Chapter 1

Introduction, background, and overview

1.1 Introduction

Non-negative multivariate data that satisfy a unit-sum constraint are known as compositional data. Historically, lacking good methods for analyzing such data, the unit-sum constraint was often simply disregarded. However, since the relevant sample space is the positive unit simplex, this type of data could be analyzed parametrically by adopting an appropriate distribution on the positive simplex.

Usually, compositional data results from normalizing data whose sample space is the positive orthant,

$$\Re_+^{n+1} = \{(z_1, \ldots, z_{n+1}) : z_1 > 0, \ldots, z_{n+1} > 0\}.$$

Let $\mathbf{z} = (z_1, \ldots, z_{n+1}) \in \Re_+^{n+1}$. A compositional vector $\mathbf{x} = (x_1, ..., x_n)$ can be formed by letting

$$\mathbf{x} = \left(z_1/t, \ldots, z_n/t\right), \text{ where } t = \textstyle\sum_{i=1}^{n+1} z_i .$$

Following the standard terminology, we will use the term "basis" to refer to the vector **z,** and the term "size" to refer to the quantity *t.* As can be clearly seen, a basis is uniquely specified by the corresponding size and composition, and conversely.

Compositional data arise in many fields; for example, Aitchison (1986) mentions chemistry, geology, biology, medicine, ecology, hydrology, manufacturing design, and economics. As a simple illustration, Leser (1963) divides U.S. farmers' expenditures into twelve categories, including food, tobacco, home maintenance, etc., and finds the proportion of each farmer's income spent on each of these categories. More examples will be given in chapter 2.

The correct statistical analysis is critically important for compositional data. Aitchison (1986) was the first to systematically consider alternatives to the Dirichlet distribution, which he felt was insufficient to model most compositional data sets, due to its inability to model positive correlations and its strong conditional independence assumptions. These two inadequacies caused us to think about extending this distribution. Also, they give some criteria that a "good" simplex distribution must possess.

Knowing these inadequacies, Aitchison developed an alternate model. In his approach, the original *n* variables in the composition $(x_1,...,x_n)$ are mapped to $\Re^n$ through the use of a log-ratio transformation. The *n* transformed variables are then modeled as multivariate normal, and the needed multivariate analyses are performed in $\Re^n$. Aitchison noted that the resulting logistic normal class unfortunately does not contain the Dirichlet distribution as a special case. That is, this class does not produce densities with extreme independence properties. Aitchison suggested two possible ways to address this problem. One is to transform the original data from the simplex to $\Re^n$ and look for extensions to the logistic normal class that can model extreme independence. The other is to stay on the simplex and look for a new family of distributions that contains the Dirichlet but also contains distributions capable of modeling nontrivial dependence structures.

In this dissertation, we adopt the latter approach. Also, we consider additional criteria that a "good" simplex distribution must possess. Generally speaking, one would like a simplex distribution that has the following properties:

1) the ability to model both positive and negative covariance;

2) little or no restriction on the structure of the covariance matrix (beyond of course the requirement that it be non-negative definite);

3) a less severe independence structure than the Dirichlet distribution;

4) a clear method of estimating its parameters from the available data;

5) the property that compositions with this distribution also have subcompositions with this distribution;

6) tractability of its moments as well as its normalizing constant;

and

7) invariance with respect to what Aitchison (1986) called the "fill-up value," $1-\sum_{i=1}^{n} x_i$. In other words, if two analysts use the same simplex distribution to model the same compositional data set, but choose different variables to be the fill-up value, then they should reach the same inferences regardless of their choice of fill-up value.

Also, for Bayesian analysis one would ideally like the distribution to be:

8) closed under multinomial sampling; i.e., a conjugate prior family for multinomial sampling.

Properties 1 and 3 resulted from the comments given by Aitchison about the Dirichlet distribution. Properties 5 and 7 are suggested by Smith (1994). Properties 2, 4, 6, and 8 are our own criteria.

One should note that a "good" simplex distribution is relative; i.e., what is good for one application may not be good for another. For example, insensitivity to the choice of fill-up value is important if the underlying variables are symmetric or exchangeable, but reduces the flexibility of the distribution in cases where symmetry is not desirable. Thus, some of these criteria may be desirable for one application and not for another.

James (1981) commented that there were few tractable distributions for random proportions other than the Dirichlet. Aitchison (1986) similarly commented that "attempts to obtain a suitably rich class of distributions containing the simple Dirichlet class have so far failed." Fang et al. (1990) stated that "In analysing so-called 'compositional data,' scientists have been handicapped by the lack of known distributions to describe various patterns of variability." Hutchinson and Lai (1991) stated that with the work of Aitchison (1986), "this whole field has received new impetus and is likely to see productive interaction of data with theory over the next few years." Rayens and Srinivasan (1994) gave a generalization to the Liouville family on the unit simplex and noted, "It appears that no one has investigated the extent to which this family addresses Aitchison's challenge to find a family that contains the Dirichlet but also contains densities with sophisticated dependence structures." Later in chapter 4 we will see that this generalized Liouville distribution is still inadequate because it is difficult to construct a generalized Liouville distribution on the simplex with a desired covariance structure, and it can not model some independence concepts. In addition, it is not clear how one could estimate the large number of parameters in its density function. Joe (1997) provided an extensive discussion of non-normal multivariate models, including models based on mixtures, latent variables, and stochastic representations. He commented that "Until recently, little research had been done in the area of multivariate non-normal distributions."

Motivated by the above comments, one of our concerns in this work is to identify a family of distributions that contains the Dirichlet but also contains distributions capable of modeling nontrivial dependence structures, as well as satisfying some of the criteria given above. Specifically, we generalize the approach suggested by Krzysztofowicz and Reese (1993) (which will be discussed later) through the use of dependent ratios (rather than independent ratios, as used in their work). In addition, we also develop new families of distributions on the positive orthant $\Re_+^n$ that encompass as special cases the Liouville, conditional generalized Liouville, adaptive Dirichlet, and multiple Dickey distributions. We then use some of these new distributions in applications. The availability of methods to estimate the models' parameters in terms of the original composition and its covariance structure is one of the advantages of our new distributions. In addition, in some applications our new models appear to resolve problems where the models originally used by other investigators were not satisfactory.

1.2 Background

Connor and Mosimann (1969) proposed a generalization of the Dirichlet distribution. However, Aitchison (1986) noted that this generalization still had many independence properties.

Dickey (1968) defined the so-called scaled Dirichlet distribution, which likewise has a strong conditional independence structure. Dickey (1983) further generalized this distribution to the Dickey and multiple Dickey distributions, but it is unknown whether or not these distributions can admit positive covariance, as Smith (1994) noted.

Barndorff-Nielsen and J∅rgensen (1991) derived new classes of parametric models on the unit simplex by conditioning independent generalized inverse Gaussian random variables on their sum. This class of distributions has the Dirichlet distribution as a special case. However, the developers pointed out that their models "do not yet provide a full alternative to the logistic normal distributions as far as statistical analysis of compositional data is concerned, principally because the latter family of distributions has m(m-1)/2 covariance parameters, whereas our distributions have only one variance, or rather precision, parameter." Thus, these distributions will presumably have rather limited covariance structures. Also, Smith (1994) noted that compositions with this distribution do not have subcompositions with the same distribution. Finally, it is not clear whether this class of distributions admits positive covariances.

Grunwald, Raftery, and Guttorp (1993) developed two new distributions on the unit simplex, the Dirichlet conjugate distribution and the Dirichlet conjugate Dirichlet distribution. These are based on the Dirichlet distribution, but generalize it to allow for dependence between the proportions. They used these distributions to model the fractions of world automobile sales by Japan, the USA, and the rest of the world. They considered time series data, and focused on the proportions of total sales over time, rather than the actual amounts. Again, it is unclear whether these distributions satisfy the properties we have identified as being desirable.

Krzysztofowicz and Reese (1993) proposed the so-called adaptive Dirichlet (AD) distributions on the unit simplex. They presented a number of examples illustrating their approach, and showed that the Connor-Mosimann distribution can be derived as a special case of their models. As noted by its developers, "The family of distributions characterized herein constitutes the ultimate generalization of the Dirichlet distribution that can be obtained through an independent bifurcation process. This generalization does not totally relax the constraints on the correlation structure of fractions that the Connor-Mosimann and standard Dirichlet distributions impose. It nonetheless offers a much richer model of the stochastic dependence among fractions." Later in this dissertation, we will show the need to relax the independent bifurcation assumption considered by Krzysztofowicz and Reese, to come up with a new family of distributions with more general correlation structures than those imposed by the AD distributions.

Smith (1994) examined numerous known distributions, to evaluate whether or not they could be considered good simplex distributions. He showed that Liouville distributions are able to model positive covariance, but have many of the same independence assumptions as the Dirichlet distribution. The generalized Liouville distributions of Rayens and Srinivasan (1994) again admit positive covariance and can model more general dependence structures, but are not invariant with respect to the choice of the fill-up value, one of Smith's criteria. Therefore, Smith proposed a new class of simplex distributions, the conditional generalized Liouville (CGL) distributions. Smith noted that the scaled Dirichlet distribution and Dickey distribution are conditional generalized Liouville distributions, while the multiple Dickey distribution is not. CGL distributions are invariant with respect to the choice of the fill-up value, admit positive covariance, and provide a fair degree of flexibility in being able to model some forms of independence without others. However, it is currently difficult to construct either generalized Liouville or CGL distributions with specified covariance structures. In addition, it is not clear how to estimate the large number of parameters in the densities of those distributions.

Gupta and Richards (1995) defined a class of distributions, containing the classical Dirichlet and Liouville distributions, in which the random variables are defined on a locally compact Abelian group or semigroup. They showed that this class has many properties of the Dirichlet and Liouville distributions, and discussed other properties of the marginal and conditional distributions. Gupta and Richards presented a number of examples illustrating the general theory, and showed that Barndorff-Nielsen's and J∅rgensen's model can be derived as a special case of their model. This work was mainly theoretical, and the authors did not give any application of their theory or test their distributions against the properties we have suggested as desirable.

To summarize, there has been a great deal of interest in constructing new families of distributions to allow more general dependence structures on the unit simplex and on the positive orthant. Some researchers (e.g., Smith, 1994) were interested in this problem primarily from a mathematical point of view. Others (e.g., Krzysztofowicz and Reese, 1993) had real-world problems (in their case, modeling snowmelt runoff) that caused them to construct new models. In this work our primary interest is theoretical, but we also consider applications of our results to two different real-world problems.

Figure 1.1 summarizes some of the existing extensions to the Dirichlet and Liouville distributions, which are also reviewed in Chapters 3 and 4. Figure 1.2 highlights the new extensions that we propose in Chapters 5, 6, and 7.