CHAPTER 2

Compositional Data

2.1 Introduction

A compositional data set is a set of vectors $(x_1,...,x_{n+1})$, where $x_i \geq 0$ for all i and

$$x_1+...+x_{n+1}=1. \qquad (2.1)$$

Compositional data arises in many fields, so it is clearly important to have valid methods of analyzing such data. Unfortunately, the unit-sum constraint (2.1) is sometimes just ignored in practice (Aitchison, 1986). In other cases, the Dirichlet distribution is used to model compositional data, even when the conditional independence properties of the Dirichlet are not appropriate to the data-generating process. Either of these problems can lead to results of questionable validity.

This chapter introduces some examples of compositional data in section 2.2, and presents some relevant background concepts from Smith (1994) in section 2.3 to 2.6. One should note that the terms "simplex" and "unit simplex" will be used equivalently in this work. Also, the term "fraction" will be used equivalently with "proportion."

2.2 Examples of compositional data

Compositional data arise naturally in many areas of science. For example, Aitchison (1986) mentions chemistry, geology, biology, ecology, hydrology, manufacturing design, medicine, and economics. Some examples will be given below.

Example 2.1 (*Reliability systems*)

A *Bernoulli system* is a coherent system made up of components, each of which can either succeed or fail according to a Bernoulli process. (e.g., a *k*-out-of-*N* system that operates if and only if at least *k* of its *N* components operate successfully, for k >1). See Azaiez (1993) for more rigorous definition.

Let

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_N) | \, \omega_i \in \{0,1\} \quad \forall \, i = 1,2,\ldots,N \} \tag{2.2}$$

be the set of all attainable *N*-dimensional binary vectors,

E_ω be the event that, for all *i,* component *i* fails if $\omega_i = 0$ and succeeds if $\omega_i = 1$, for ω in Ω, and

x_ω be the probability of event E_ω, for any ω in Ω.

From the above we have:

$$\sum_{\omega \in \Omega} x_\omega = 1 \tag{2.3}$$

So, the set $\{x_\omega | \omega \in \Omega\}$ is a composition. Also, the system failure probability is given by

$$P_f = \sum_{\omega \in \Psi} x_\omega , \tag{2.4}$$

where $\Psi \subset \Omega$ is the set of states that cause the system to fail.

We are interested in finding joint prior distributions for $\{x_\omega | \omega \in \Omega\}$ so that the posterior system failure probability (i.e., the posterior distribution of P_f) obtained by updating this prior with data on the occurrences of the various system states is the same as if we instead used system-level success and failure data. This property has been described as "perfect aggregation" (Azaiez, 1993). A full discussion of this problem will be given in chapter 7.

Example 2.2 (*Diagnosis problem*)

Consider two serious illnesses of the lungs, pneumonia and emphysema. These two illnesses may be alternative ways of accounting for some of the same symptoms. In this context the "ratios" may be quantities such as the conditional probability that a patient with particular symptoms has emphysema, given that the patient does or does not have pneumonia. Since these are conditional probabilities for the same event under somewhat different conditions, it is clear that they might be correlated. Because of that, it seems likely that none of the existing simplex distributions can adequately model this kind of problem.

Example 2.3 (*Segmented market*)

Markets will in general have multiple brands of the same product, each one competing to maximize its market share. Since the total of all market shares must equal 1, modeling market shares may entail the use of a suitable simplex distribution.

Moreover, markets may contain several submarkets (e.g., premium, regular, and economy versions of the same product), with products competing primarily against others in the same submarket (Day et al., 1979). In this case, the Dirichlet distribution would clearly not be an adequate model for market shares. For further discussion of the complexities that can arise (e.g., overlapping submarkets), see Arabie et al. (1981), Srivastava et al. (1984), and Queen et al. (1994).

2.3 The sample space and terminology

As mentioned earlier, compositional data are constrained data. Suppose that a composition has n+1 elements. If n of the elements of that composition are known, then one can determine the last element. Thus, a composition with n+1 elements is referred to as an *n*-

dimensional composition. The sample space for an n-dimensional composition is the n-dimensional simplex, defined by

$$S^n = \{(x_1,\ldots,x_n): x_1 > 0,\ldots,x_n > 0;\ x_1+\ldots+x_n < 1\}.$$

If $\mathbf{x} \in S^n$, then $x_{n+1} = 1 - x_1 - \cdots - x_n$ will be called the *fill-up value*.

Such data frequently arise by normalizing data on 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 $\mathbf{x} = (x_1,\ldots,x_n)$ can be formed by letting

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

We will use the term "*basis*" to refer to the vector **z,** and the term "*size*" to refer to the quantity *t*. The compositional operator $\mathbb{C}$ will be defined as

$$\mathbb{C}(\mathbf{z}) = (\frac{z_1}{\sum_{i=1}^{n+1} z_i},\ldots,\frac{z_n}{\sum_{i=1}^{n+1} z_i}).$$

It is important to note the one-to-one transformation between the basis space $\Re^{n+1}$ and the size-composition space $\Re_+^1 \times S^n$; i.e., any basis can be uniquely specified by the corresponding size and composition, and conversely.

2.4 Amalgamations and subcompositions

A three-dimensional graph can be used to represent a composition consisting of three elements. However, in representing higher-dimensional compositional data, it may be desirable to decrease its dimensionality while retaining the unit-sum constraint. For further motivation, consider the following.

Suppose that there exists a composition $\mathbf{x} \in S^n$, but that we are interested in only a subset of the elements. This is typical in compositional data analysis; for example, in example 2.1, we are interested in estimating the failure probability of the system, which is the sum of a subset of the elements of that composition. This example motivates the following definitions.

Definition 2.1 Let $\mathbf{x} = (x_1,...,x_n) \in S^n$, and let $1 \le n_1 < n_2 < \cdots < n_k \le n$. Then the vector $\boldsymbol{t} = (t_1, t_2, \cdots t_k)$ is said to be a (k-1)-dimensional **amalgamation** of $\mathbf{x}$ if $t_1 = x_1 + \cdots + x_{n_1}, t_2 = x_{n_1+1} + \cdots + x_{n_2}, \cdots, t_k = x_{n_k+1} + \cdots + x_{n+1}$.

Definition 2.2 Let $\mathbf{x} = (x_1,...,x_n) \in S^n$, and let $x_{n+1} = 1 - x_1 - \cdots - x_n$. Consider the subvector $\mathbf{z} = (x_{i_1}, x_{i_2}, ..., x_{i_k})$ where k<n+1, i_j is some integer $\in \{1,...,n+1\}$ for all $j \in \{1,..., k\}$, and $i_j \ne i_r$ for all $j \ne r$. Then the vector

$$\boldsymbol{s} = \mathbb{C}(\mathbf{z})$$

is called a (k-1)-dimensional **subcomposition** of $\mathbf{x}$.

Definition 2.3 Let $0 = a_0 < a_1 < a_2 < \cdots < a_k < a_{k+1} = n+1$, $\mathbf{x} = (x_1,...,x_n) \in S^n$, $x_{n+1} = 1 - x_1 - \cdots - x_n$, and $\boldsymbol{t} = (t_1, t_2, \cdots t_{k+1})$, where $t_i = x_{a_{i-1}+1} + \cdots + x_{a_i}$, for $i = 1,2,\cdots, k+1$. Then the set $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_{k+1}, \boldsymbol{t}\}$, where $\boldsymbol{s}_i = \mathbb{C}(x_{a_{i-1}+1}, \cdots, x_{a_i})$, for $i = 1,2,\cdots, k+1$, is said to form a **partition of order *k***.

2.5 The Dirichlet distribution

The Dirichlet distribution is the most familiar of the simplex distributions. In this section, we will define the Dirichlet distribution, give some related theorems by Mosimann (1962),

and discuss some properties of this distribution that make it appropriate only for modeling compositional vectors that exhibit forms of extreme independence as well as negative covariance.

Definition 2.4 The random vector $\mathbf{y} \in S^n$, is said to have a ***Dirichlet distribution*** if its density function is given by

$$g(y_1,...,y_n) = \Gamma(\alpha^*)\prod_{i=1}^{n+1}\frac{y_i^{\alpha_i - 1}}{\Gamma(\alpha_i)},$$

where $\alpha^* = \sum_{i=1}^{n+1}\alpha_i$, $y_{n+1} = 1 - y_1 - ... - y_n$ and $\alpha_i > 0 \quad i = 1,...,n+1$.

As noted by Fang et al. (1990), if $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_n,\alpha_{n+1})$ is the parameter vector of a Dirichlet distribution, then this distribution can be represented either as a distribution on the hyperplane $H_{n+1} = \{(y_1,...,y_n,y_{n+1}) | \sum_{i=1}^{n+1} y_i = 1\}$ in $\Re_+^{n+1}$. (In which case we write $\mathbf{y} \sim D_{n+1}(\boldsymbol{\alpha})$ on H_{n+1}), or as a distribution inside the simplex S^n in $\Re_+^n$ written $\mathbf{y} \sim D_n(\alpha_1,...,\alpha_n;\alpha_{n+1})$ on S^n.

Mosimann (1962) gives the following two theorems regarding the Dirichlet distribution:

Theorem 2.1 *Let $z_1, z_2, ..., z_{n+1}$ be independently distributed gamma random variables with location parameters α_i and equal scale parameters. Then the composition formed from this basis has a Dirichlet distribution that is independent of its size.*

Theorem 2.2 *Let $z_1, z_2, ..., z_{n+1}$ be a set of independent variables. Then $t = \sum_{i=1}^{n+1} z_i$ and $\mathbf{y} = (z_1/t, ..., z_n/t)$ are independent if and only if the z_i have gamma distributions with the same scale parameters.*

This is a generalization of a result due to Lukacs (1955) for the two-dimensional case. Note, however, composition of a basis can be Dirichlet distributed and independent of the size even if the elements of the basis are dependent, as pointed out by Smith (1994). Such basis have so-called Liouville distributions (Fang et al., 1990), which will be discussed later in this dissertation.

As stated above, the Dirichlet distribution has two properties that make it a poor distribution for modeling compositional data (Fang et al., 1990). First, the covariances are strictly negative. In particular, if $\mathbf{y} \in S^n$ have a Dirichlet distribution, then we have

$$\operatorname{cov}(y_i, y_j) = \frac{-\alpha_i \alpha_j}{(\alpha_i + \alpha_j)^2(\alpha_i + \alpha_j + 1)} \quad \text{for all } i \neq j .$$

Because of the unit-sum constraint given in (2.1), compositional data will often tend to be negatively correlated, but some data sets can nonetheless exhibit positive correlations between particular variables (see for example Reese and Krzysztofowicz, 1991). Thus, to model compositional data successfully, distributions on the simplex should be able to model both positive and negative covariance.

Second, the Dirichlet distribution has very strong conditional independence assumptions. In particular, if $\mathbf{y} \in S^n$ has a Dirichlet distribution and $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_k, \boldsymbol{t}\}$ is a partition of order k-1 as in definition 2.3, then $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_k, \boldsymbol{t}\}$ is an independent set. This result is independent of how the $y_1, \cdots, y_{n+1}$ are ordered, where $y_{n+1} = 1 - y_1 - \cdots - y_n$. Hence, the Dirichlet distribution has very strong independence properties. Aitchison (1982) goes so far as to say that "the Dirichlet class has so much independence structure built into its definition that it represents not a convenient modeling class for compositional data but the ultimate in

independence hypothesis." Thus, to improve on the Dirichlet, a successful simplex distribution should have more general correlation structures and weaker independence properties, as stated in chapter 1.

2.6 Dependence and independence concepts

Surprisingly, since compositional data arise very frequently in practice, Aitchison (1982) was the first to systematically define and study alternative forms of independence for distributions on the simplex. More recently, Smith (1994) in his Ph.D. thesis examined various concepts of independence that have been proposed for compositional data. In this section, some of the independence concepts that have been proposed to date will be summarized.

Aitchison (1986) stated that for compositional data, "there must be at least one negative element on each row of the crude covariance matrix." He proved that as follows. Let $\mathbf{x}=(x_1,...,x_n)\in S^n$, and let $x_{n+1}=1-x_1-\cdots-x_n$. Since $\text{cov}(x_i,\sum_{j=1}^{n+1}x_j)=0,$ we must have

$$\sum_{j\neq i}\text{cov}(x_i,x_j)=-\text{var}(x_i) \qquad (2.5)$$

Clearly the right-hand side of (2.5) can never be positive, so there must be at least one negative covariance on the left-hand side.

Due to the unit-sum constraint on compositional data, its correlation structure is difficult to interpret. This also makes the usual concept of independence meaningless; the elements of a composition absolutely cannot be independent, since their sum must equal one.

Realizing this, Aitchison (1982) suggested looking at dependence among subcompositions and amalgamations. Let $\mathbf{x} = (x_1,...,x_n) \in S^n$, and suppose that we are interested in breaking $\mathbf{x}$ into two parts: $\mathbf{x}_1 = (x_1,...,x_d)$; and $\mathbf{x}_2 = (x_{d+1},...,x_n, 1-\sum_{i=1}^{n} x_i)$. A partition of order one would consist of the set $\{\boldsymbol{s}_1, \boldsymbol{s}_2, t\}$, where $\boldsymbol{s}_1 = \left(\frac{x_1}{\sum_{i=1}^{d} x_i},..., \frac{x_{d-1}}{\sum_{i=1}^{d} x_i} \right)$, $\boldsymbol{s}_2 = \left(\frac{x_{d+1}}{1-\sum_{i=1}^{d} x_i},..., \frac{x_n}{1-\sum_{i=1}^{d} x_i} \right)$, and $t = \sum_{i=1}^{d} x_i$.

Taking $\perp$ to mean "statistically independent," Aitchison (1982, 1986) defines the following independence concepts for a partition of order 1:

1) Partition independence	$(\boldsymbol{s}_1, \boldsymbol{s}_2, t)$ mutually independent
2) Subcompositional invariance	$(\boldsymbol{s}_1, \boldsymbol{s}_2) \perp t$
3) Right neutrality	$(\boldsymbol{s}_1, t) \perp \boldsymbol{s}_2$
4) Left neutrality	$\boldsymbol{s}_1 \perp (\boldsymbol{s}_2, t)$
5) First subcompositional invariance	$\boldsymbol{s}_1 \perp t$
6) Second subcompositional invariance	$\boldsymbol{s}_2 \perp t$
7) Subcompositional independence	$\boldsymbol{s}_1 \perp \boldsymbol{s}_2$

Note that any partition of order 1 from the Dirichlet distribution satisfies all of these forms of independence. Smith (1994) points out several obvious relationships among these concepts:

i) (1) $\Rightarrow$ (2), (3), and (4).

ii) (4) $\Rightarrow$ (5) and (7).

iii) (3) $\Rightarrow$ (6) and (7).

iv) (2) $\Rightarrow$ (5) and (6).

Connor and Mosimann (1969) say that x_j is *neutral* with respect to x_i if $x_j \perp \frac{x_i}{1-x_j}$. They also give a vector generalization of this definition. For instance, let $(x_1,...,x_n)$ be a random vector on S^n. Then x_1 is said to be neutral with respect to $(x_2,...,x_n)$ if $x_1 \perp (\frac{x_2}{1-x_1},\frac{x_3}{1-x_1},...,\frac{x_n}{1-x_1})$. Similarly, $(x_1,...,x_d)$ is said to be neutral with respect to $(x_{d+1},...,x_n)$ if $(x_1,...,x_d) \perp (\frac{x_{d+1}}{1-x_1-\cdots-x_d},...,\frac{x_n}{1-x_1-\cdots-x_d})$. Finally, Connor and Mosimann say that a compositional vector $(x_1,...,x_n)$ has *complete neutrality* if and only if

$$x_1 \perp (\frac{x_2}{1-x_1},...,\frac{x_n}{1-x_1})$$

$$(x_1,x_2) \perp (\frac{x_3}{1-x_1-x_2},...,\frac{x_n}{1-x_1-x_2})$$

$$\vdots$$

$$(x_1,...,x_{n-1}) \perp \frac{x_n}{1-x_1-\cdots-x_{n-1}}.$$

Smith (1994) noted that, since $(x_1,...,x_d)$ is equivalent to $(\boldsymbol{s}_1,t)$ and $(\frac{x_{d+1}}{1-x_1-\cdots-x_d},...,\frac{x_n}{1-x_1-\cdots-x_d})$ is $\boldsymbol{s}_2$, then neutrality of $(x_1,...,x_d)$ with respect to $(x_{d+1},...,x_n)$ is equivalent to right neutrality; i.e., Connor and Mosimann's concept of neutrality is equivalent to Aitchison's concept of right neutrality. Thus, Smith refers to Connor and Mosimann's concept of complete neutrality as *complete right neutrality*.

Connor and Mosimann introduce complete neutrality as an analogy to complete independence in unconstrained data. The analogy is not perfect, however, since a composition can be completely neutral for some permutations of its elements but not others. Because of this, extreme neutrality defined below as a neutrality concept that does not depend on order.

A compositional data set is said to have *extreme neutrality* if it has complete neutrality for any permutation of its elements. This form of independence has been frequently discussed, but was apparently not given a name in the literature until Smith (1994). According to Connor and Mosimann (1969), W. Kruskal showed that extreme neutrality is essentially equivalent to the Dirichlet distribution. Kruskal apparently never published his proof, but Smith (1994) proved this property using published results.

As stated earlier, Connor and Mosimann's concept of neutrality is equivalent to Aitchison's concept of right neutrality, and Smith (1994) refers to complete neutrality as *complete right neutrality*. Aitchison (1986) expands the list of complete independence concepts to include complete partition independence, complete left neutrality, complete subcompositional invariance, complete subcompositional independence, complete first subcompositional invariance, and complete second subcompositional invariance, as follows: Consider an n-dimensional composition **x**. The *partition at level c* of **x** is simply the partition of order 1 formed from the single division $(x_1,\cdots,x_c;x_{c+1},\cdots,x_{n+1})$. Then let I be an independence property, such as right neutrality or subcompositional invariance, associated with a partition of order 1. Let I_k denote the presence of that independence property for a

partition of order 1 at level k. Then the composition **x** is said to have independence property I *up to level c*, written I^c, if I_k holds for k=1,....,c. The composition **x** is said to have the *complete* independence property I if I_k holds for partitions at all possible levels; i.e., if the composition **x** has independence property I^{n+1}.

Smith (1994) notes that complete subcompositional independence has been inconsistently defined in the literature, so he uses the term "extreme subcompositional independence" to refer to the stronger form of Aitchison's "subcompositional independence" (Aitchison, 1986, p. 235-236). Thus, in Smith's terminology, a compositional vector is said to have "extreme subcompositional independence" if the subcompositions $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_{k+1}\}$ formed from a partition of any order 1, 2, 3,...,etc. form an independent set. I would note that Smith used the term "extreme" in his definition of "extreme neutrality" to mean that a compositional vector set has complete neutrality for any permutation of its elements. However, he later used "extreme" in the definition of "extreme subcompositional independence" to mean that the subcompositions $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_{k+1}\}$ formed from a partition of any order (1, 2, 3,...,etc.) constitute an independent set. Note that "extreme subcompositional independence" is defined for a particular permutation. Thus, the term "extreme" is used in two different senses. A suggested change in terminology to avoid this problem will be proposed in the next section.

The Dirichlet distribution has every form of independence that has been defined in this section, and in fact Aitchison (1982) calls the Dirichlet distribution "the ultimate in independence hypothesis." Moreover, since extreme neutrality is achieved only by the

Dirichlet distribution, extreme neutrality is in some sense the strongest possible independence concept for compositional data.

We now introduce an independence concept that involves both the basis and its resulting composition. In fact, since a basis completely determines both its composition and its size it is natural to ask whether the size of the basis and the composition are independent. Aitchison (1986) says that a basis has *compositional invariance* if this property holds. Note from theorems (2.1) and (2.2) above that if **z** is an (n+1)-element basis and z_i, $i = 1,\ldots,n+1$, are independent gamma random variables with equal scale parameters, then **z** has compositional invariance. Furthermore, the composition **x**=$\mathbb{C}$(**z**) has a Dirichlet distribution in this case.

2.7 Some new independence concepts

From the above discussion, one should note that all of the independence concepts discussed here have been defined for a partition of order one, except for extreme subcompositional independence, which was defined for a partition of any order (1, 2, 3,...,etc.). These concepts are used as measures of independence for compositional data. Thus, one interesting question is whether it is beneficial in applications to extend the independence concepts to partitions of higher order. In what follows, we extend some concepts of independence to partitions of higher order, and also propose some new terminology.

Throughout this section we will use the term "complete" to denote independence at any possible level for a specific permutation, while the term "extreme" will be used to denote complete independence for any permutation.

Consider a partition $\{\mathbf{s}_1,\mathbf{s}_2,\ldots,\mathbf{s}_{k+1},\mathbf{t}\}$ of order k of the compositional vector $\mathbf{x}\in S^n$ as in definition 2.3. We say that a compositional vector **x** has *complete k-partition independence* if $\{\mathbf{s}_1,\mathbf{s}_2,\ldots,\mathbf{s}_{k+1},\mathbf{t}\}$ is an independent set of vectors for a particular permutation of the composition **x**, but for partitions of any order up to and including k. Analogous to the definition of extreme neutrality, we will say that a compositional vector **x** has *extreme k-partition independence* if it has complete k-partition independence for any permutation of its elements, with "extreme" denoting any permutation, and "k" denoting a partition of any order up to and including k .

From section 2.5, we know that if **x** has a Dirichlet distribution, then $\{\boldsymbol{s}_1,\boldsymbol{s}_2,\ldots,\boldsymbol{s}_{k+1},\boldsymbol{t}\}$ will be an independent set of vectors for any $k\in\{1,2,\ldots,n\}$, regardless of how the elements of **x** are ordered. This form of independence has been frequently discussed, but was apparently not given a name in the literature. The above definition suggests *extreme n-partition independence* as a name for this independence concept. Note that if a compositional vector **x** has complete (extreme) k-partition independence, then it must have complete (extreme) r-partition independence for any $r\le k$. By convention, complete 1-partition independence is equivalent to complete partition independence; in general, if a prefix is not affixed to an independence concept, then it should be understood that we are talking about a partition of order one.

Similarly, we say that a compositional vector has *complete k-subcompositional independence* if the subcompositions $\{\mathbf{s}_1,\mathbf{s}_2,\ldots,\mathbf{s}_{k+1}\}$ is an independent set of vectors for a particular permutation of the composition **x**, but for partitions of any order up to and including k. Also, we will say that a compositional vector **x** has *extreme k-subcompositional*

independence if it has complete k-subcompositional independence for any permutation of its elements.

Note that if k=n, we get complete n-subcompositional independence, which is equivalent to what Smith (1994) inconsistently called extreme subcompositional independence (the Aitchison's stronger form of "subcompositional independence"). Thus, throughout this thesis complete n-subcompositional independence will be used to refer to Aitchison's stronger form of "subcompositional independence". Once again, by convention, complete 1-subcompositional independence is equivalent to complete subcompositional independence.

It is clear from the above that subcompositional independence and partition independence have now been defined for higher-order partitions, the first by Aitchison (1986), and the second one in this dissertation. Analogous to that, could one do the same thing for other independence concepts? Are such extensions beneficial? Have such generalized independence concepts been discussed in the literature? The answers to these questions are left for the interested researcher.