Chapter 5

Adaptive Dirichlet distributions with dependent ratios

5.1 Introduction

Krzysztofowicz and Reese (1993) modeled a vector of fractions $\mathbf{x} = (x_1, \cdots, x_n, x_{n+1})$ via a sequence of random variables $y_1, \cdots, y_n$, called ratios of fractions, with $0<y_i<1$ for all *i* as shown in Figures 3.1-3.3. The mutual independence of these ratios was one of the main assumptions that led to the adaptive Dirichlet distributions that they propose. In this chapter we will relax the assumption of independence and propose a new model for compositional data via a method that has become popular in recent years, copulas. This method allows us to construct joint densities with arbitrary marginal distributions and arbitrary correlation coefficients. Moreover, the correlation coefficient remains unchanged even with changes in the marginal distribution functions.

As discussed in section 3.4, let T be the one-to-one transformation from the vector of ratios $\mathbf{y} = (y_1, \cdots, y_n)$ to the compositional vector $\mathbf{x} = (x_1, \cdots, x_n) \in S^n$. Hence, given a joint density $g_0(\mathbf{y})$, the joint density $g(\mathbf{x})$ is specified by

$$g(\mathbf{x}) = g_0(y_1, \cdots, y_n) J(T),$$

where $J(T)$ is the Jacobian of the transformation *T*. Note that the ratios can have different dependence structures. For example, the joint density of ratios y_1, y_2, and y_3 can have one of the following forms: $g_0(y_1, y_2, y_3)$; $g_1(y_1) g_{23}(y_2, y_3)$; $g_2(y_2) g_{13}(y_1, y_3)$; or $g_3(y_3) g_{12}(y_1, y_2)$, where g_0, g_{23}, g_{13}, and g_{12} are multivariate densities and g_1, g_2, and

g_3 are univariate densities. The multivariate densities g_0, g_{23}, g_{13}, and g_{12} can be constructed in several ways, including by means of copulas, as mentioned above.

Section 5.2 gives a motivation for our new model via some examples, as well as via a comparison between the correlation sign structures resulting from adaptive Dirichlet distributions with independent and dependent ratios. Section 5.3 will introduce our approach for achieving dependent ratios by means of bivariate copulas.

5.2 Motivation for our new model

5.2.1 Examples

a) 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 medical 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, none of the existing adaptive Dirichlet distributions can model this kind of problem. Along the same lines, one could also consider lung cancer and bronchitis, since it is known that both diseases are caused by smoking.

b) Snowmelt runoff problem:

Reese and Krzysztofowicz (1991) applied the adaptive Dirichlet distribution to snowmelt runoff. In particular, they analyzed monthly snow runoff data for 14 western United States rivers. They were interested particularly in the fraction of total annual runoff occurring in

each month. In describing the statistical nature of snowmelt runoff patterns, they found that the correlations among these fractions were a reasonable way to characterize the seasonal patterns at particular rivers. Reese and Krzysztofowicz hypothesized that the vector of fractions was generated from an adaptive Dirichlet distribution, and tests of that assumption were performed. They found the model valid for about one-half to two-thirds of the rivers tested, and therefore concluded that the adaptive Dirichlet distribution was useful, but not sufficiently general, because it was restricted by two structural assumptions: first, that the composition is stochastically independent of the size of the basis (i.e., the total seasonal runoff); and second, that the ratios are mutually independent. In fact, one or both of these assumptions are violated by some rivers. Thus, they dispute the modelers to construct multivariate distributions on the simplex that could fit the remaining rivers whose seasonal runoff patterns exhibit different dependence structures than can be represented using their model. In chapter 8 of this proposal, we will discuss the latter problem in more detail and outline our proposed approach for resolving it.

5.2.2 Correlation sign structure

Table 5-1 below, given by Krzysztofowicz and Reese (1991), shows the correlation sign structures obtainable from the topology in Figure 5.1 in the case where the ratios y_1, y_2, and y_3 are mutually independent. Table 5-2 shows the correlation sign structures obtainable from the same topology when $\mathrm{cov}(y_1, y_3)>0$. In addition, the sign structure

$$\begin{matrix} + & + & - \\ & - & + \\ & & + \end{matrix}$$

may also be achievable, but we have not been able to either confirm or refute this. It is clear from these tables that the adaptive Dirichlet distribution that would result

from Figure 5.1 with dependent ratios is an improvement over the one that would result from the same figure with independent ratios, in the sense that a much broader range of correlation sign structures can be obtained.

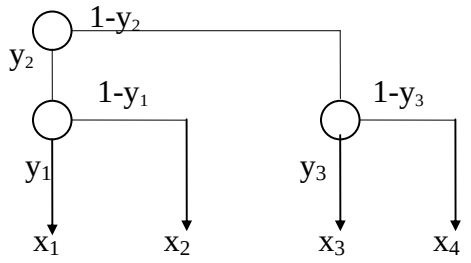

Figure 5.1 Double-cascaded bifurcation topology of four fractions (taken from Krzysztofowicz and Reese, 1991).

Table 5-1

Correlation sign structures obtainable from Figure 5.1

when the y_i, i=1,2,3, are independent

+	−	−	−	−	−	+	−	−	−	−	−
	−	−		−	−		−	−		−	−
		+			+			−			−

Table 5-2

Correlation sign structures known to be obtainable from Figure 5.1

when y_2 and (y_1, y_3) are independent and cov(y_1,y_3) > 0

+	−	−	−	+	−	+	−	−	+	+	−	+	+	−		
	−	+		−	+		−	+		−	−		−	+		
		−			+			+			+			−		
+	+	−	−	−	−	−	+	−	+	−	−	−	+	−		
	−	−		−	+		−	+		−	−		−	−		
		−			+			−			+			+		
−	−	−	+	−	−	−	+	−	−	−	−	−	−	−		
	−	−		−	−		−	−		−	+		−	−		
		+			−			−			−			−		

We have

$x_1 = y_1 y_2$,

$x_2 = (1 - y_1) y_2$,

$x_3 = (1 - y_2) y_3$, and

$x_4 = (1 - y_2)(1 - y_3)$. (5.1)

This system of four equations has a unique inverse:

$y_1 = \frac{x_1}{x_1 + x_2}$,

$y_2 = x_1 + x_2$,

$y_3 = \frac{x_3}{x_3 + x_4}$, and

$x_1 + x_2 + x_3 + x_4 = 1$.

Assuming that y_1 and y_3 are correlated, but independent of y_2 and using equations (5.1) above, we get

$$\operatorname{cov}(x_1, x_2) = \mathrm{E}(y_1(1-y_1))\mathrm{E}(y_2^2) - \mathrm{E}(y_1)\mathrm{E}(1-y_1)\mathrm{E}(y_2)^2$$

$$\operatorname{cov}(x_1, x_3) = \mathrm{E}(y_2(1-y_2))\mathrm{E}(y_1y_3) - \mathrm{E}(y_2)\mathrm{E}(1-y_2)\mathrm{E}(y_1)\mathrm{E}(y_3)$$

$$= \mathrm{E}(y_2(1-y_2))\operatorname{cov}(y_1, y_3) - \mathrm{E}(y_1)\mathrm{E}(y_3)\operatorname{var}(y_2)$$

$$\operatorname{cov}(x_1, x_4) = \mathrm{E}(y_1(1-y_3))\mathrm{E}(y_2(1-y_2)) - \mathrm{E}(y_1)\mathrm{E}(1-y_3)\mathrm{E}(y_2)\mathrm{E}(1-y_2)$$

$$= -\operatorname{cov}(x_1, x_3) - \mathrm{E}(y_1)\operatorname{var}(y_2)$$

$$\operatorname{cov}(x_2, x_3) = \mathrm{E}(y_2(1-y_2))\mathrm{E}(y_3(1-y_1)) - \mathrm{E}(y_2)\mathrm{E}(1-y_2)\mathrm{E}(y_3)\mathrm{E}(1-y_1)$$

$$= -\operatorname{cov}(x_1, x_3) - \mathrm{E}(y_3)\operatorname{var}(y_2)$$

$$\operatorname{cov}(x_2, x_4) = \mathrm{E}((1-y_1)(1-y_3))\mathrm{E}(y_2(1-y_2)) - \mathrm{E}(1-y_1)\mathrm{E}(1-y_3)\mathrm{E}(y_2)\mathrm{E}(1-y_2)$$

$$= \operatorname{cov}(x_1, x_3) - (1-\mathrm{E}(y_1) - \mathrm{E}(y_3))\operatorname{var}(y_2)$$

$$\operatorname{cov}(x_3, x_4) = \mathrm{E}(y_3(1-y_3))\mathrm{E}((1-y_2)^2) - \mathrm{E}(y_3)\mathrm{E}(1-y_3)\,\mathrm{E}(1-y_2)^2 \qquad (5.2)$$

Utilizing the above equations, one can show analytically that if $\operatorname{cov}(y_1,y_3) > 0$, then the general matrix of signs must be of the form

$$\begin{matrix} \pm & \pm & - \\ & - & \pm \\ & & \pm \end{matrix}$$

We have demonstrated by example (Table B-0) that 15 of the 16 possible sign structures (those shown in Table 5-2) are in fact achievable. (As mentioned earlier, we have not been able to confirm whether the 16th case is achievable.).

If $\operatorname{cov}(y_1,y_3) = 0$, then we have

$$\operatorname{cov}(x_1, x_3) = -\mathrm{E}(y_1)\mathrm{E}(y_3)\operatorname{var}(y_2) < 0\,,$$

$$\operatorname{cov}(x_2, x_3) = -\mathrm{E}(y_3)(1-\mathrm{E}(y_1))\operatorname{var}(y_2) < 0$$

$\operatorname{cov}(x_1, x_4) = -\mathrm{E}(y_1)(1-\mathrm{E}(y_3))\operatorname{var}(y_2) < 0$, and

$$\operatorname{cov}(x_2, x_4) = -(1-\mathrm{E}(y_1))(1-\mathrm{E}(y_3))\operatorname{var}(y_2) < 0. \qquad (5.3)$$

Thus, in this case we just get Table 5-1, confirming the results given by Krzysztofowicz and Reese.

Finally, if $\operatorname{cov}(y_1,y_3) < 0$ (i.e., $\mathrm{E}(y_1y_3) < \mathrm{E}(y_1)\mathrm{E}(y_3)$), then the general matrix of signs must be of the form

$$\begin{matrix} \pm & - & \pm \\ & \pm & - \\ & & \pm \end{matrix}$$

Once again, we have been able to demonstrate by example that many of these sign structures are achievable, but are unsure whether all 16 cases are possible.

5.3 Copula model for achieving dependent ratios

5.3.1 Introduction

Copulas are functions that can be used to specify multivariate distributions with any specific set of univariate marginal distributions. For the 2-dimensional case, if F(x) and G(y) are cumulative distribution functions (cdf's) for random variables X and Y, then there exists a 2-dimensional copula C such that the multivariate cumulative distribution H(x,y) can be expressed in the form

$$\mathrm{H}(x,y) = \mathrm{C}(\mathrm{F}(x),\mathrm{G}(y)) \qquad (5.4)$$

for all x and y in $\Re$ (Schweizer and Sklar, 1983). Hence, the function C has domain $[0,1]^2$ and range [0,1]. Suppose that X (respectively, Y) is beta distributed with parameters α_1 and β_1 (respectively, α_2 and β_2). Then, the joint probability density function satisfies

$$\begin{aligned} \mathrm{h}(x,y) &= \frac{\partial^2}{\partial x \partial y}\mathrm{H}(x,y) \\ &= \mathrm{c}(\mathrm{F}(x),\mathrm{G}(y))\mathrm{f}(x)\mathrm{g}(y) \\ &\propto \mathrm{c}(\mathrm{F}(x),\mathrm{G}(y))x^{\alpha_1-1}(1-x)^{\beta_1-1}y^{\alpha_2-1}(1-y)^{\beta_2-1}, \end{aligned} \tag{5.5}$$

where $\mathrm{c}(s,t) = \partial^2 \mathrm{C}(s,t)/\partial s \partial t$ is the copula density function.

Copulas allow us to assess joint distributions with particular values for the marginal moments, because the marginals of H(x,y) are simply equal to F(x) and G(y), so the marginal moments of the joint distribution H(x,y) are equivalent to those of F(x) and G(y). Also, some copulas are capable of representing dependence structures ranging from perfect negative to perfect positive correlation.

In the next section, we use Frank's copula (Frank, 1979; Genest, 1987) to show the feasibility of achieving dependent rather than independent ratios in Krzysztofowicz and Reese's model. However, in the literature, there are also many other kinds of bivariate copulas, such as these proposed by Gumbel (1961), Cook and Johnson (1981), and Ali, Mikhail, and Haq (1978).

Two important issues in choosing the copula C are: 1) the ability to describe strong positive or negative correlations; and 2) simplicity of computations. For the application that will be given in chapter 8, a suitable range of correlation coefficients is more important than

computational convenience. Therefore, we use Frank's copula here since it allows a full range of correlations from -1 to 1 (Frank, 1979; Genest, 1987; Yi, 1997).

5.3.2 An approach using Frank's copula

This section introduces a specific method of constructing generalized adaptive Dirichlet distributions with dependent ratios. Dependency among the ratios is represented by means of Frank's copula (Frank, 1979; Genest 1987). Some of the properties of Frank's copula are summarized.

The class of bivariate distributions of the form

$$P(X \le x, Y \le y) = H_\alpha(x, y) = \log_\alpha \left\{ 1 + \frac{(\alpha^x - 1)(\alpha^y - 1)}{\alpha - 1} \right\} \quad (\alpha \neq 1), \tag{5.6}$$

where X and Y are uniformly distributed over [0,1] and $\log_\alpha(t)$ denotes logarithm to the base $\alpha>0$, was discovered by Frank (1979). Each member of the family (5.6) is absolutely continuous and has full support over the unit square except when $\alpha=0$ or ∞, in which case $Y=X$ or $Y=1-X$, respectively. When $0<\alpha<\infty$, the density corresponding to H_α is given by

$$h_\alpha(x, y) = \frac{(\alpha - 1)\log(\alpha)\alpha^{x+y}}{\{(\alpha - 1) + (\alpha^x - 1)(\alpha^y - 1)\}^2} \quad (0 < x, y < 1), \tag{5.7}$$

where $h_\alpha(x, y)$ tends to 1 $\forall x, y$ as α approaches 1 (corresponding to the case where x and y are uncorrelated).

It is possible to construct families of bivariate distributions with non-uniform marginals by means of the method of translation (Nataf, 1962), as Genest (1987) has noted. The general form of the resulting cumulative distribution function (cdf) is given by

$$H_\alpha(x,y) = \log_\alpha\left\{1+\frac{(\alpha^{F(x)}-1)(\alpha^{G(y)}-1)}{\alpha-1}\right\} \quad (\alpha\neq 1), \tag{5.8}$$

with arbitrary marginals $F(x)$ and $G(y)$. Depending on the choice of the parameter α, these distributions allow for highly positive or negative correlation between X and Y.

Consider the topology given in Figure 5.1. Assume that y_1 depends on y_3 and that both are independent of y_2. Hence, given a joint density $g_{13}(y_1,y_3)$ and a density $g_2(y_2)$, the joint density $g(\mathbf{x})$ would be given by

$$g(\mathbf{x}) = g_2(y_2)g_{13}(y_1,y_3)J(\mathbf{y}\rightarrow\mathbf{x}), \tag{5.9}$$

where $J(\mathbf{y}\rightarrow\mathbf{x}) = \left(\frac{1}{x_1+x_2}\right)\left(\frac{1}{x_3+x_4}\right)$ is the Jacobian of the transformation from $\mathbf{y}\rightarrow\mathbf{x}$.

Lett $g_2(y_2)$ be a beta density and

$$g_{13}(y_1,y_3) = \frac{(\delta-1)\log(\delta)\delta^{G_1(y_1)+G_3(y_3)}}{\{(\delta-1)+(\delta^{G_1(y_1)}-1)(\delta^{G_3(y_3)}-1)\}^2}\, g_1(y_1)g_3(y_3) \tag{5.10}$$

(i.e., let $g_{13}(y_1,y_3)$ be one of Frank's family of distributions), where the marginals $G_1(y_1)$ and $G_3(y_3)$ are the cumulative distribution functions (cdf's) of beta distributions (i.e., $y_i \sim Be(\alpha_i,\beta_i)$, i=1,2,3). Then the joint density of $\mathbf{x}$ is given by

$$g(\mathbf{x}) = \prod_{i=1}^{3}\frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}\frac{(\alpha-1)\log(\alpha)\alpha^{G_1(\frac{x_1}{x_1+x_2})+G_3(\frac{x_3}{x_3+x_4})}}{\{(\alpha-1)+(\alpha^{G_1(\frac{x_1}{x_1+x_2})}-1)(\alpha^{G_3(\frac{x_3}{x_3+x_4})}-1)\}^2}$$

$$x_1^{\alpha_1-1}x_2^{\beta_1-1}x_3^{\alpha_3-1}x_4^{\beta_3-1}(x_1+x_2)^{\alpha_2-\alpha_1-\beta_1}(x_3+x_4)^{\beta_2-\alpha_3-\beta_3}. \tag{5.11}$$

As one can see, this joint density function is closed form (except for the dependence on G_1 and G_3), but not very simple or computationally tractable.