dirichlet model: a simulation study', *Applied Stochastic Models and Data Analysis (ASMDA)* .

Azzalini, A. and Valle, A. D. (1996), 'The multivariate skew-normal distribution', *Biometrika* **83**(4), 715–726.

Baddeley, M. C., Curtis, A. and Wood, R. (2004), 'An introduction to prior information derived from probabilistic judgements: elicitation of knowledge, cognitive bias and herding', *Geological Society, London, Special Publications* **239**(1), 15–27.

Bar-Hillel, M. (1973), 'On the subjective probability of compound events', *Organizational behavior and human performance* **9**(3), 396–406.

Bdiri, T. and Bouguila, N. (2011), An infinite mixture of inverted dirichlet dis-

tributions, *in* 'International Conference on Neural Information Processing', Springer, pp. 71–78.

Bdiri, T., Bouguila, N. and Ziou, D. (2014), 'Object clustering and recognition using multi-finite mixtures for semantic classes and hierarchy modeling', *Expert systems with applications* **41**(4), 1218–1235.

Bedford, T. and Cooke, R. M. (2002), 'Vines–a new graphical model for dependent random variables', *The Annals of Statistics* **30**(4), 1031 – 1068.

Bedford, T., Daneshkhah, A. and Wilson, K. J. (2016), 'Approximate uncertainty modeling in risk analysis with vine copulas', *Risk Analysis* **36**(4), 792–815.

Bélisle, C. J. (1992), 'Convergence theorems for a class of simulated annealing algorithms on d', *Journal of Applied Probability* **29**(4), 885–895.

Bolger, F. (2018), *The selection of experts for (probabilistic) expert knowledge elicitation. In Elicitation (pp. 393-443).*

Bourouis, S., Alharbi, A. and Bouguila, N. (2021), 'Bayesian learning of shifted-scaled dirichlet mixture models and its application to early covid-19 detection in chest x-ray images', *Journal of Imaging* **7**(1), 7.

Box, G. E. P. and Cox, D. R. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society. Series B (Methodological)* **26**(2), 211–252.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**, 123–140.

Breiman, L. (2001), Random forests, *in* 'Machine Learning', pp. 5–32.

Brown, B. B. (1968), Delphi process: a methodology used for the elicitation of opinions of experts, Technical report, Rand Corp Santa Monica CA.

Bunn, D. W. (1978), 'Estimation of a dirichlet prior distribution', *Omega* **6**(4), 371–373.

Butler, A. and Glasbey, C. (2008), 'A latent gaussian model for compositional data with zeros', *Journal of the Royal Statistical Society Series C* **57**, 505–520.

Caelen, O. (2017), 'A bayesian interpretation of the confusion matrix', *Annals of Mathematics and Artificial Intelligence* **81**(3), 429–450.

Cappellini, V., Sommers, H.-J., Bruzda, W. and Życzkowski, K. (2009), 'Random bistochastic matrices', *Journal of Physics A: Mathematical and Theoretical* **42**(36), 365209.

Caron, R. M., Li, X., Mikusiński, P., Sherwood, H. and Taylor, M. D. (1996), 'Nonsquare "doubly stochastic" matrices', *Lecture Notes-Monograph Series* pp. 65–75.

Cartinhour, J. (1990), 'One-dimensional marginal density functions of a truncated multivariate normal density function', *Communications in Statistics-Theory and Methods* **19**(1), 197–203.

Carvalho, C., Polson, N. and Scott, J. (2012), 'Handling sparsity via the horseshoe', *Journal of Machine Learning Research* **5**, 73–80.

Chaloner, K. and Duncan, G. T. (1987), 'Some properties of the dirichlet-multinomial distribution and its use in prior elicitation', *Communications in Statistics-Theory and Methods* **16**(2), 511–523.

Chayes, F. (1960), 'On correlation between variables of constant sum', *Journal of Geophysical Research* **65**.

Clemen, R. T. and Reilly, T. (1999), 'Correlations and copulas for decision and risk analysis', *Management Science* **45**(2), 208–224.

Connor, R. and Mosimann, J. (1969), 'Concepts of independence for proportions with a generalization of the dirichlet distribution', *Journal of The American Statistical Association* **64**, 194–206.

Cumming, J. and Wooff, D. A. (2007), 'Dimension reduction via principal variables', *Computational Statistics Data Analysis* **52**, 550–565.

Dalal, S. R. and Hall, W. J. (1983), 'Approximating priors by mixtures of natural conjugate priors', *Journal of the Royal Statistical Society: Series B (Methodological)* **45**(2), 278–286.

Dallow, N., Best, N. and Montague, T. H. (2018), 'Better decision making in drug development through adoption of formal prior elicitation', *Pharmaceutical statistics* **17**(4), 301–316.

Daneshkhah, A. and Oakley, J. E. (2010), 'Eliciting multivariate probability distributions', *Rethinking risk measurement and reporting* **1**, 23.

Darroch, J. N. (1969), 'Null correlation for proportions', *Journal of the International Association for Mathematical Geology* **1**(2), 221–227.

Darroch, J. N. and Ratcliff, D. (1978), 'No-association of proportions', *Journal of the International Association for Mathematical Geology* **10**(4), 361–368.

DEFRA (2018*a*), 'A green future: our 25 year plan to improve the environment', *HM Government London* .

DEFRA (2018*b*), 'Tree health resilience strategy'.

Dennis III, S. Y. (1991), 'On the hyper-dirichlet type 1 and hyper-liouville distributions', *Communications in Statistics-Theory and Methods* **20**(12), 4069–4081.

Diaconis, P. and Ylvisaker, D. (1985), 'Quantifying prior opinion. in bayesian statistics 2 (eds j. m. bernardo, m. h. degroot, d. v. lindley, a. f. m. smith', *Proceedings of the Second Valencia International Meeting* .

Dickey, J. M. (1968), 'Three multidimensional-integral identities with bayesian applications', *The Annals of Mathematical Statistics* pp. 1615–1628.

Dickey, J. M. (1983), 'Multiple hypergeometric functions: Probabilistic interpretations and statistical uses', *Journal of the American Statistical Association* **78**(383), 628–637.

Dobigeon, N. and Tourneret, J.-Y. (2007), 'Truncated multivariate gaussian distribution on a simplex.', *Technical report 2007a, University of Toulouse.* .

Donovan, C., Harwood, J., King, S., Booth, C., Caneco, B. and Walker, C. (2016), Expert elicitation methods in quantifying the consequences of acoustic disturbance from offshore renewable energy developments, *in* 'The Effects of Noise on Aquatic Life II', Springer, pp. 231–237.

Duran, B. S. and Booker, J. M. (1988), 'A bayes sensitivity analysis when using the beta distribution as a prior', *IEEE transactions on reliability* **37**(2), 239–247.

Dymarski, P. (2011), *Hidden Markov models: Theory and applications*, BoD–Books on Demand.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barcelo-Vidal, C. (2003), 'Isometric logratio transformations for compositional data analysis', *Mathematical geology* **35**(3), 279–300.

Egozcue, J. and Pawlowsky-Glahn, V. (2008), Compositional data and simpson's paradox.

Elfadaly, F. G. and Garthwaite, P. H. (2016), 'On eliciting logistic normal priors for multinomial models', *preparation. http://mcs-brains. open. ac. uk/elicitation/Logistic% 20Paper. pdf* .

Elfadaly, F. G. and Garthwaite, P. H. (2017), 'Eliciting dirichlet and gaussian copula prior distributions for multinomial models', *Statistics and Computing* **27**(2), 449–467.

Elfadaly, F. G. and Garthwaite, P. H. (2020), 'On quantifying expert opinion about multinomial models that contain covariates', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **183**(3), 959–981.

Embrechts, P. (1983), 'A property of the generalized inverse gaussian distribution with some applications', *Journal of Applied Probability* **20**(3), 537–544.

Erb, I. (2020), 'Partial correlations in compositional data analysis', *Applied Computing and Geosciences* **6**, 100026.

Evans, J. S. B. T. (1988), 'The knowledge elicitation problem: a psychological perspective', *Behaviour & Information Technology* **7**(2), 111–130.

Evans, M., Guttman, I. and Li, P. (2017), 'Prior elicitation, assessment and inference with a dirichlet prior', *Entropy* **19**(10), 564.

Ezbakhe, F. and Pérez Foguet, A. (2019), Wash your data off: navigating statistical uncertainty in compositional data analysis, *in* 'Proceedings of the 8th International Workshop on Compositional Data Analysis (CoDaWork2019): Terrassa, 3-8 June, 2019', pp. 57–62.

Ferrers, N. M. (1886), *An Elementary Treatise on Trilinear Coordinates.*, London: Macmillan.

Fiedler, M. (2011), *Matrices and graphs in geometry*, number 139, Cambridge University Press.

Fix, E. and Hodges, J. L. (1951), 'Discriminatory analysis, non parametric discrimination: Consistency properties'.

French, S. (1983), *Group consensus probability distributions: A critical survey*, University of Manchester. Department of Decision Theory.

Frigyik, B., Gupta, M. and Chen, Y. (2010), 'Shadow dirichlet for restricted probability modelling', *Advances in Neural Information Processing Systems* **23**, 613–621.

Fry, J. M., Fry, T. R. L. and McLaren, K. R. (2000), 'Compositional data analysis and zeros in micro data', *Applied Economics* **32**(8), 953–959.

Frye Jr, R. M. (2012), 'Use of expert elicitation at the us nuclear regulatory commission', *Alb. LJ Sci. & Tech.* **23**, 309.

Garthwaite, P. H. and Dickey, J. M. (1992), 'Elicitation of prior distributions for variable-selection problems in regression', *The Annals of Statistics* **20**(4), 1697–1719.

Garthwaite, P. H. and Dickey, J. M. (1996), 'Quantifying and using expert opinion for variable-selection problems in regression', *Chemometrics and Intelligent Laboratory Systems* **35**(1), 1–26.

Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005), 'Statistical methods for eliciting probability distributions', *Journal of the American Statistical Association* **100**(470), 680–701.

Gavrilova, T. and Andreeva, T. (2012), 'Knowledge elicitation techniques in a knowledge management context.', *Journal of Knowledge Management* **16**(4), 523–537.

Gislason, P. O., Benediktsson, J. A. and Sveinsson, J. R. (2006), 'Random forests for land cover classification', *Pattern Recognition Letters* **27**(4), 294–300.

Goldstein, D. G. and Rothschild, D. (2014), 'Lay understanding of probability distributions.', *Judgment & Decision Making* **9**(1).

Goldstein, M. (2006), 'Subjective bayesian analysis: Principles and practice.', *Bayesian Analysis* **1**(3), 403–420.

Good, I. J. and Mittal, Y. (1987), 'The Amalgamation and Geometry of Two-by-Two Contingency Tables', *The Annals of Statistics* **15**(2), 694 – 711.

Gosling, J. P. (2008), 'Elicitation: a nonparametric view'.

Granger, M. M. (2014), 'Use (and abuse) of expert elicitation in support of decision making for public policy.', *Proceedings of the National Academy of Sciences* **111**(20), 7176–7184.

Greenacre, M. (2002), 'Ratio maps and correspondence analysis'.

Greenacre, M. (2018), *Compositional Data Analysis in Practice.*, Chapman and Hall/CRC.

Grigore, B., Peters, J., Hyde, C. and Stein, K. (2016), 'A comparison of two methods for expert elicitation in health technology assessments', *BMC medical research methodology* **16**(1), 1–11.

Grunsky, E., Kjarsgaard, B., Egozcue, J. J., Pawlowsky-Glahn, V. and Henestrosa, S. (2008), 'Studies in stoichiometry with compositional data'.

Guillotte, S. and Perron, F. (2012), 'Bayesian estimation of a bivariate copula using the jeffreys prior', *Bernoulli* **18**(2), 496–519.

Gupta, A. and Upadhyay, S. K. (2019), 'Subjective elicitation of dirichlet hyperparameters using past data: A study of ovarian cancer patients', *Austrian Journal of Statistics* **48**(3), 1–14.

Hejblum, G., Ioos, V., Vibert, J.-F.and Böelle, P.-Y., Chalumeau-Lemoine, L., Chouaid, C., Valleron, A.-J. and Guidet, B. (2008), 'A web-based delphi study on the indications of chest radiographs for patients in icus', *Chest* **133**(5), 1107–1112.

Heremans, S., Bossyns, B., Eerens, H. and Orshoven, J. V. (2011), 'Efficient collection of training data for sub-pixel land cover classification using neural networks.', *International Journal of Applied Earth Observation and Geoinformation* **13**, 657–667.

Hijazi, R. H. and Jernigan, R. W. (2009), 'Modelling compositional data using dirichlet regression models', *Journal of Applied Probability & Statistics* **4**(1), 77–91.

Hodgetts, R. M. (1977), 'Applying the delphi technique to management gaming', *Simulation* **29**(1), 209–212.

Huang, H., Tosun, A. B., Guo, J., Chen, C., Wang, W., Ozolek, J. A. and Rohde, G. K. (2014), 'Cancer diagnosis by nuclear morphometry using spatial information', *Pattern recognition letters* **42**, 115–121.

Hughes, G. and Madden, L. V. (2002), 'Some methods for eliciting expert knowledge of plant disease epidemics and their application in cluster sampling for disease incidence', *Crop Protection* **21**(3), 203–215.

Huguenin, R. L., Karaska, M. A., Van Blaricom, D. and Jensen, J. R. (1997), 'Subpixel classification of bald cypress and tupelo gum trees in thematic mapper imagery', *Photogrammetric Engineering and Remote Sensing* **63**(6), 717–724.

Iyengar, M. and Dey, D. (1998), 'Box–cox transformations in bayesian analysis of compositional data', *Environmetrics* **9**, 657–671.

Johnson, N. L. and Kotz, S. (1970), 'Continuous univariate distributions i'.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S. and Peters, S. C. (1980), 'Interactive elicitation of opinion for a normal linear model', *Journal of the American Statistical Association* **75**(372), 845–854.

Kadane, J. and Wolfson, L. J. (1998), 'Experiences in elicitation: [read before the royal statistical society at a meeting on'elicitation 'on wednesday, april 16th,

1997, the president, professor afm smith in the chair]', *Journal of the Royal Statistical Society: Series D (The Statistician)* **47**(1), 3–19.

Katsis, A., Martzoukos, S. and Yannacopoulos, A. (2003), 'Expert opinion elicitation in option pricing: A bayesian approach', *Tech. Report 03–06, HERMES Center of Excellence on Computational Finance & Economics* .

Kent, J. T. (1982), 'The fisher-bingham distribution on the sphere', *Journal of the Royal Statistical Society: Series B (Methodological)* **44**(1), 71–80.

Kieschnick, R. and McCullough, B. D. (2003), 'Regression analysis of variates observed on (0, 1): percentages, proportions and fractions', *Statistical modelling* **3**(3), 193–213.

Kork, J. O. (1977), 'Examination of the chayes-kruskal procedure for testing correlations between proportions', *Journal of the International Association for Mathematical Geology* **9**(6), 543–562.

Kurowicka, D. and Joe, H., eds (2010), *Dependence Modeling:Vine Copula Handbook*, World Scientific Publishing Co. Pte. Ltd.

Lad, F. (1996), *Operational subjective statistical methods: A mathematical, philosophical, and historical introduction*, Vol. 315, Wiley-Interscience.

Lamb, R., Aspinall, W., Odbert, H. and Wagener, T. (2017), 'Vulnerability of bridges to scour: insights from an international expert elicitation workshop', *Natural Hazards and Earth System Sciences* **17**(8), 1393–1409.

Lancaster, H. O. (1965), 'The helmert matrices', *American Mathematical Monthly* **72**.

Lee, L.-F. (1979), 'On the first and second moments of the truncated multi-normal distribution and a simple estimator', *Economics Letters* **3**(2), 165–169.

Leininger, T. J., Gelfand, A. E., Allen, J. M. and Silander, J. A. (2013), 'Spatial regression modeling for compositional data with many zeros.', *Journal of Agricultural, Biological, and Environmental Statistics* **18**(3), 314–334.

León, C. J., Vázquez-Polo, F. J. and González, R. L. (2003), 'Elicitation of expert opinion in benefit transfer of environmental goods', *Environmental and Resource Economics* **26**(2), 199–210.

Li, H. (2015), 'Microbiome, metagenomics, and high-dimensional compositional data analysis', *Annual Review of Statistics and Its Application* **2**, 73–94.

Lindley, D. V. (1964), 'The bayesian analysis of contingency tables', *The Annals of Mathematical Statistics* pp. 1622–1643.

Lindley, D. V. (1985), 'Reconciliation of discrete probability distributions', *Bayesian statistics* **2**(375-390), 375–390.

List of Latin Botanical Tree Names, Genus and Species (2020).
URL: *https://www.treenames.net/ti/*

Liu, Y., Chen, J. and Shan, C. (2014), Dirichlet-tree distribution enhanced random forests for facial feature detection, *in* '2014 IEEE International Conference on Image Processing (ICIP)', IEEE, pp. 234–238.

Louck, J. D. (1997), 'Doubly stochastic matrices in quantum mechanics', *Foundations of Physics* **27**(8), 1085–1104.

Lu, D. and Weng, Q. (2007), 'A survey of image classification methods and techniques for improving classification performance.', *International Journal of Remote Sensing* **28**(5).

Maier, M. J. (2014), 'Dirichletreg: Dirichlet regression for compositional data in r'.

Mamon, R. S. and Elliott, R. J. (2007), *Hidden Markov models in finance*, Vol. 4, Springer.

Mao, J. and Ma, L. (2020), 'Dirichlet-tree multinomial mixtures for clustering microbiome compositions', *arXiv preprint arXiv:2008.00400* .

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Barceló-Vidal, C. (2005), 'The additive logistic skew-normal distribution on the simplex', *Stochastic Environmental Research and Risk Assessment* **19**(3), 205–214.

Mateu-Figueras, G. and Tolosana-Delgado, R. (2006), 'Using the dirichlet distribution to model geochemical data'.

Matheron, G. (1963), '"principles of geostatistics"', *Economic Geology* **58**, 1246–1266.

McAlister, D. (1879), 'Xiii. the law of the geometric mean', *Proceedings of the Royal Society of London* **29**(196-199), 367–376.

Miesch, A. T. (1969), The constant sum problem in geochemistry, *in* 'Computer applications in the earth sciences', Springer, pp. 161–176.

Minka, T. (1999), 'The dirichlet-tree distribution', *http://research. microsoft. com/˜ minka/papers/dirichlet/minka-dirtree. pdf* .

Monti, G. S., Mateu i Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J. (2011), 'The shifted-scaled dirichlet distribution in the simplex'.

Moody, J., Will, R. and Blanton, J. (1996), 'Enhancing knowledge elicitation using cognitive interview', *Expert Systems with Applications* **10**, 127–133.

Morgan, M. G. (2014), 'Use (and abuse) of expert elicitation in support of decision making for public policy', *Proceedings of the National academy of Sciences* **111**(20), 7176–7184.

Morris, D. E., Oakley, J. E. and Crowe, J. A. (2014), 'A web-based tool for eliciting probability distributions from experts', *Environmental Modelling & Software* **52**, 1–4.

Mosimann, J. E. (1962), 'On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions', *Biometrika* **49**(1/2), 65–82.

Nallapati, R., Ahmed, A., Cohen, W. and Xing, E. (2007), Sparse word graphs: A scalable algorithm for capturing word correlations in topic models, *in* 'Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)', IEEE, pp. 343–348.

Nallapati, R., Minka, T. and Robertson, S. (2007), The smoothed-dirichlet distribution: A new building block for generative topic models.
URL: *https://www.microsoft.com/en-us/research/publication/smoothed-dirichlet-distribution-new-building-block-generative-topic-models/*

Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.

Ng, K., Tian, G.-L. and Tang, M.-L. (2011), 'Dirichlet and related distributions: Theory, methods and applications', **888**.

Ng, K. W., Tang, M.-L., Tan, M. and Tian, G.-L. (2008), 'Grouped dirichlet distribution: A new tool for incomplete categorical data analysis', *Journal of Multivariate Analysis* **99**(3), 490–509.

Ng, K. W., Tang, M.-L., Tian, G.-L. and Tan, M. (2009), 'The nested dirichlet distribution and incomplete categorical data analysis', *Statistica Sinica* pp. 251–271.

Oakley, J. E. and O'Hagan, A. (2004), 'Probabilistic sensitivity analysis of complex models: a bayesian approach', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**(3), 751–769.

Oakley, J. E. and O'Hagan, A. (2010), 'Shelf: the sheffield elicitation framework (version 2.0)', *School of Mathematics and Statistics, University of Sheffield, UK (http://tonyohagan. co. uk/shelf)* .

O'Hagan, A. (1998), 'Eliciting expert beliefs in substantial practical applications: [read before the royal statistical society at ameeting on'elicitation 'on wednesday, april 16th, 1997, the president, professor afm smithin the chair]', *Journal of the Royal Statistical Society: Series D (The Statistician)* **47**(1), 21–35.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and T., R. (2006), *Uncertain Judgements: Eliciting Experts' Probabilities. Statistics in Practice.*, Wiley.

Ongaro, A. and Migliorati, S. (2013), 'A generalization of the dirichlet distribution', *Journal of Multivariate Analysis* **114**, 412–426.

Ongaro, A. and Migliorati, S. (2014), A dirichlet mixture model for compositions allowing for dependence on the size, *in* 'Advances in Latent Variables', Springer, pp. 101–111.

Ortego, M. I., M. I. and Egozcue, J. J. (2013), Spurious copulas, *in* 'Proceedings of the 5th International Workshop on Compositional Data Analysis (CoDaWork 2013), June 3-7, 2013, Vorau, Austria', pp. 123–130.

Ottosen, T.-B., Petch, G., Hanson, M. and Skjøth, C. A. (2020), 'Tree cover mapping based on sentinel-2 images demonstrate high thematic accuracy in

europe', *International Journal of Applied Earth Observation and Geoinformation* **84**, 101947.

O'Hagan, A. (2019*a*), 'Expert knowledge elicitation: Subjective but scientific.', *The American Statistician* **73**, 69–81.

O'Hagan, A. (2019*b*), 'Expert knowledge elicitation: subjective but scientific', *The American Statistician* **73**(sup1), 69–81.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2011), 'Exploring compositional data with the coda-dendrogram', *AUSTRIAN JOURNAL OF STATISTICS Volume* **40**, 103–113.

Pearson, K. (1897), 'Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurement of organs.', *Proceedings of the Royal Society,* (60), 489–498.

Pelletier, C., Valero, S., Inglada, J., Champion, N. and Dedieu, G. (2016), 'Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas', *Remote Sensing of Environment* **187**, 156–168.

Pham-Gia, T. and Turkkan, N. (1992), 'Sample size determination in bayesian analysis', *Journal of the Royal Statistical Society: Series D (The Statistician)* **41**(4), 389–397.

Rayens, W. S. and Srinivasan, C. (1994), 'Dependence properties of generalized liouville distributions on the simplex', *Journal of the American Statistical Association* **89**(428), 1465–1470.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M. and Rigol-Sanchez, J. P. (2012), 'An assessment of the effectiveness of a random forest classifier for land-cover classification', *ISPRS Journal of Photogrammetry and Remote Sensing* **67**, 93–104.

Rodriguez, J. C. (2007), 'Measuring financial contagion: A copula approach', *Journal of Empirical Finance* **14**(3), 401–423.

Rossi, S. H., Blick, C., Nathan, P., Nicol, D., Stewart, G. D. and Wilson, E. C. F. (2019), 'Expert elicitation to inform a cost-effectiveness analysis of screening for renal cancer', *Value in Health* **22**(9), 981–987.

Ročková, V. and George, E. (2018), 'The spike-and-slab lasso', *Journal of the American Statistical Association* **113**, 0.

Rubin (1981), *The Bayesian Bootstrap*.

Salter-Townshend, M. and Haslett, J. (2006), Modelling zero inflation of compositional data.

Scealy, J. and Welsh, A. (2011), 'Regression for compositional data by using distributions defined on the hypersphere.', *Journal of the Royal Statistical Society, Series B, Statistical Methodology,* **73**, 351–375.

Scealy, J. and Welsh, A. (2014), 'Fitting kent models to compositional data with small concentration', *Statistics and Computing* **24**.

Scealy, J. and Wood, A. (2020), 'Score matching for compositional distributions'. **URL:** *https://arxiv.org/abs/2012.12461*

Schaefer, R. E. and Borcherding, K. (1973), 'The assessment of subjective probability distributions: A training experiment', *Acta Psychologica* **37**(2), 117–129.

Shanteau, J. (1992*a*), 'Competence in experts: The role of task characteristics', *Organizational behavior and human decision processes* **53**(2), 252–266.

Shanteau, J. (1992*b*), The psychology of experts an alternative view, *in* 'Expertise and decision support', Springer, pp. 11–23.

Sharp, W. (2006), 'The graph median–a stable alternative measure of central tendency for compositional data sets', *Mathematical Geology* **38**, 221–229.

Sinkhorn, R. (1964), 'A relationship between arbitrary positive matrices and doubly stochastic matrices', *The annals of mathematical statistics* **35**(2), 876–879.

Sklar, A. (1973), 'Random variables, joint distribution functions, and copulas', *Kybernetika* **9**(6), 449–460.

Small, C. (1997), 'Multidimensional medians arising from geodesics on graphs', *The Annals of Statistics* **25**.

Soares, M. and Bojke, L. (2018), *Expert Elicitation to Inform Health Technology Assessment*, pp. 479–494.

Sood, V. and Gupta, S. (2018), A comparative review on different sub-pixel classification algorithms.

Stephens, M. (1982), 'Use of the von mises distribution to analyse continuous proportions', *Biometrika* **69**.

Stewart, C. and Field, C. (2011), 'Managing the essential zeros in quantitative fatty acid signature analysis', *Journal of Agricultural Biological and Environmental Statistics* **16**, 45–69.

Stone, M. (1961), 'The opinion pool', *The Annals of Mathematical Statistics* pp. 1339–1342.

Takahashi, Y. (1969), Markov chains with random transition matrices, *in* 'Kodai Mathematical Seminar Reports', Vol. 21, Department of Mathematics, Tokyo Institute of Technology, pp. 426–447.

Tallis, G. M. (1961), 'The moment generating function of the truncated multinormal distribution', *Journal of the Royal Statistical Society: Series B (Methodological)* **23**(1), 223–229.

Tallis, G. M. (1963), 'Elliptical and radial truncation in normal populations', *The Annals of Mathematical Statistics* **34**(3), 940–944.

Tallis, G. M. (1965), 'Plane truncation in normal populations', *Journal of the Royal Statistical Society: Series B (Methodological)* **27**(2), 301–307.

Tiao, G. G. and Cuttman, I. (1965), 'The inverted dirichlet distribution with applications', *Journal of the American Statistical Association* **60**(311), 793–805.

Tsagris, M. (2015), 'Regression analysis with compositional data containing zero values', *Chilean journal of statistics* **6**, 47–57.

Tsagris, M., Alenazi, A. and Stewart, C. (2021), 'Non-parametric regression models for compositional data'.
URL: *https://arxiv.org/pdf/2002.05137.pdf*

Tsagris, M., Preston, S. and Wood, A. (2011), 'A data-based power transformation for compositional data', *Proceedings of the 4th international workshop on Compositional Data Analysis, Girona, Spain, May 2011* .

Tsagris, M. and Stewart, C. (2020), 'A folded model for compositional data analysis', *Australian & New Zealand Journal of Statistics* **62**(2), 249–277.

Tu, K. (2016), Modified dirichlet distribution: Allowing negative parameters to induce stronger sparsity, *in* 'Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing', pp. 1986–1991.

Verbeiren, S., Eerens, H., Piccard, I., Bauwens, I. and Orshoven, J. V. (2008), 'Sub-pixel classification of spot-vegetation time series for the assessment of regional crop areas in belgium.', *International Journal of Applied Earth Observation and Geoinformation* **10**, 486–497.

Wang, H., Meng, J. and Tenenhaus, M. (2010), *Regression Modelling Analysis on Compositional Data*, pp. 381–406.

Wang, X., Nie, F. and Huang, H. (2016), Structured doubly stochastic matrix for graph based clustering: Structured doubly stochastic matrix, *in* 'Proceedings of the 22nd ACM SIGKDD International conference on Knowledge discovery and data mining', pp. 1245–1254.

Weiler, H. (1965), 'The use of incomplete beta functions for prior distributions in binomial sampling', *Technometrics* **7**(3), 335–347.

Werner, C., Bedford, T. and Quigley, J. (2018), 'Sequential refined partitioning for probabilistic dependence assessment', *Risk Analysis* **38**(12), 2683–2702.

Wilson, E. C. F. (2017), 'Fitting a modified connor-mosimann distribution to elicited quantiles of multinomial probabilities'.

Wilson, E. C. F., Usher-Smith, J. A., Emery, J., Corrie, P. G. and Walter, F. M. (2018), 'Expert elicitation of multinomial probabilities for decision-analytic modeling: an application to rates of disease progression in undiagnosed and untreated melanoma', *Value in Health* **21**(6), 669–676.

Wilson, K. J. (2018), 'Specification of informative prior distributions for multinomial models using vine copulas', *Bayesian Analysis* **13**(3), 749–766.

Winkler, R. L. (1967), 'The assessment of prior distributions in bayesian analysis', *Journal of the American Statistical association* **62**(319), 776–800.

Yule, G. (1903), 'Notes on the theory of association of attributes in statistics', *Biometrika* **2**, 121–134.

Zapata-Vázquez, R. E., O'Hagan, A. and Soares Bastos, L. (2014), 'Eliciting expert judgements about a set of proportions', *Journal of Applied Statistics* **41**(9), 1919–1933.

Zass, R. and Shashua, A. (2006), 'Doubly stochastic normalization for spectral clustering', *Advances in neural information processing systems* **19**, 1569–1576.

Zellner, A. (1977), 'Maximal data information prior distributions', *New developments in the applications of Bayesian methods* pp. 211–232.

Zellner, A. (1996), 'Models, prior information, and bayesian analysis', *Journal of Econometrics* **75**(1), 51–68.

Zellner, A. and Min, C.-K. (1992), *Bayesian analysis, model selection and prediction*, University of Chicago, Graduate School of Business, Department of Economics.

Zhang, M. and Shi, W. (2020), 'Systematic comparison of five machine-learning methods in classification and interpolation of soil particle size fractions using different transformed data'.

Zucchini, W. and MacDonald, I. L. (2009), *Hidden Markov models for time series: an introduction using R*, Chapman and Hall/CRC.

Chapter 9

Appendices

9.1 Appendix A

Figures 9.1 through to 9.12 present diagnostic plots for models considered for compositional tree species regression approaches in Chapter 3.

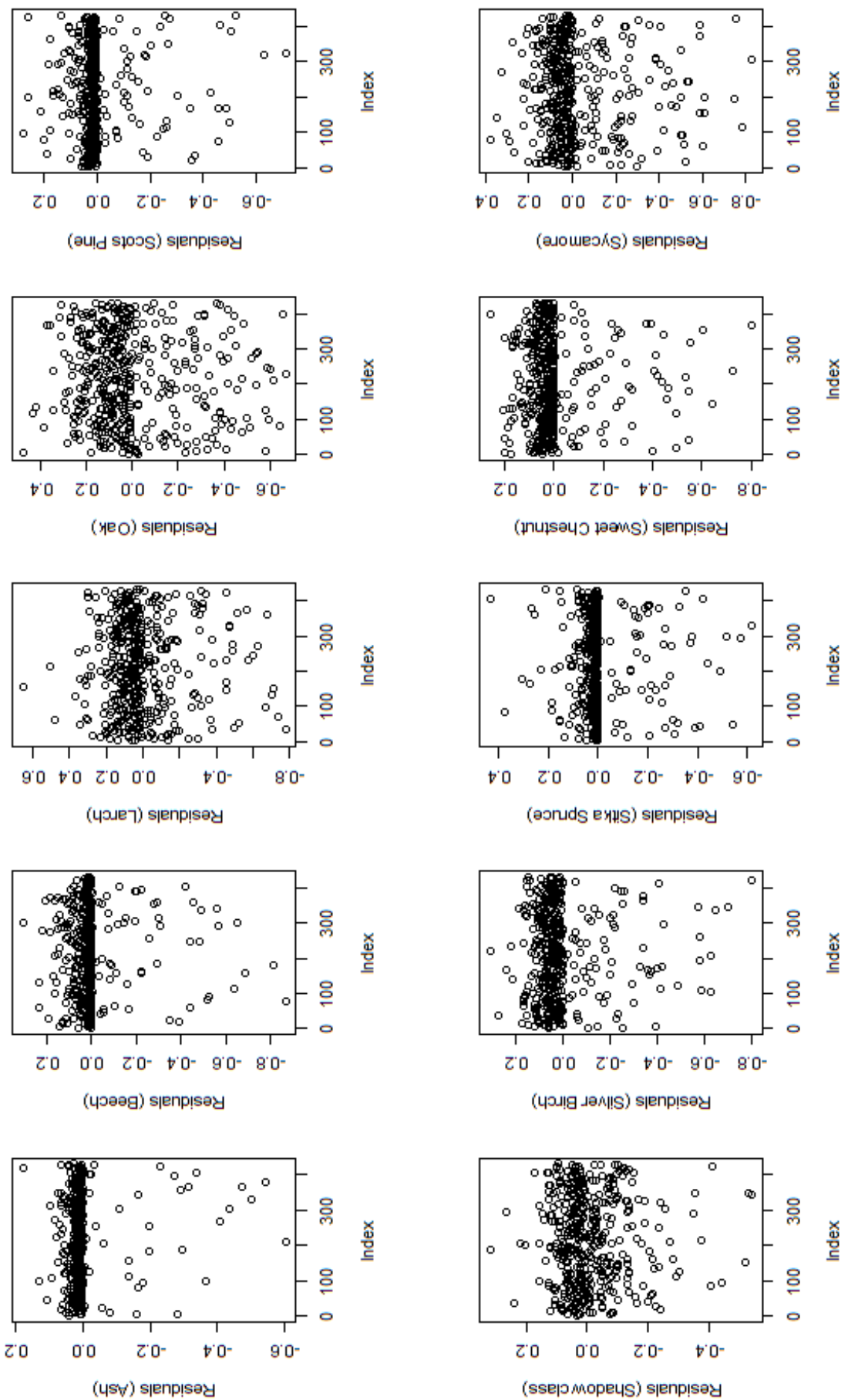


Figure 9.1: Residual scatter plots for each tree type, random forest model

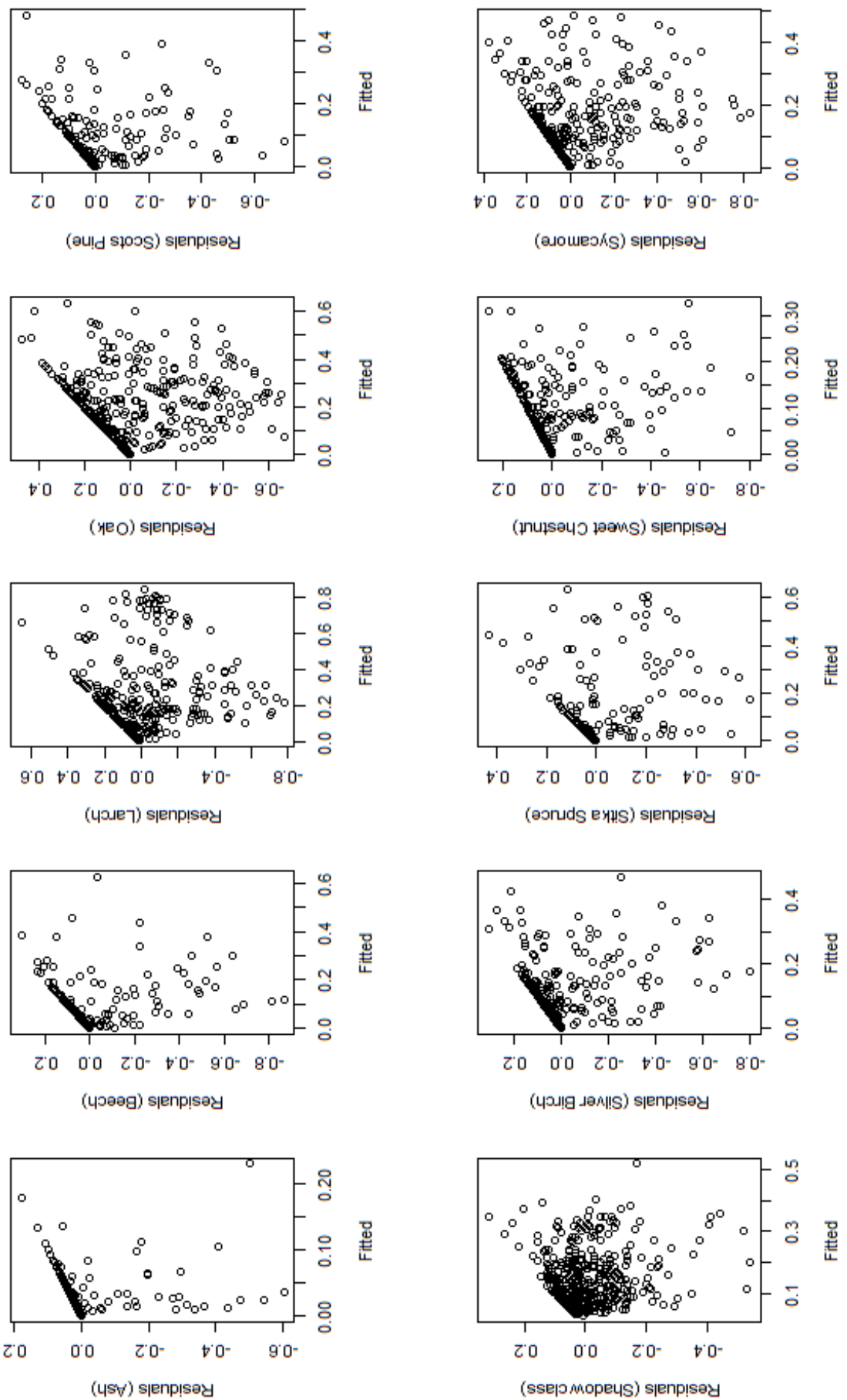


Figure 9.2: Residual vs. fitted values scatter plots for each tree type, random forest model

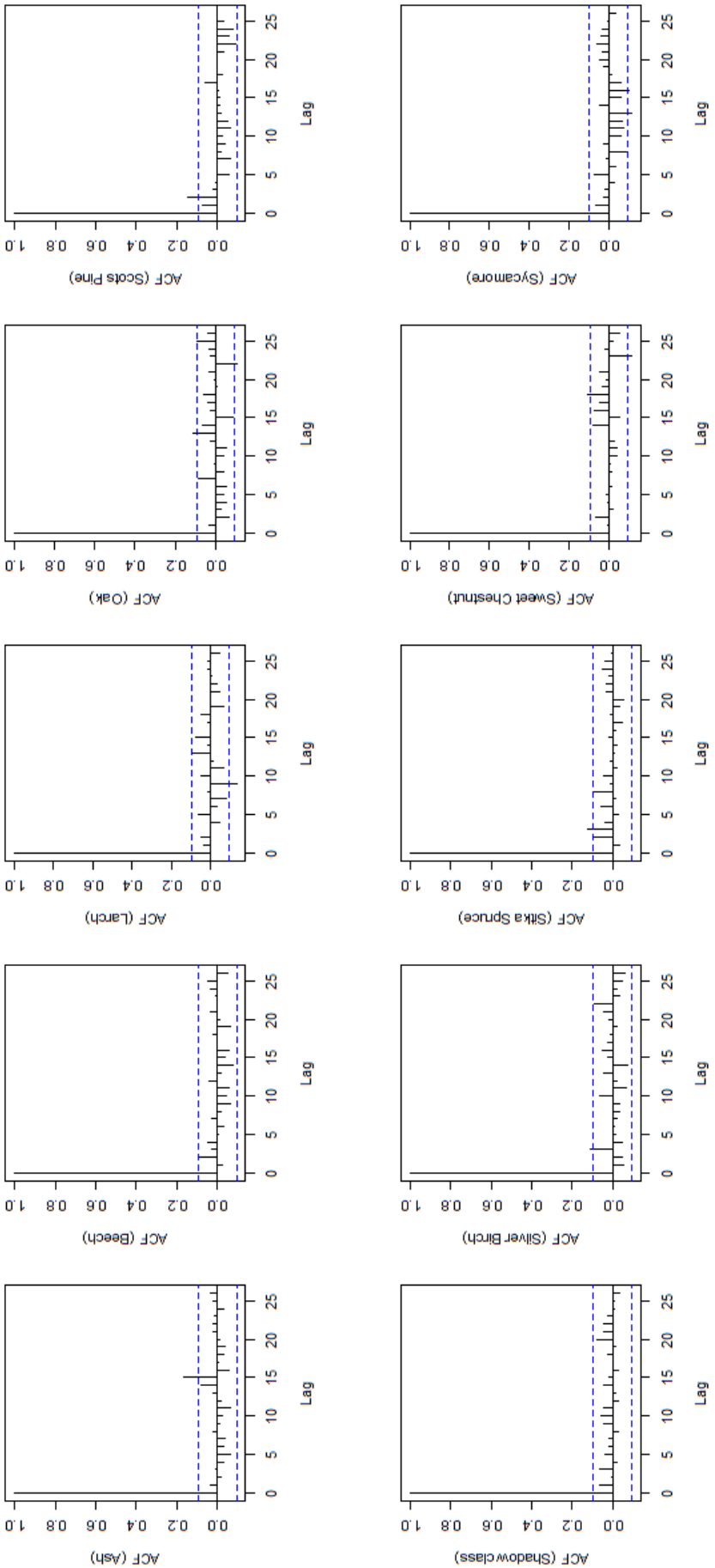


Figure 9.3: Residual ACF plots for each tree type, random forest model

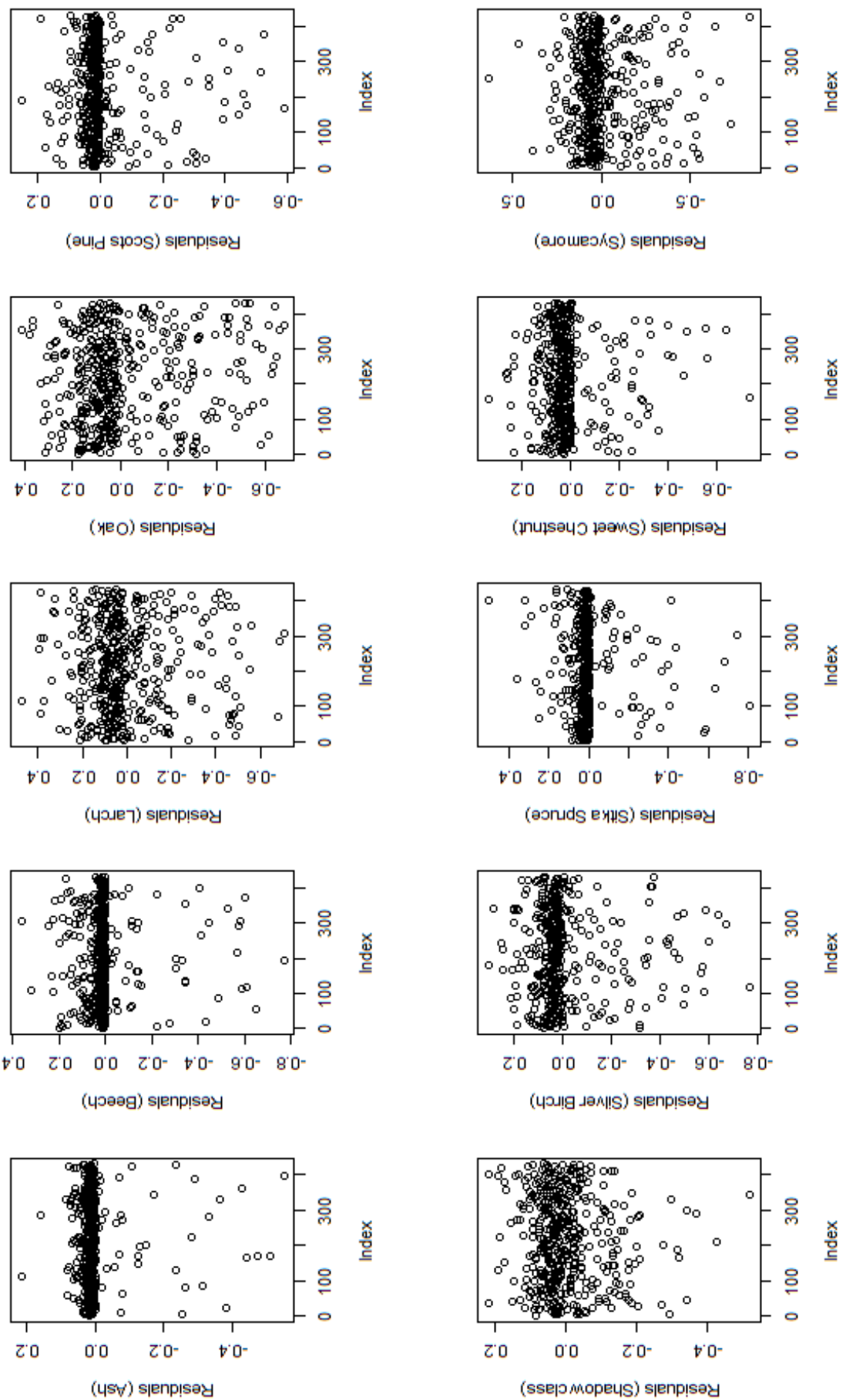


Figure 9.4: Residual scatter plots for each tree type, random forest model with spatial coordinates

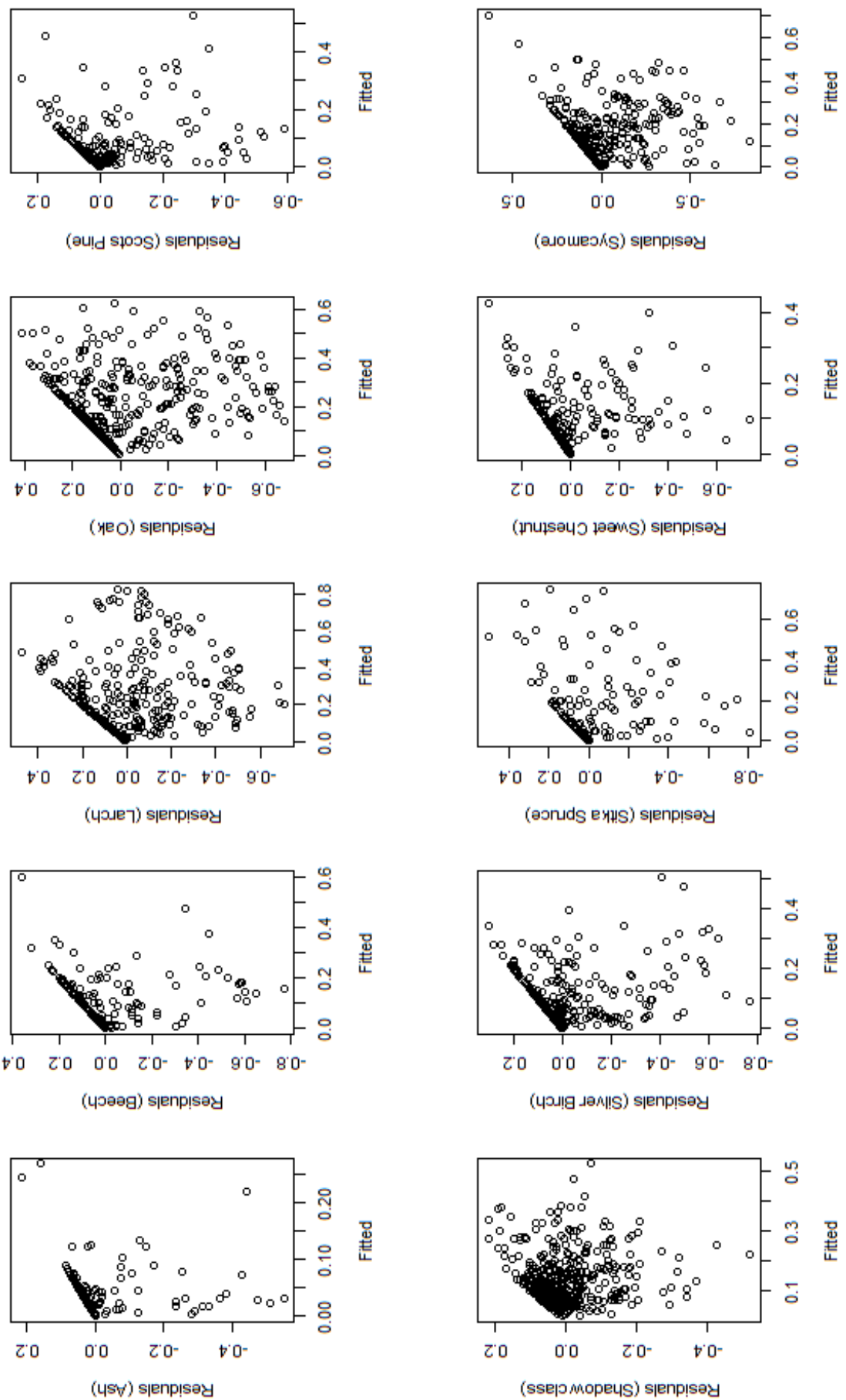


Figure 9.5: Residual vs. fitted values scatter plots for each tree type, random forest model with spatial coordinates

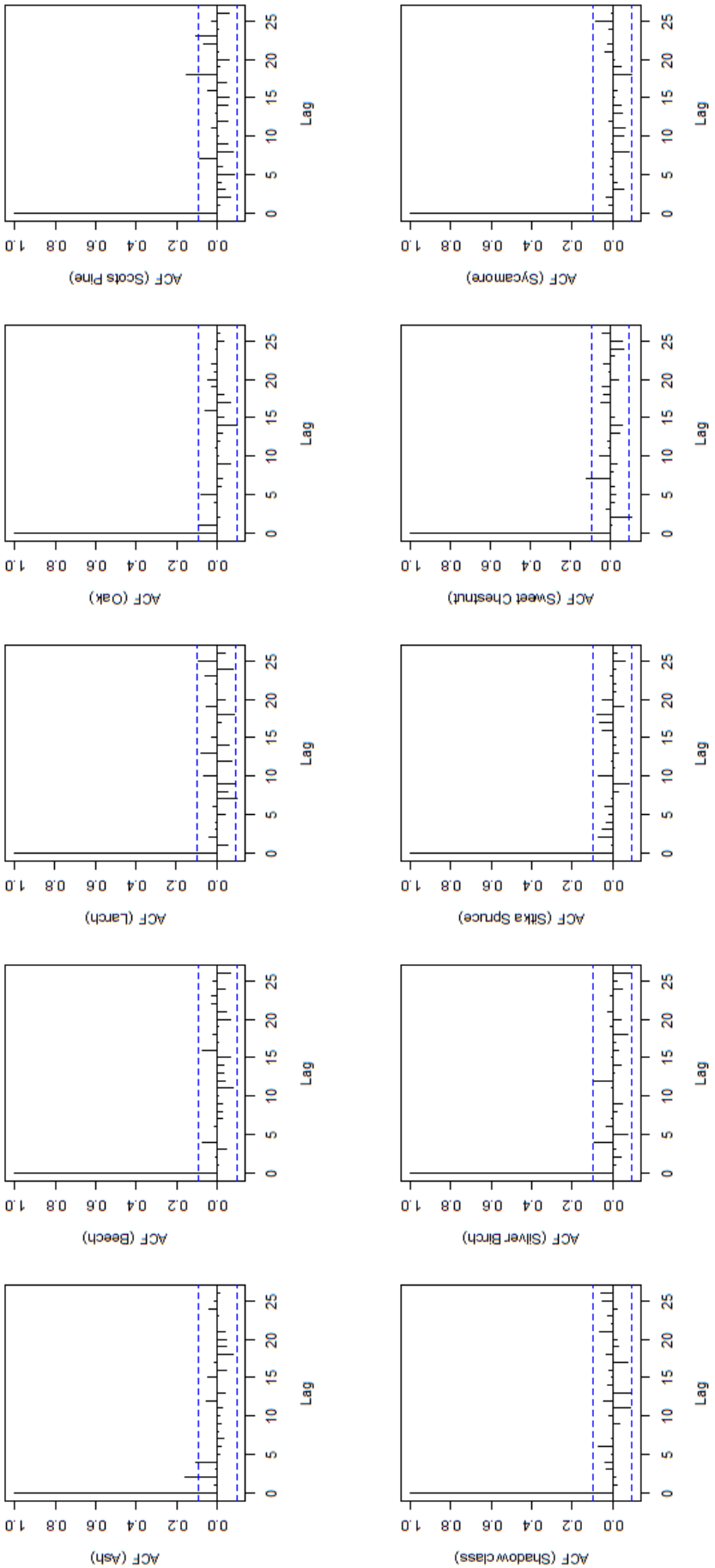


Figure 9.6: Residual ACF plots for each tree type, random forest model with spatial coordinates

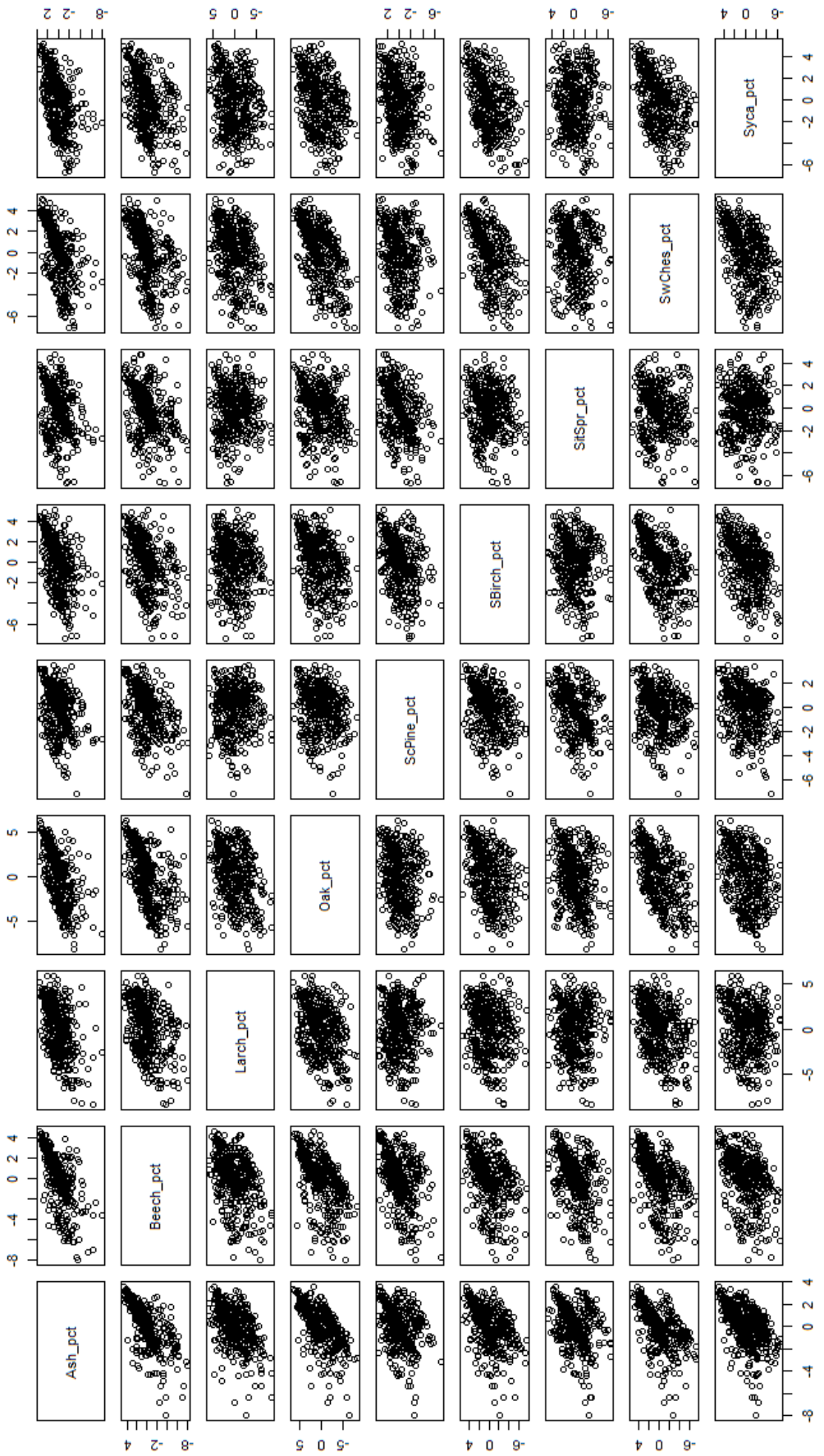


Figure 9.7: Residual plots for each tree type, alr-transformed multivariate Normal model

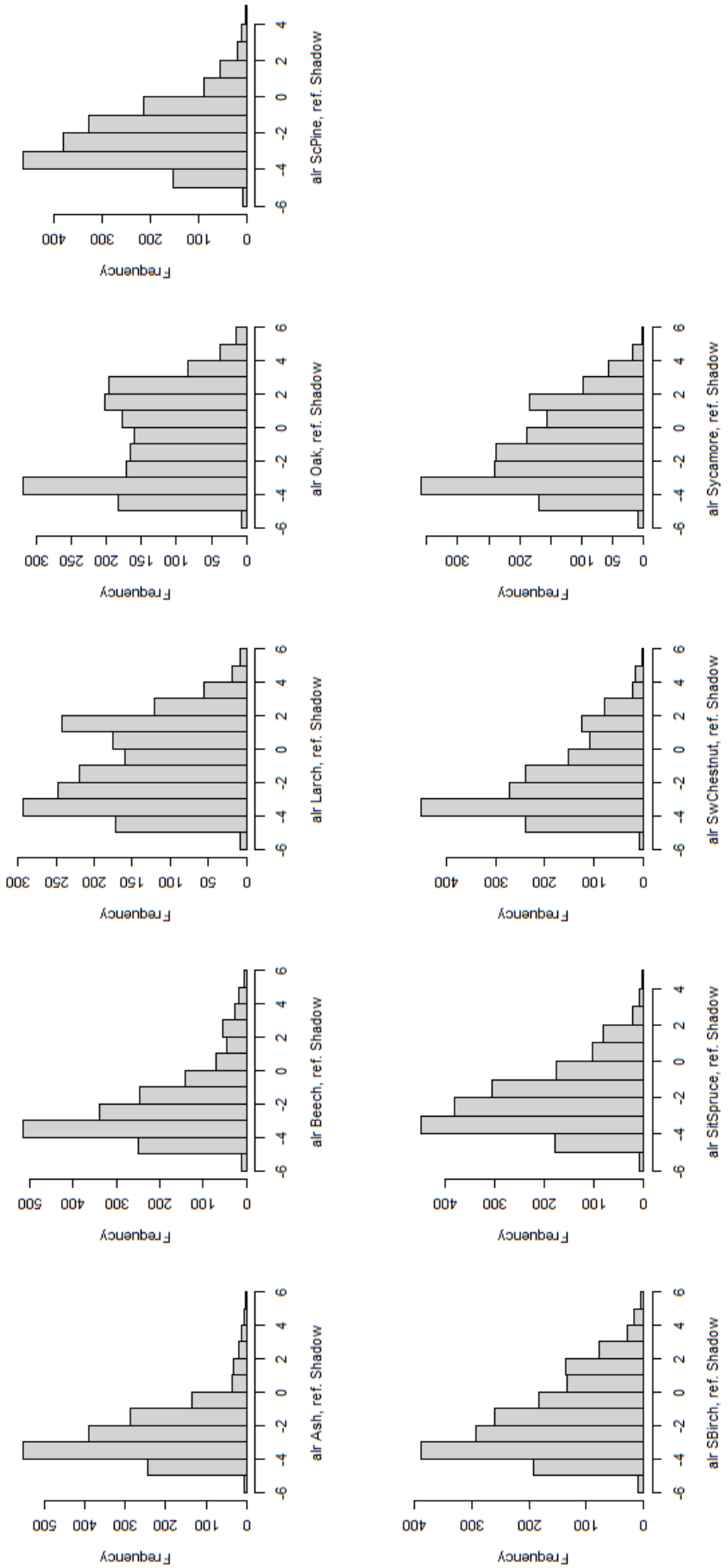


Figure 9.8: Histograms for alr-transformed tree type data

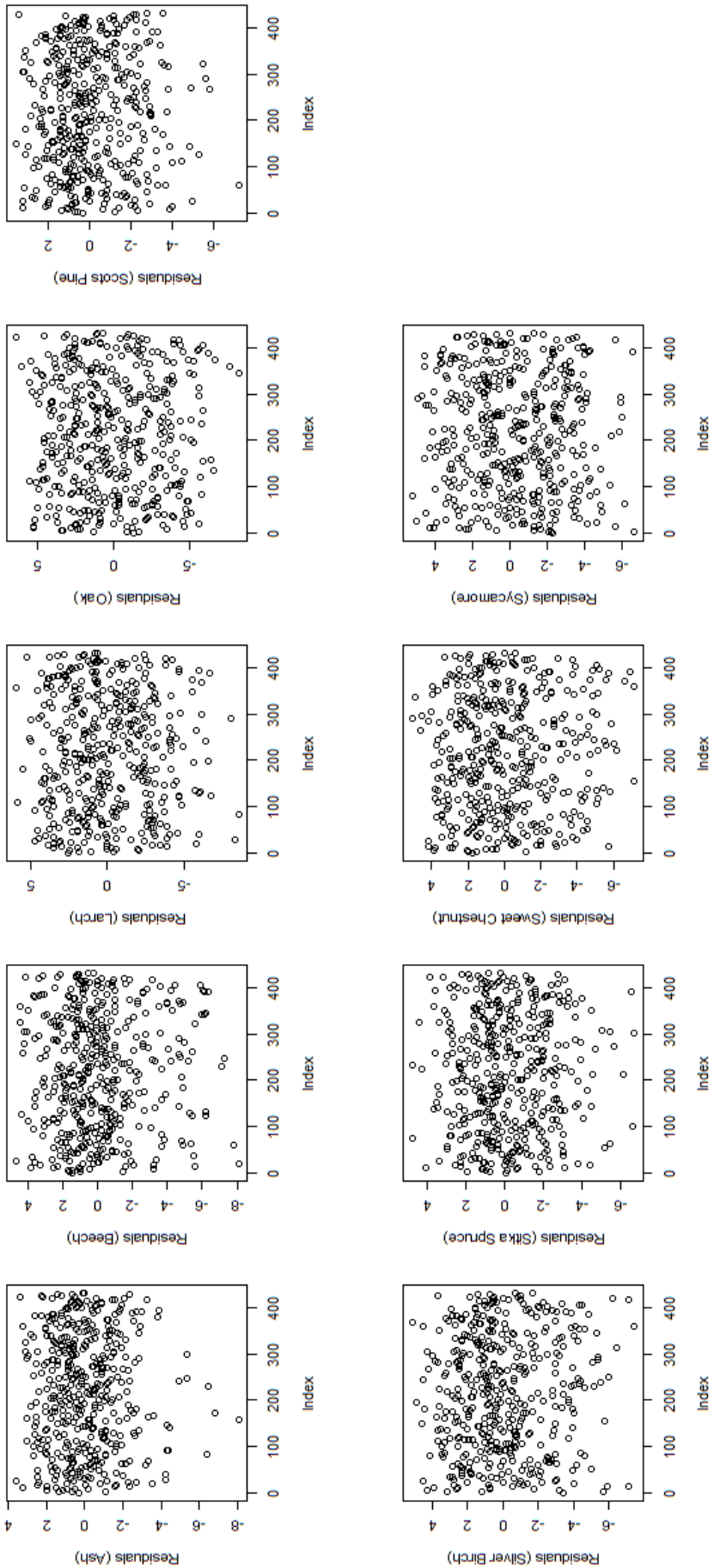


Figure 9.9: Residual scatter plots for each tree type, alr-transformed data with multivariate Normal model

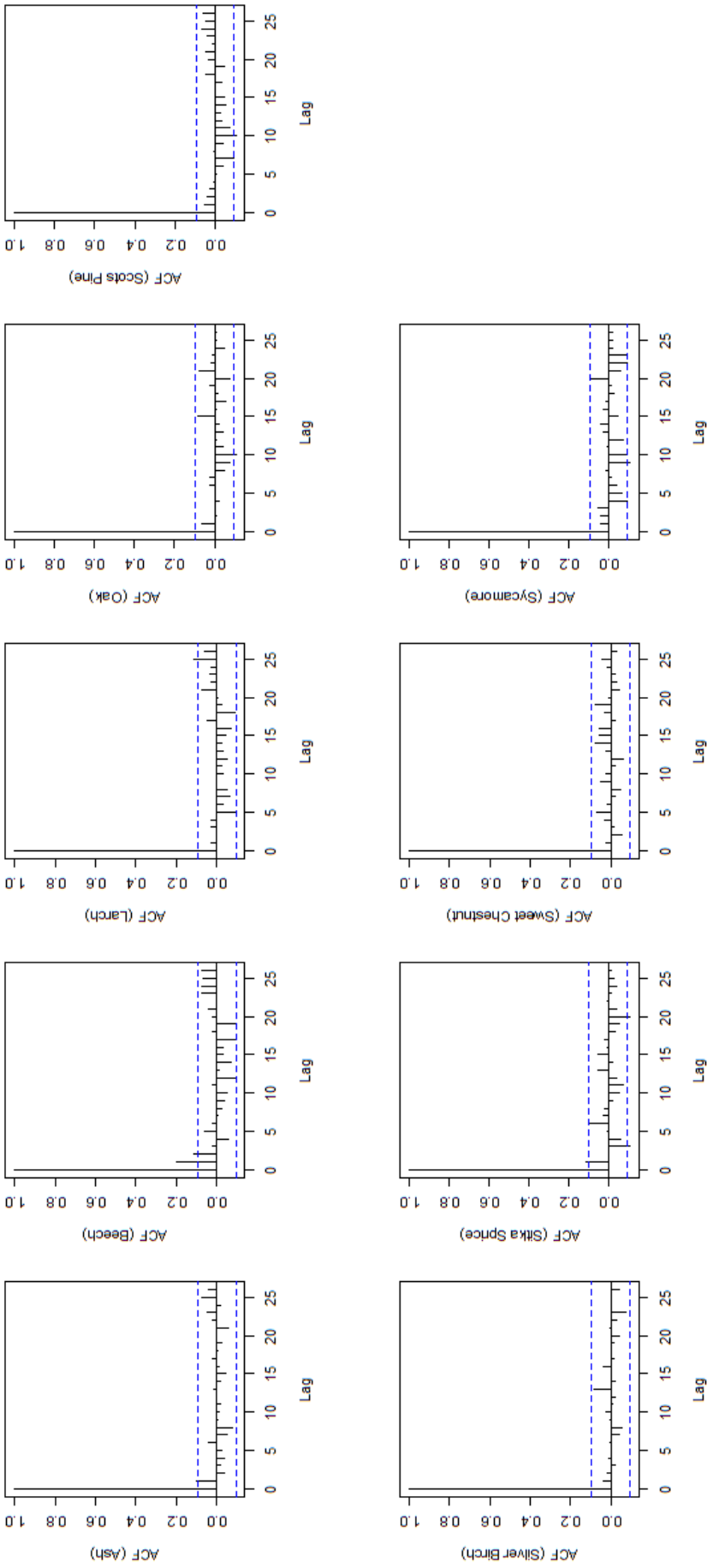


Figure 9.10: Residual ACF plots for each tree type, alr-transformed data with multivariate Normal model

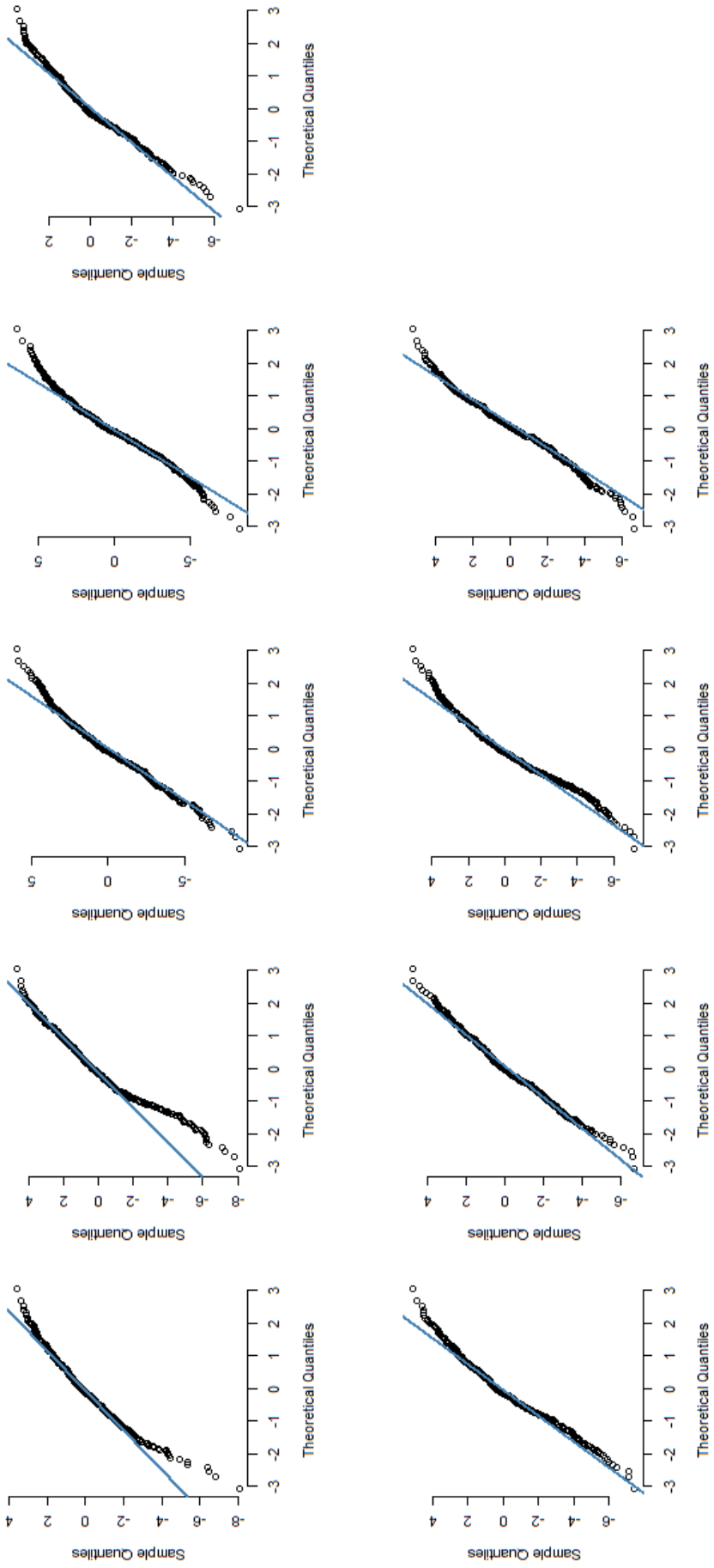


Figure 9.11: Q-Q plots for each tree type, alr-transformed data with multivariate Normal model

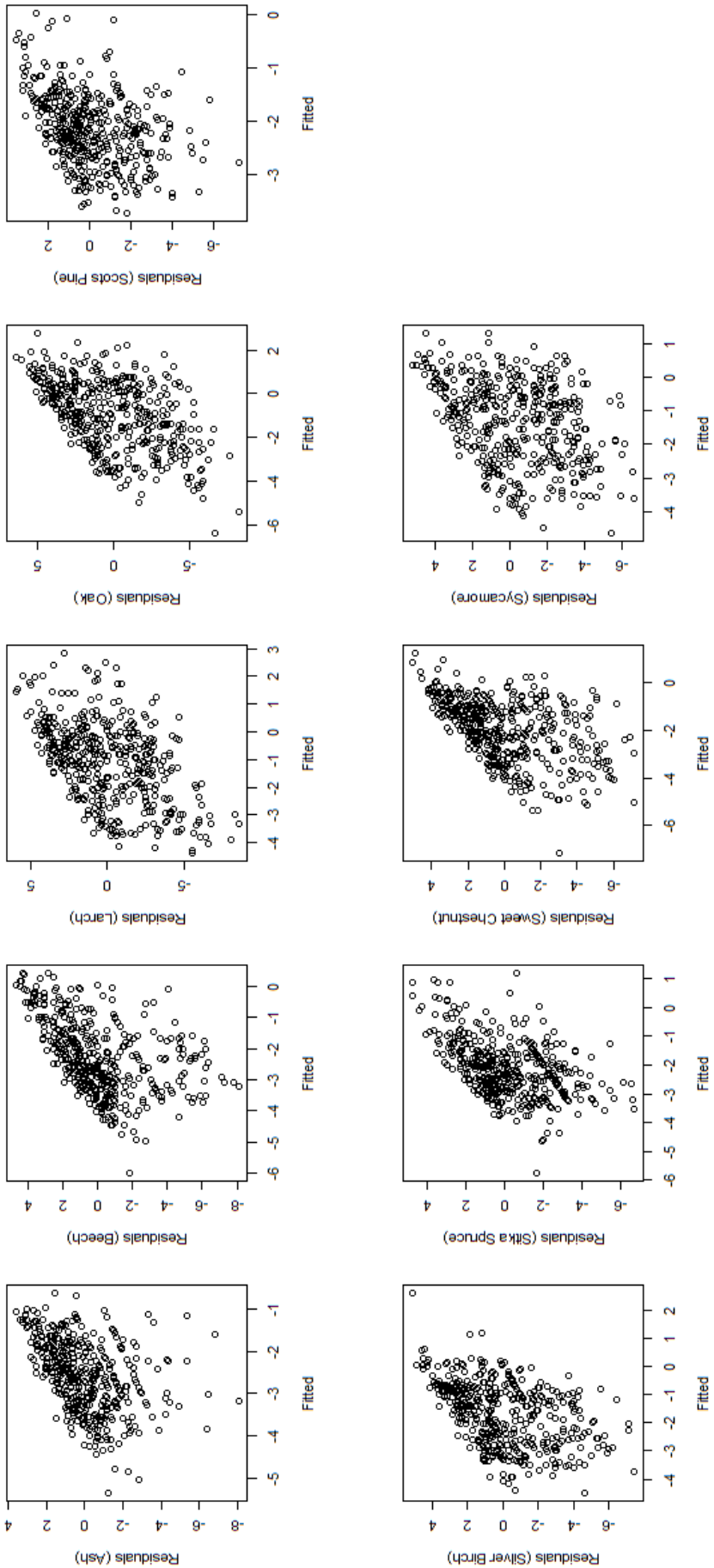


Figure 9.12: Residual vs. fitted values plots for each tree type, alr-transformed multivariate Normal model

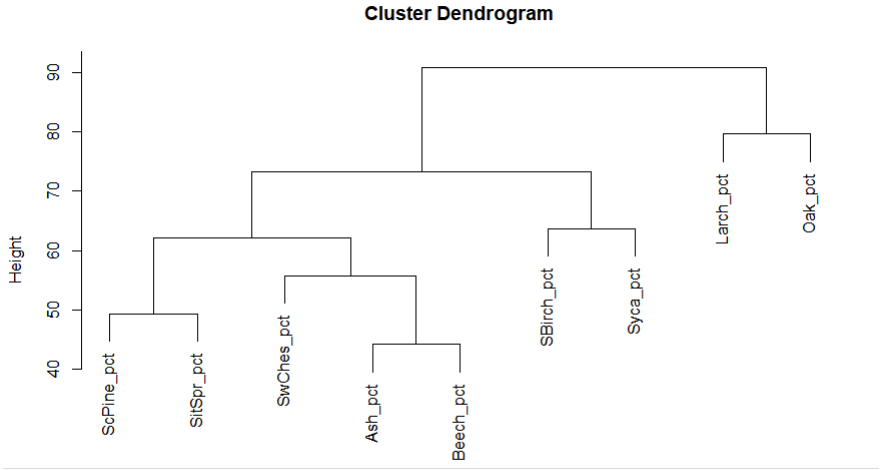
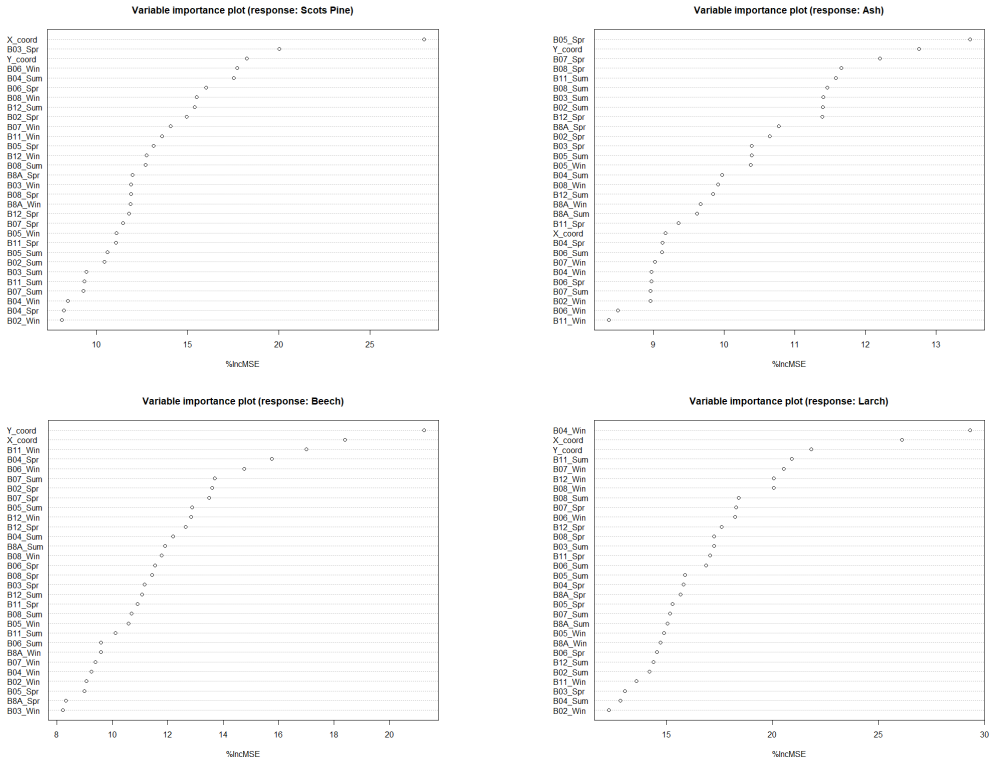


Figure 9.13: Cluster dendrogram for alr-transformed tree species data.



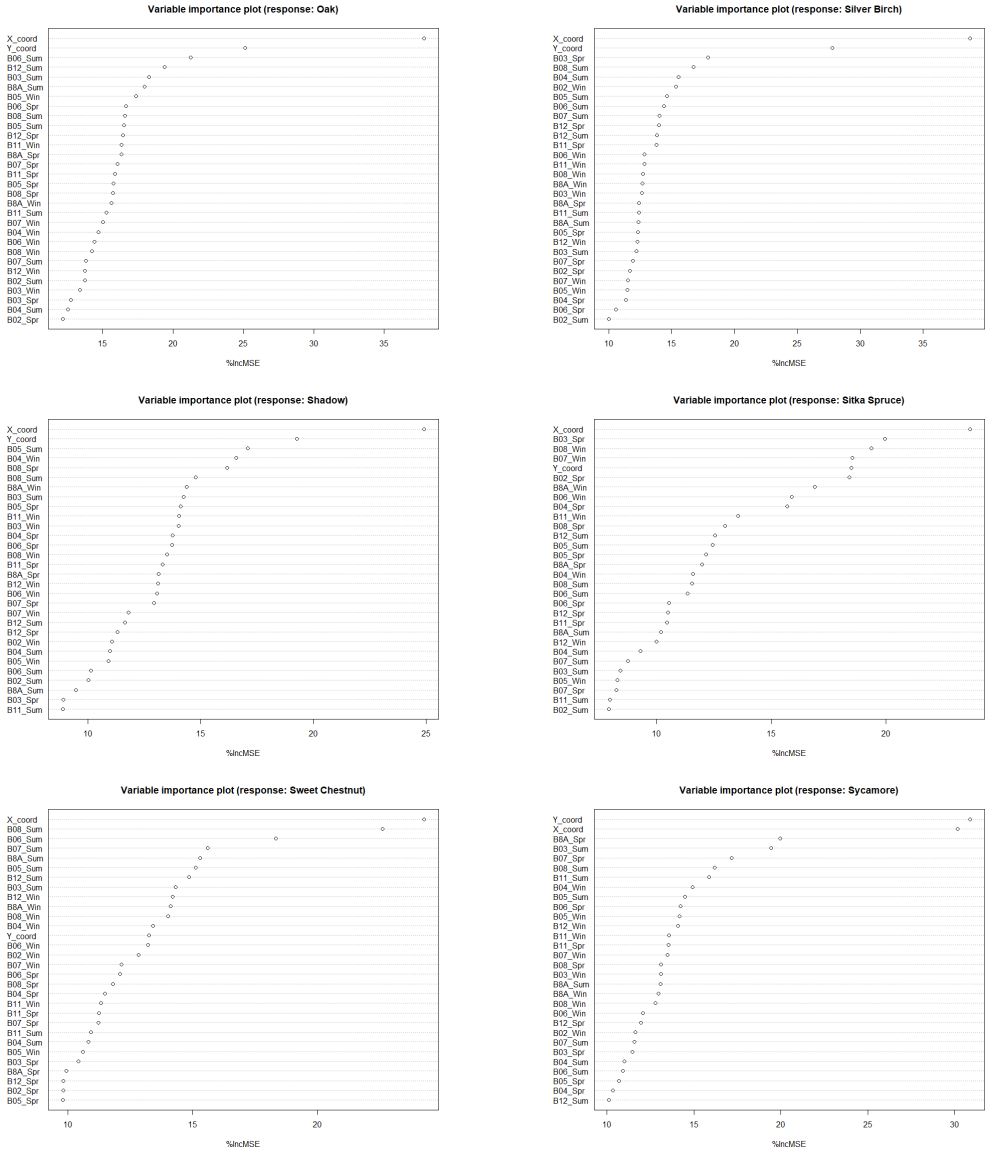
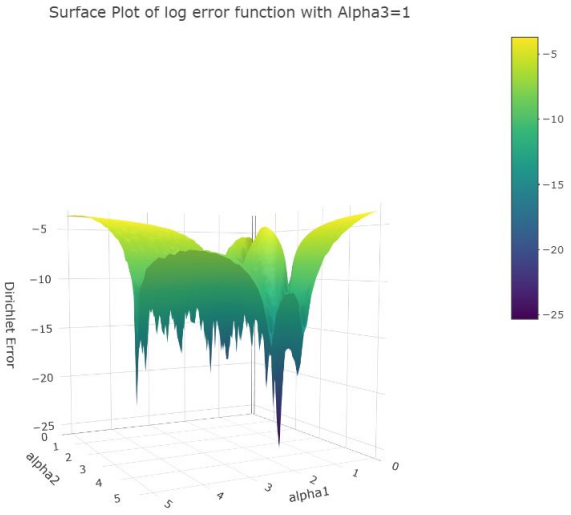
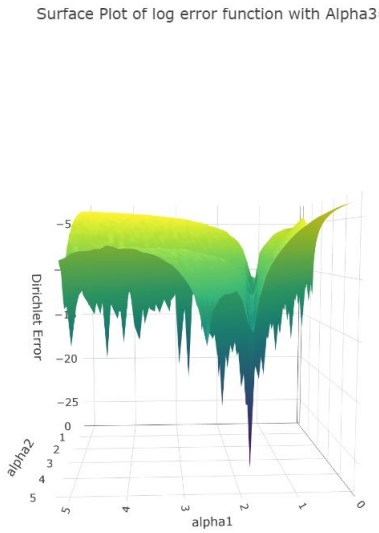


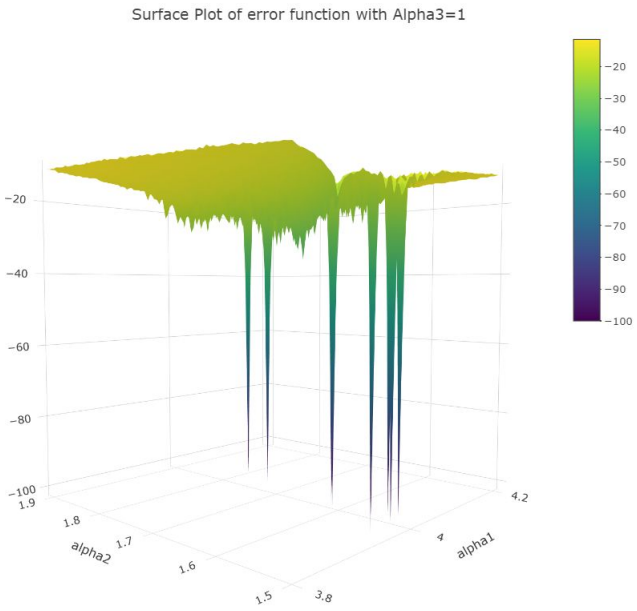
Figure 9.14: Random forest variable importance plots by tree type.

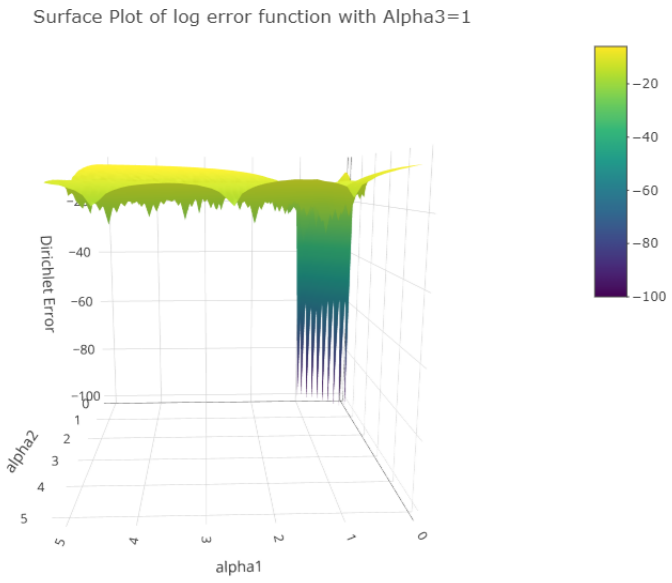
9.2 Appendix B

The following plots illustrate error functions of the simplex-partition (simplex dissection) fitting procedure to elicited statements in Chapter 7, Section 7.2. The graphs depict scenarios in a 3-dimensional setting, with 3-part compositional data. However, for visualisation purposes, one of the parameters of the Dirichlet distribution has been fixed (below plots depict $\alpha_3 = 1$, but others have too been trialled with similar results). Two axis represent values of α_1 and α_2, where a fitting to elicitation judgements is sought, and the vertical axis represents error fit. The colour scale on the right-hand side represents size of the error.






Surface Plot of error function with Alpha3=1
−20
−30
−40
−50
−60
−70
−80
−90
−100
−20
−40
−60
−80
−100
1.9
1.8
1.7
1.6
1.5
alpha2
3.8
4
4.2
alpha1


Surface Plot of log error function with Alpha3=1
−20
−40
−60
−80
−100
Dirichlet Error
−40
−60
−80
−100
0
1
2
3
4
5
alpha2
5
4
3
2
1
0
alpha1

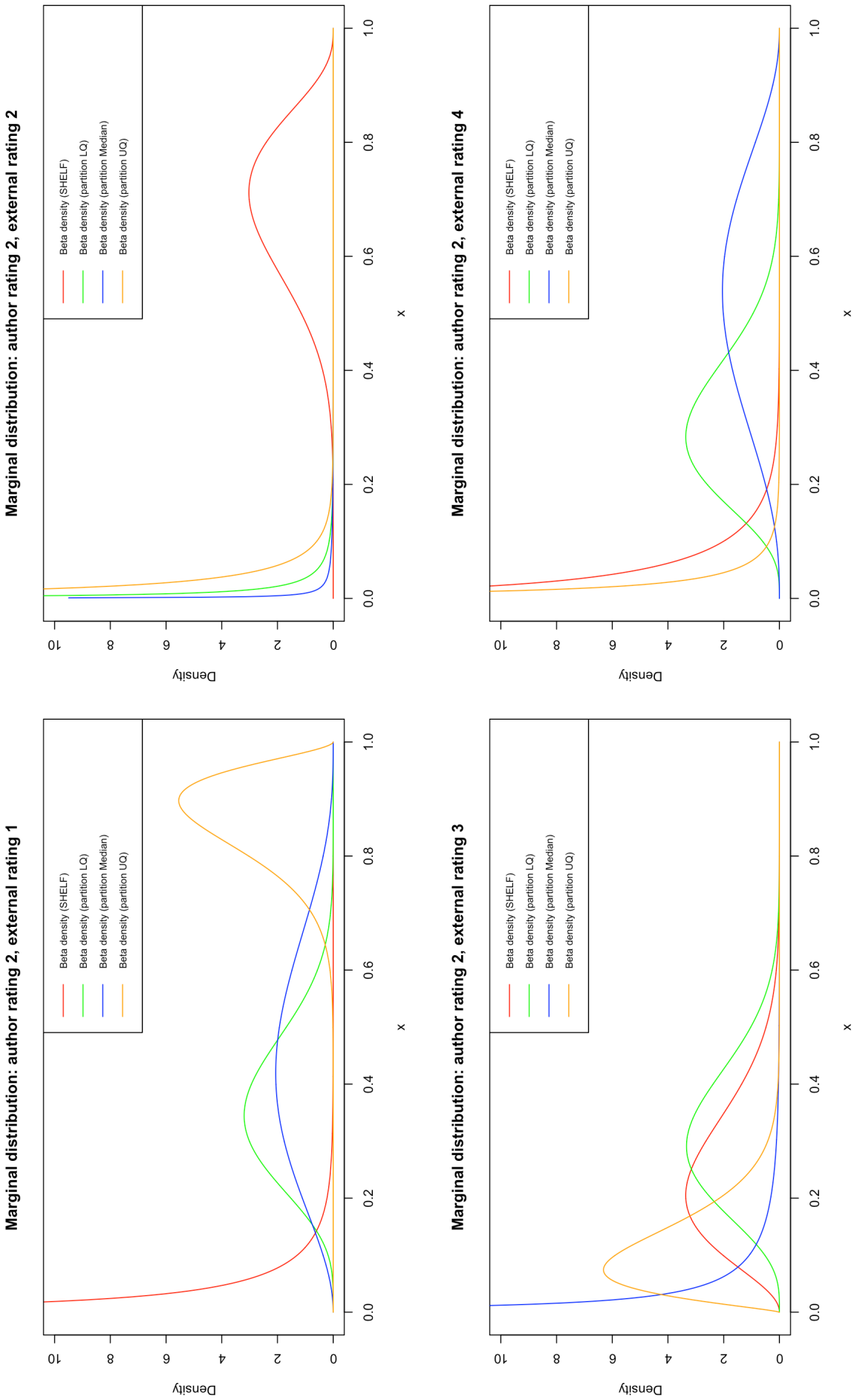


Figure 9.15: Marginal Beta distribution plots of ratings misclassifications.

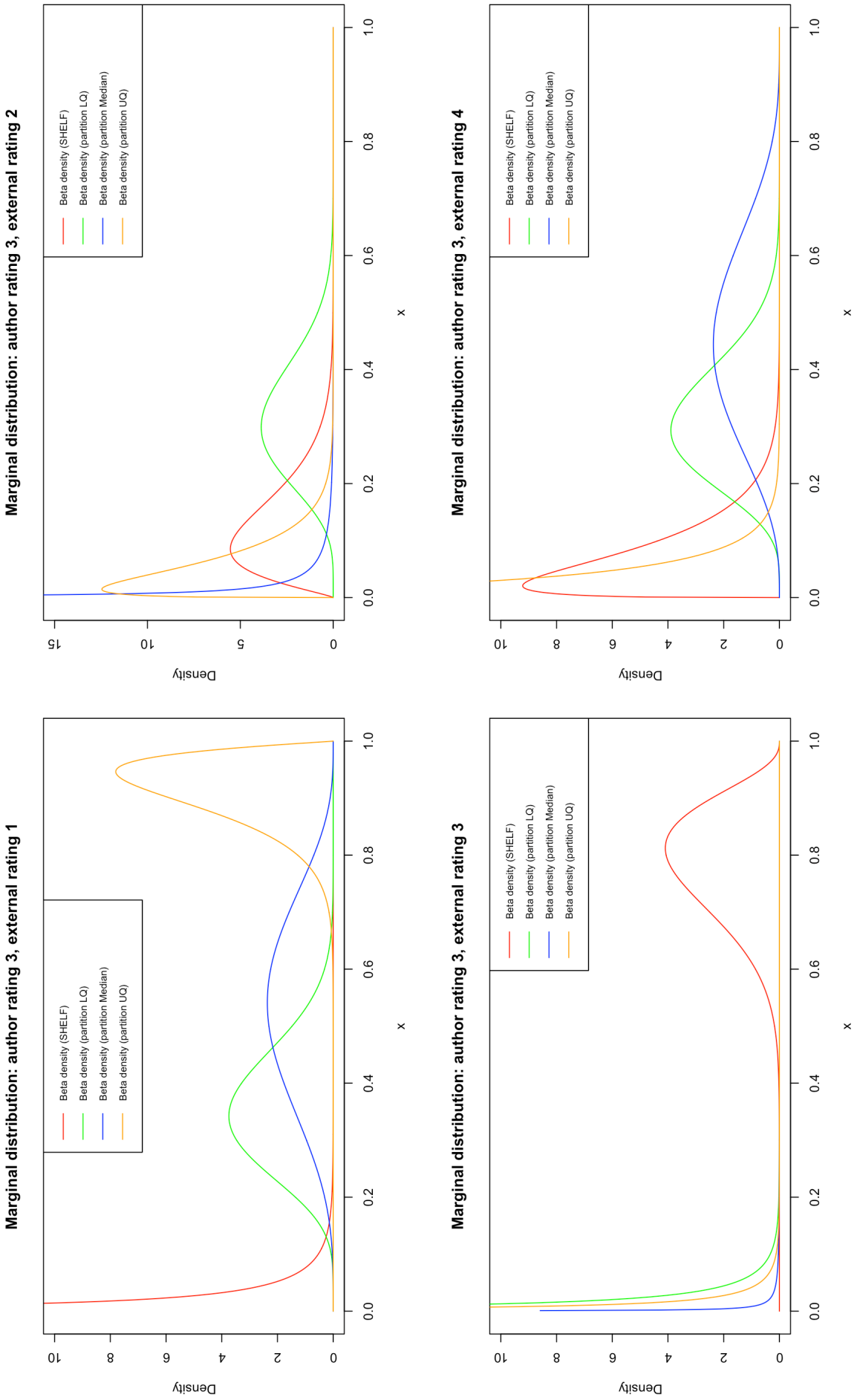


Figure 9.16: Marginal Beta distribution plots of ratings misclassifications.

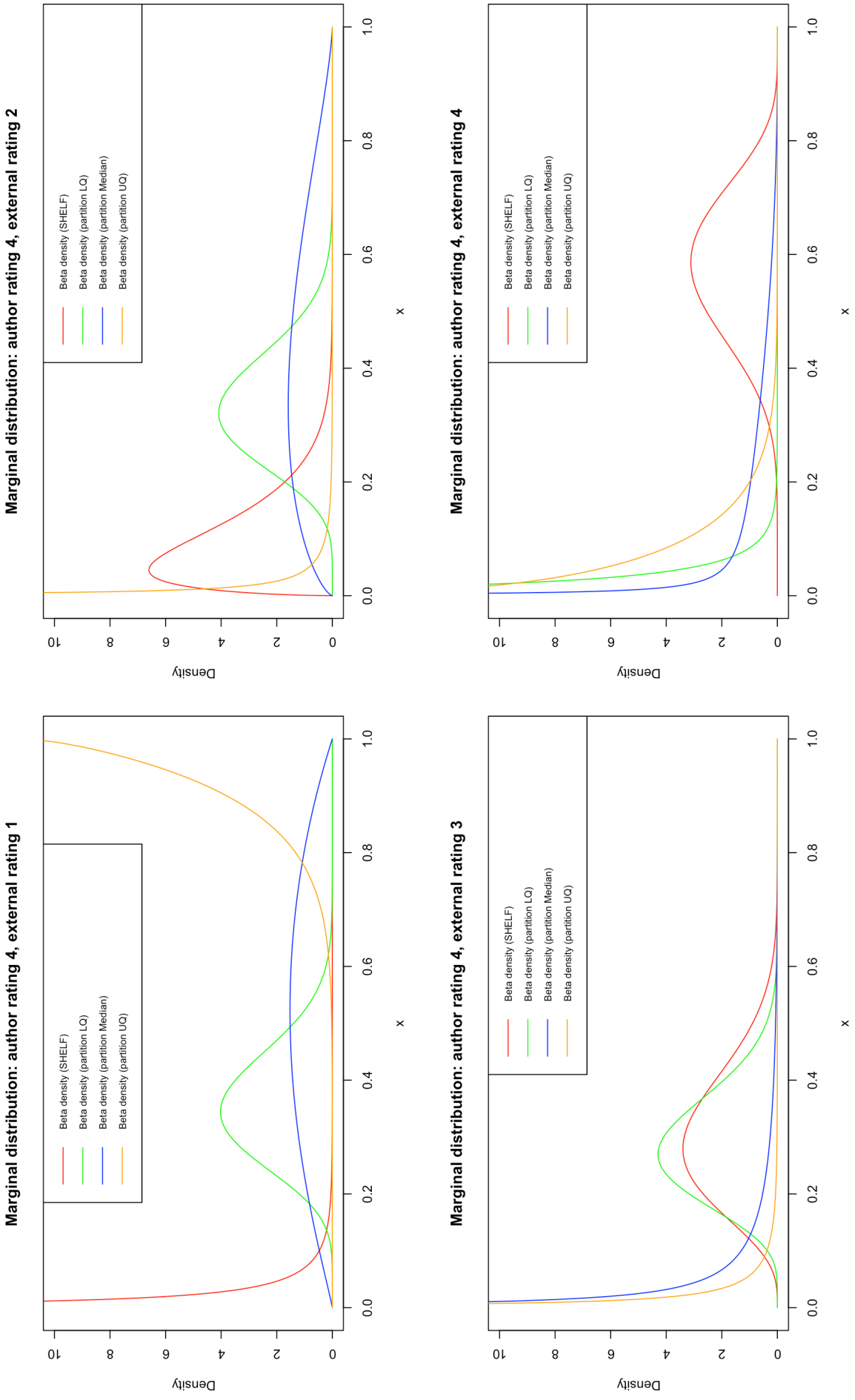


Figure 9.17: Marginal Beta distribution plots of ratings misclassifications.