9.1 Strategy of analysis

We now describe the procedure used to analyze the simulated data:

1. Using the parameters of the four independent beta random variates, z_1, z_2, z_3, and z_4, we calculate the true correlation matrices of both the fractions x_i and the ratios y_i.
2. Based on the topology given in Figure 9.1, we transform each simulated vector of fractions **x** into a vector of ratios **y** via equations (9.1). For the resulting vectors of ratios, we then calculate the empirical correlation matrix.
3. All empirical correlations of the y_i are tested to see whether they are statistically significant; the null hypothesis is that a given correlation coefficient ρ is equal to zero. The two-tailed test uses the Student-Fisher *t* distribution with *k*-2 degrees of freedom:

$$t = |\hat{\rho}| \left(\frac{k-2}{1-\hat{\rho}^2} \right), \tag{9.2}$$

where $\hat{\rho}$ is the empirical sample correlation coefficient and *k* is the sample size. The hypothesis that $\rho = 0$ is rejected at the significance level *p* if $t > t(p, k-2)$, a tabulated critical value (Hoshmand, 1988). We test this null hypothesis at three levels of significance: 0.05; 0.01; and 0.001.

4. From the simulated vectors of the ratios **y**, we estimate the mixed moments needed to compute the means, variances, and covariances of the fractions x_i according to equations (8.7-8.9). These moments are then used to compute the correlations of the fractions x_i for both the model with independent ratios and the three models with dependent ratios. This method of fitting is the same as the one that we used in the snowmelt runoff problem, and is also the same as that used by Reese and Krzysztofowicz (1989).

5. Finally, for each case, we compute the Euclidean distances, as given by (8.2), between the correlation matrix of the x_i corresponding to each fitted model and the corresponding empirical and true correlation matrices.

The results of our simulation are given in Appendices B-E.

9.2 Analysis and results of our simulation

Below we list the questions to be answered by our analysis of each simulated data set:

1. What are the two most strongly correlated ratios (empirically)?
2. Are the signs of the true (respectively, the empirical) correlations of the x_i preserved under the independent ratios model?
3. Are the signs of the true (respectively, the empirical) correlations of the x_i preserved under the dependent ratios model?
4. Is the Euclidean distance between the true (respectively, the empirical) correlations and the correlations of the dependent ratios model at least as small as the distance between the true (respectively, the empirical) correlations and the correlations of the independent ratios model?
5. Is the Euclidean distance of the dependent ratios model actually used to generate the data (for those data sets with non-zero true correlations) at least as small as the Euclidean distances of the other two dependent ratios models?
6. For those data sets in which the true correlations of the y_i are all zero (i.e., in Appendix E) is the reduction in the Euclidean distance obtained by the dependent ratios model smaller than for those data sets in which the y_i are assumed to be correlated?

The detailed answers to the above questions for the simulated data can be found in the tables in Appendices B-E.

The data for Tables B-1 to B-15 were generated assuming dependence between y_1 and y_3. Three of these (Tables B-1, B-3, and B-8) have the signs of the true correlations preserved under all four models. Eight of them (Tables B-2, B-4 to B-7, B-11, B-14 and B-15) have the signs of the true correlations preserved only under the model where y_1 and y_3 are dependent. In Tables B-10 and B-13, the signs of the true correlations are preserved only under the independent ratios model and the model where y_1 and y_2 are dependent. In Table B-12, the signs of the true correlations are preserved only under the independent ratios model and the model where y_2 and y_3 are dependent. In the remaining case (Table B-9), the signs of the true correlations are not preserved under any four of the models. Thus, the model where y_1 and y_3 are assumed to be dependent preserves the true correlation sign structures for eleven of the fifteen data sets, while the independent ratios model accomplishes this for only six of the data sets. When the signs of the empirical correlations (i.e., the correlations among the x_i in the simulated data) are taken to be the reference point in comparing the correlation sign structures (rather than the signs of the true correlations), the model where y_1 and y_3 are assumed to be dependent preserves the empirical correlation sign structures for twelve of the fifteen data sets, while the independent ratios model accomplishes this for only four of the data sets.

The results in Table B-12 and Table B-13 are not as expected. In particular, the model that was actually used to generate the data did not preserve the true sign structure, even though the independent ratios model did. Since the independent ratios model is actually just a

special case of the dependent ratios model, clearly the observed performance does not represent a true limitation of the dependent ratios model, but rather is due to the fitting method that was used (which was described above). A suggested alternative fitting method is discussed at the end of this chapter.

The data for Tables C-1 to C-4 were generated assuming dependence between y_1 and y_2. In Table C-3, the signs of the true correlations are preserved under all models except the model where y_1 and y_2 are assumed to be correlated, which again is unexpected and is apparently due to the limitations of our method for fitting the dependent ratios models. In the remaining cases, the signs of the true correlations are not preserved under any four of the models. When the signs of the empirical correlations are taken to be the reference point rather than the signs of the true correlations, then all four models preserve the signs of the empirical correlations in Tables C-1 and C-2. In Tables C-3 and C-4, the signs of the empirical correlations are preserved only under the model where y_1 and y_2 are dependent. Thus, the model that was actually used to generate the data preserves the empirical correlation sign structure for all four data sets, while the independent ratios model preserves the empirical correlation sign structure for only two of the data sets.

The data for Tables D-1 to D-4 were generated assuming dependence between y_2 and y_3. One of these (Table D-3) has the signs of the true correlations preserved under all four models. In the remaining cases (Tables D-1, D-2, and D-4), the signs of the true correlations are not preserved under any four of the models. Thus, none of the models are particularly successful at preserving the true correlation sign structures for these data sets. When the signs of the empirical correlations are taken to be the reference point rather than the signs of

the true correlations, Table D-3 has the signs of the empirical correlations preserved under all four models. Tables D-1 and D-2 have the signs of the empirical correlations preserved only under the model where y_2 and y_3 are dependent. In the remaining case (Table D-4), the signs of the empirical correlations are not preserved under any four of the models. Thus, the model that was actually used to generate the data preserves the empirical correlation sign structure for three of the four data sets, while the independent ratios model accomplishes this for only one of the data sets.

Finally, the data for Tables E-1 to E-4 were generated assuming independence of the ratios. Three of these (Tables E-1, E-2, and E-4) have the signs of the true correlations preserved under all four models. In Table E-3, the signs of the true correlations are preserved under all models except the model where y_1 and y_2 are dependent (note that the results for Table E-3 again reflect the limitations of our method for fitting the dependent ratios model). Thus, as expected, the independent ratios model does about as well as the dependent ratios models for these data sets. Similar results are obtained when the signs of the empirical correlations are taken to be the reference point in comparing the correlation sign structures, instead of the true correlations.

A few additional cases were run with the same correlation sign structures as the cases in Tables C-2, C-4, D-2, and E-2. The results were roughly similar to those presented here.

Thus, in our simulation, the dependent ratios model that was actually used to generate the simulated data preserved the true correlation sign structure in 12 of the 23 data sets in Appendices B-D, while the independent ratios model did this in only eight of the 23 data sets. As noted above, the performance of the dependent ratios model appears to be limited by our fitting method, and might be substantially better with a different fitting method. Note also

that the dependent ratios model performed better for the data sets given in Appendix B (where most of the true sign structures are not achievable with independent ratios) than for the data sets in Appendices C and D (where over half of the true sign structures are achievable with independent ratios). Finally, in all of the cases in which the dependent ratios models were not able to preserve the signs of the true correlations, the empirical correlation matrix had a different sign structure than the true correlation matrix.

When the empirical correlation matrix is taken to be the reference point rather than the true correlation matrix, the dependent ratios model that was actually used to generate the data preserved the empirical correlation sign structures for more than three fourths (19 out of 23) of the simulated data sets. By contrast, the independent ratios model preserved the empirical sign structures of the correlations in less than one third (seven out of 23) of the data sets. In practice, the true correlation sign structure will frequently not be known. Thus, if the aim is to match the empirical correlation sign structure of a given data set, our model appears to do substantially better than the independent ratios model.

The Euclidean distances between the true correlations and the correlations induced by the particular dependent ratios model that was actually used to generate the data are at least as small as the distances between the true correlations and the correlations of the independent ratios model in 11 of the 23 data sets created using dependent ratios (i.e., Tables B-1 through B-5, B-7, B-8, B-10, B-11, B-15, and C-1). Note that this is only about as good as would be expected due to chance alone. With regard to the Euclidean distances from the empirical correlations, the dependent ratios model does a better job, achieving smaller Euclidean distances than the independent ratios model in all of the 23 data sets generated by the dependent ratios model.

In fact, choosing the model that yields the smallest Euclidean distance between the model correlations and the true correlations in our data would have correctly identified the true model in 21 of the 23 dependent cases. While choosing the model in this way would not have correctly identified the four independent cases, the difference in Euclidean distance between the true (independent) model and the model with the smallest Euclidean distance is relatively small (less than a factor of 2 in all four cases). Therefore, one potentially promising strategy for model selection would be to choose the independent model unless one of the dependent models yields at least some pre-specified (e.g., 33%) reduction in Euclidean distance. Further investigation of this approach might be worthwhile.

In practice, we will frequently be interested in fitting data sets where the true correlations of the x_i are unknown and only the empirical correlations are available, as in the snowmelt runoff problem. Thus, 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. This is illustrated by the generally smaller Euclidean distances in Table 9-6 (distances from the empirical correlation matrix) than in Table 9-5 (distances from the true correlation matrix).

Remember, however, that these results should be regarded as only preliminary. A rigorous study of small-sample behavior would need to include more replication of each analysis (with different simulation seed values), as well as data sets of different lengths. In particular, we would expect our model to do substantially better for very long data sets, in which the empirical correlation matrix would converge to the true correlation matrix.

Table 9-5

Euclidean distances between the true correlations and

the correlations of the fitted models.

True correlated pair	Table #	Fitted model			
		Indep. ratios	y_1 dep. y_2	y_1 dep. y_3	y_2 dep. y_3
y_1 , y_3	B-1	0.20	0.19	**0.19**	0.20
	B-2	0.16	0.14	**0.12**	0.17
	B-3	0.10	0.12	**0.09**	0.11
	B-4	0.21	0.22	**0.16**	0.24
	B-5	0.11	0.13	**0.06**	0.12
	B-6	0.39	0.43	0.40	**0.36**
	B-7	0.16	0.16	**0.14**	0.17
	B-8	0.09	0.10	**0.05**	0.11
	B-9	**0.09**	0.10	0.09	0.12
	B-10	0.26	0.27	**0.25**	0.25
	B-11	0.13	0.15	**0.08**	0.13
	B-12	**0.01**	0.35	0.04	0.01
	B-13	0.02	**0.01**	0.05	0.17
	B-14	**0.12**	0.15	0.17	0.13
	B-15	0.22	0.25	**0.17**	0.23
y_1 , y_2	C-1	0.84	**0.81**	0.84	0.84
	C-2	0.06	0.09	**0.06**	0.07
	C-3	0.10	0.16	**0.09**	0.10
	C-4	0.48	0.95	0.48	**0.48**
y_2 , y_3	D-1	0.49	0.53	**0.49**	0.84
	D-2	0.33	0.36	**0.30**	0.39
	D-3	**0.14**	0.19	0.16	0.16
	D-4	0.58	**0.56**	0.59	0.60
None	E-1	**0.07**	0.09	0.07	0.07
	E-2	**0.07**	0.08	0.07	0.09
	E-3	**0.01**	0.05	0.02	0.01
	E-4	0.15	0.15	0.15	**0.13**

Note: Bold face used to indicate the smallest Euclidean distance for each data set.

Table 9-6

Euclidean distances between the empirical correlations and the correlations of the fitted models.

True correlated pair	Table #	Fitted model			
		Indep. ratios	y_1 dep. y_2	y_1 dep. y_3	y_2 dep. y_3
y_1 , y_3	B-1	0.14	0.15	**0.04**	0.14
	B-2	0.20	0.18	**0.06**	0.20
	B-3	0.09	0.08	**0.04**	0.08
	B-4	0.12	0.13	**0.06**	0.11
	B-5	0.12	0.12	**0.04**	0.12
	B-6	0.47	0.44	**0.13**	0.49
	B-7	0.14	0.16	**0.03**	0.14
	B-8	0.07	0.07	**0.05**	0.05
	B-9	0.08	0.07	**0.06**	0.07
	B-10	0.14	0.16	**0.03**	0.14
	B-11	0.12	0.12	**0.04**	0.12
	B-12	0.17	0.25	**0.17**	0.17
	B-13	0.17	0.17	0.17	**0.03**
	B-14	0.20	0.22	**0.06**	0.18
	B-15	0.12	0.11	**0.06**	0.13
y_1 , y_2	C-1	0.22	**0.04**	0.22	0.21
	C-2	0.11	**0.08**	0.11	0.10
	C-3	0.10	**0.04**	0.09	0.10
	C-4	0.62	**0.01**	0.62	0.62
y_2 , y_3	D-1	0.50	0.49	0.50	**0.05**
	D-2	0.58	0.58	0.58	**0.06**
	D-3	0.14	0.14	**0.07**	0.14
	D-4	0.19	0.19	0.15	**0.14**
None	E-1	0.04	**0.03**	0.04	0.04
	E-2	0.09	0.09	0.08	**0.05**
	E-3	0.03	0.06	**0.03**	0.04
	E-4	0.06	**0.04**	0.05	0.07

Note: Bold face used to indicate the smallest Euclidean distance for each data set.

9.3 Alternative fitting methods

As mentioned above, the performance of the dependent ratios model in this simulation study appears to be limited by the method used to fit the model to the simulated data sets. In particular, the fitting method we used, which was also used by Reese and Krzysztofowicz (1989), matches the moments of the y_i, which may not be optimal if the ultimate goal is to match the correlations of the x_i. This is clearly not the only possible fitting technique. For example, one could attempt to match the moments of the x_i (while also ensuring that the means and variances of the y_i stay within the bounds imposed by the fact that $0 \le y_i \le 1$). However, this approach turns out to be complicated computationally, because of extensive nonlinearities in the system of equations to be solved, and is also over-constrained, with 7 free parameters (the means, variances, and non-zero covariance of the y_i), but many more constraints (three for the means of the x_i, four for the variances of the x_i, and a number of inequality constraints for the means, variances, and non-zero covariance of the y_i, respectively).

Another interesting approach is to explicitly attempt to minimize the Euclidean distance between the empirical and fitted correlation matrices. That approach would in some sense put all of the models on the best possible footing for comparison purposes. The resulting optimization problem has a nonlinear objective function and a mix of linear and nonlinear constraints.

In this optimization problem, for the model where y_1 and y_3 are allowed to be dependent, we have 7 free parameters (the means and the variances of the y_i, and

$\text{cov}(y_1, y_3)$), and 10 constraints (three each to ensure that the means and variances of the y_i are within acceptable bounds, and four constraints that must be satisfied by the covariance of y_1 and y_3). This optimization problem is as follows:

$$\begin{aligned}
&\text{Minimize } \sum_{i=1}^{3}\sum_{j=i+1}^{4}[\text{cor}(x_i, x_j) - \hat{\text{cor}}(x_i, x_j)]^2 \\
&\quad \text{Subject to} \\
&\qquad 0 \le \text{E}(y_i) \le 1 \text{ for } i = 1,...,3. \\
&\qquad 0 \le \text{var}(y_i) \le [1 - \text{E}(y_i)]\text{E}(y_i) \text{ for } i = 1,...,3 \\
&\qquad -\sigma(y_1)\sigma(y_3) \le \text{cov}(y_1, y_3) \le \sigma(y_1)\sigma(y_3) \\
&\qquad \text{cov}(y_1, y_3) \le \text{E}(y_1)[1 - \text{E}(y_3)] \\
&\qquad \text{cov}(y_1, y_3) \le \text{E}(y_3)[1 - \text{E}(y_1)],
\end{aligned} \tag{9.2}$$

where $\text{cor}(x_i, x_j)$ is the correlation between x_i and x_j in the fitted model, $\hat{\text{cor}}(x_i, x_j)$ is the empirical correlation obtained from the given data set, and $\sigma(y_i) = \sqrt{\text{var}(y_i)}$.

Let $\text{cor}(x_i, x_j) = w_{ij}$ and $\hat{\text{cor}}(x_i, x_j) = \hat{w}_{ij}$, for $i = 1,2,3$ and $j = i + 1,...,4$. Then the objective function of the optimization problem in (9.2) is

$$\begin{aligned}
&\text{Minimize } [w_{12} - \hat{w}_{12}]^2 + [w_{13} - \hat{w}_{13}]^2 + [w_{14} - \hat{w}_{14}]^2 + \\
&\qquad [w_{23} - \hat{w}_{23}]^2 + [w_{24} - \hat{w}_{24}]^2 + [w_{34} - \hat{w}_{34}]^2
\end{aligned}$$

Writing the w_{ij} in terms of the y_i by means of equations 9.1 gives for example

$$w_{12} = \frac{[\text{E}(y_1) - \text{var}(y_1) - \text{E}(y_1)^2][\text{var}(y_2) + \text{E}(y_2)^2] - (1 - \text{E}(y_1))\text{E}(y_1)\text{E}^2(y_2)}{\sqrt{[\text{E}(y_1^{\,2})\text{E}(y_2^{\,2}) - \text{E}^2(y_1)\text{E}^2(y_2)][\text{E}((1 - y_1)^2)\text{E}(y_2^{\,2}) - \text{E}^2(1 - y_1)\text{E}^2(y_2)]}}$$

However, solution of this optimization problem was considered beyond the scope of this dissertation. Further exploration of alternative fitting methods is left for future work.