Chapter 10

Conclusion and suggestions for future research

10.1 Conclusion

In this work, our primary interest has been theoretical, but we also considered applications of our results to two different real-world problems. In particular, in this dissertation, we developed new families of distributions on the unit simplex and the positive orthant, proved a theorem that helps in identifying classes of distributions that satisfy the conditions for perfect aggregation in Bayesian reliability analysis, and applied the adaptive Dirichlet distribution with dependent ratios (one of our new families) to the problem of snowmelt runoff.

In chapter 1, we mentioned many fields in which compositional data arise, and noted that due to the unit sum constraint, this data is difficult to analyze correctly. Also, we demonstrated the need to generalize the set of existing unit simplex distributions. In particular, existing distributions are insufficient to model many compositional data sets, due to the strong independence structures built into their definitions and their inability to model compositional data with general covariance structures. 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 chapter 2, we introduced some notation for compositional data and gave some examples. We also presented some relevant background about concepts of dependence from Smith (1994).

In chapter 3, we surveyed the known classes of unit simplex distributions, and showed that the adaptive Dirichlet distribution (Krzysztofowicz and Reese, 1993) is a subclass of the multiple Dickey family of distributions (Dickey, 1983 and 1987). We devoted chapter 4 to the standard Liouville distribution and its known generalizations on the positive orthant. We defined these distributions using the approach of Fang et al. (1990), which is helpful for investigating the covariance structures of Liouville random vectors. We showed that the possible sign structures obtainable using the Liouville distribution are very restrictive. Gupta and Richards (1987, 1991, 1992, and 1995) had previously been unable to determine the covariance structures obtainable using this distribution.

In chapter 5, we introduced the adaptive Dirichlet distribution with dependent ratios, where dependency among the ratios is represented by means of a joint distribution such as Frank's copula. The correlation matrix of this distribution is shown to be much less restrictive than the correlation matrix of the adaptive Dirichlet distribution with independent ratios developed by Krzysztofowicz and Reese (1993). However, its joint density function is not as simple or computationally tractable.

In chapter 6, we developed a new family of distributions on the positive orthant $\Re_{+}^{n}$ that encompasses as special cases the Liouville, adaptive Dirichlet with independent ratios, and multiple Dickey distributions. This new family of distributions (which we named the multiple Dickey extended Liouville distribution) gives us more flexibility in modeling correlated data on the positive orthant, since it has significantly more general covariance and dependence structures than the standard Liouville distribution. We gave closed-form expressions for the means, variances, covariances, and correlation sign structures of two special types of multiple Dickey extended Liouville distributions, and showed that these correlation sign structures are much less restrictive than those of the standard Liouville distribution. We note here that four previous attempts to extend the standard Liouville distribution (Rayens and Srinivasan, 1994; Smith, 1994; Gupta and Richards, 1995; and Gupta and Song, 1996), the authors were unable to determine the covariance structures of the resulting distributions.

In chapter 7, we were concerned with identifying families of unit simplex distributions that satisfy perfect aggregation in Bayesian reliability analysis. In other words, we were interested in identifying joint prior distributions for the system state probabilities such that the posterior system failure probability obtained by updating one of these priors with component-level data would be the same as if we instead used system-level data. By means of our theorems 7.2 and 7.4, we were able to tell relatively easy if a given prior distribution satisfied perfect aggregation. Before this, there was no means to determine this easily. In this chapter, we were therefore able to identify many existing unit simplex distributions that satisfy perfect aggregation for both Bernoulli and Poisson systems. Also, we introduced two approaches for developing additional families of unit simplex distributions that facilitate the study of perfect aggregation. One of these approaches allowed us to construct a large number of new prior distributions that satisfy the conditions of theorem 7.2 for perfect aggregation.

In chapter 8, we discussed the application of the adaptive Dirichlet distribution with dependent ratios to a snowmelt runoff problem, where Krzysztofowicz and Reese's original models were not entirely satisfactory. In particular, Krzysztofowicz and Reese found their model valid for about one-half to two-thirds of the rivers tested, and therefore concluded that the adaptive Dirichlet distribution is useful but insufficient, because it is restricted by independence assumptions. In this dissertation, we reanalyzed this problem using the adaptive Dirichlet distribution with dependent ratios, as given in chapter 5. In this study, we fit the means and variances of the ratios, and the covariance of a pair of ratios. We then compared the means, variances, and correlation matrices of the fractions obtained from their model and ours with the empirical data as a reference point. Through this comparison, we found that the Euclidean distance between our model and the empirical correlations was less than the distance between Reese and Krzysztofowicz's model and the empirical correlations for all 14 rivers. However, our model still did not preserve the empirical correlation sign structures for those rivers where Reese and Krzysztofowicz’s model did not preserve the signs. In particular, even when the sign structure of a particular correlation matrix was in principle achievable by the dependent ratios model, it was not always possible to achieve the desired sign structure while also matching the observed moments of the data set. This may be due to limitations of the fitting method that was used.

In chapter 9, we performed a simulation study to gain more insight into the behavior of the adaptive Dirichlet distribution with dependent ratios. In particular, we generated simulated data sets from adaptive Dirichlet distributions with and without dependent ratios. Four of these data sets were generated under the assumption that all the ratios were independent, while in the remaining data sets, one pair of ratios was assumed to be correlated. In our study, we attempted to fit these data sets using both the independent ratios model and models where different ratios were assumed to be dependent. As in chapter 8, we fit the means and variances of the ratios, and the covariance of a pair of ratios. Our results show that, if the goal is to accurately model an observed correlation matrix (rather than to correctly detect the underlying process used to generate those correlations), the dependent ratios model appears to be a substantial improvement over the independent ratios model. Therefore, the added flexibility of the adaptive Dirichlet model with dependent ratios seems to be a significant advantage in accurately describing a wide range of possible correlation structures. However, some limitations in the method used to fit the dependent ratios model to the simulated data sets became apparent.

To summarize, this dissertation has significantly expanded the sets of existing distributions on both the unit simplex and the positive orthant. Previous authors, including Gupta and Richards (1987, 1990, 1991, 1992, and 1995) and Rayens and Srinivasan (1994), have pointed out the importance of knowing the sign structures achievable by a given model, but information on correlation sign structures was not available for many existing distributions. With respect to the criteria suggested in chapter 1, the distributions we have developed can represent both positive and negative correlations, allow reasonably general sign structures for their correlation matrices, and are therefore clearly much less restrictive than the Dirichlet distribution. In general, our new families are substantially more flexible than the distributions that existed before this work. Also, we were able to give closed-form expressions for their moments in a number of cases. However, future work is needed to develop better methods of parameter estimation, and to explore additional criteria suggested in chapter 1.

10.2 Suggestions for future research

Our suggestions for future research encompass three lines of research: analyzing our models for the ranges of dependence (as well as the types of dependence structures) that are covered; further extensions of the families of distributions developed in this thesis; and further applications of those distributions. With respect to the range of dependence achievable by a given model, we observed some cases in both the snowmelt runoff data and our simulation study in which adaptive Dirichlet models with dependent ratios were not able to reproduce a particular sign structure, even though that sign structure was in principle achievable by the model. This may be because the model was not able to reproduce the numerical ranges of the correlations represented in the data, especially while also matching other moments. Alternatively, it may be due to the limitations of the fitting method used in this dissertation. Further investigation of this issue could shed more light on the strengths and limitations of our new models.

With regard to analyzing our models for the types of dependence structures that are covered, in this distribution we focused primarily on the sign structures of the correlation matrix. Therefore, future work could explore additional criteria mentioned in chapter 1 (e.g., the property that compositions with a particular distribution also have subcompositions with the same distribution, natural conjugacy), and additional dependence concepts reviewed in chapter 2 (e.g., partition independence, right neutrality, subcompositional independence). Joe (1997), Gupta and Richards (1987, 1990, 1991, 1992, and 1995), and Fang et al. (1990) would be valuable references for such work.

With regard to possible extensions, it is worth noting that in modeling compositional data using the adaptive Dirichlet distribution with dependent ratios, exact results for all means, variances, and covariances can be obtained by allowing dependence among all the ratios (rather than only two). This can be done by using multivariate rather than bivariate copulas (see for example Mardia, 1970; Cuadras and Auge, 1981; MacKenzie, 1994; Joe, 1997). Investigating the relative merits of different families of multivariate distributions for this purpose is left for future research.

The idea of generalizing the adaptive Dirichlet distribution through the use of dependent ratios could also be applied to the multiple Dickey family of distributions. In particular, for those multiple Dickey distributions obtained from successively nested partitions (as discussed in chapter 3), one could impose dependence among the ratios, as was done for the adaptive Dirichlet distribution. Also, one could use a non-Dirichlet rather than a Dirichlet distribution for the parameters of the multinomial process underlying the multiple Dickey distribution.

We believe that further experience with applications of the new distributions developed here will yield greater insight into their strengths and weaknesses. Therefore, for example, we recommend applying the extended Liouville distribution discussed in chapter 6 to reanalyze the reliability problem that was originally analyzed by Gupta and Richards (1992) using the standard Liouville distribution. As discussed in section 3.1, the covariance matrix of the Liouville distribution has a very restrictive form; namely, the signs of the covariances must be either all positive or all negative. Therefore, the extended Liouville distribution may be more suitable than the standard Liouville in cases with mixed sign structures. For example, instead of assuming that a common environment affects all components in a similar manner (as was assumed by Gupta and Richards), there might be two types of components, such that certain aspects of the environment may be beneficial for one type of component but harmful for the other. As an example, one may think of a system in which some components are adversely affected by high temperature, while others are adversely affected by low temperature, so that the lifetimes of the two types of components are negatively correlated. Our model could be useful in such cases.

Similarly, it might be worthwhile to apply some of our new distributions to models of brand choice such as those studied by Peter (1993) and Queen et al. (1994). More flexible models might permit better representations of real-world brand choice data.

Finally, with regard to further applications of the adaptive Dirichlet distribution, Krzysztofowicz and Reese (1993) point out that for each value of *n* (*n*=2,3,….), there exists a set of possible bifurcation topologies and permutations of fractions, each of which generates a family of multivariate densities on the *n*-dimensional simplex. By considering this family of structural densities, Krzysztofowicz and Reese were able to identify bifurcation models with correlation structures "close" to the empirical correlation structures evidenced by the data for most of the rivers they analyzed. In our work, we considered the same topologies and permutations of fractions used by Reese and Krzysztofowicz for each river, but in fact once one allows dependence among the ratios, some other topology than the one they chose may turn out to be optimal. Therefore, reanalysis of their data using the best possible topology and permutation of fractions may yield more favorable results than those obtained in this thesis, providing further evidence of the usefulness of adaptive Dirichlet distributions with dependent ratios. Also, as mentioned earlier, the approach we used to fit the adaptive Dirichlet distribution to the observed data set was not ideally suited to reproducing the signs of the correlations in that data. Using other methods of fitting could shed more light on the strengths and limitations of our new models, and might result in substantially better performance.