Journal of Statistics: Advances in Theory and Applications
Volume 3, Number 1, 2010, Pages 17-26

CORRELATION ANALYSIS FOR PROPORTIONS VIA UPDATING AND RESAMPLING

G. S. MONTI and G. WALTHER

Department of Statistics
University of Milan-Bicocca
Italy
e-mail: gianna.monti@unimib.it

Department of Statistics
Stanford University
Sequoia Hall, 390 Serra Mall
Stanford, CA 94305
U. S. A.

Abstract

Compositional data arise frequently in practice, but their statistical analysis is not yet well developed. Compositions arise when non-negative random vectors are mapped into the unit simplex via a closure operation, e.g., when the amounts of certain minerals in a soil sample are converted to percentages. Components in random compositions can never be stochastically independent, due to the spurious correlation introduced by taking the closure of the basis vectors. This paper aims to assess independence of the unobserved basis vectors based on the observed composition. We propose a resampling procedure that is based on an updating formula. A simulation study shows that this procedure works well in the case, where the components of the composition are roughly of the same size. We apply the procedure to a geochemical data set obtained from the Kola peninsula.

2010 Mathematics Subject Classification: 97F80, 62H20, 62G09, 62F40.

Keywords and phrases: closure data, independence for proportions, bootstrap test.

Received February 18, 2010; Revised March 30, 2010

1. Introduction

A random composition $\underline{X} = (X_1, \ldots, X_D)$ is a random vector that lies in the unit simplex:

$$S^D = \left\{ \underline{X} : X_i \geq 0, i = 1, \ldots, D; \sum_i X_i = 1 \right\}.$$

Any random vector $\underline{W} \in \mathbf{R}_+^D$ with non-negative components can be transformed into a random composition $\underline{X}$ by taking its closure: $X_i = W_i / \sum_{k=1}^D W_k$, $i = 1, \ldots, D$. A typical case, where compositions arise is when quantities, e.g., the amounts of certain minerals in a soil sample, are converted to percentages. The statistical analysis of compositional data $\underline{X}$ is challenging for several reasons: not only does the closure operation lead to a loss of information, but dividing by $\sum W_k$ introduces a spurious correlation into the components of $\underline{X}$. See Aitchison [1] for a detailed exposition on compositions.

In this paper, we are concerned with testing whether the basis components W_k are uncorrelated, given only realizations of the composition $\underline{X}$. This problem is of interest in a number of applications [1]. The above mentioned spurious correlation makes this problem deceptive and difficult, a fact that was brought to the fore by the seminal paper of Pearson [14]. Pearson realized that due to the spurious correlation, a value of zero for Pearson's correlation coefficient is not an appropriate measure of absence of correlation. Pearson attempted to correct for this with an approximation based on component means and variances. Pearson's formula was refined by Chayes [2] and by Chayes and Kruskal [3] by expressing the correlation of two components in terms of the moments of the basis variables. By expanding the identity $\mathrm{Var}(\sum_{i=1}^D X_i) = 0$, Chayes showed that equal variances of the $X_1, \ldots, X_D$ implies that, the average value of $\mathrm{cor}(X_i, X_r)$ over all pairs (i, r), $i \neq r$, is $-1/(D-1)$, so there is a negative bias in the correlations, and he attempted to separate the spurious part from the real correlation. Vistelius [16] has described the basic problem of Chayes's approach and developed some essential rules. In the same way, Darroch [5] studied the relationship among the structure of correlation of the basis and that of the composition. Connor and Mosimann [4] have treated the problem in a probabilistic, theoretical form. Some experiments demonstrate a lower limit for the applicability of Chayes's model [11] furthermore, his test for correlation is inappropriate unless the number of variables D is greater than or equal to four.

In this paper, we attempt to derive appropriate null distributions for the correlation coefficients between components of the composition by using the bootstrap and a conditioning argument. The appeal and the power of the bootstrap lie in its ability to extract information from the data that is difficult to obtain analytically. This suggests that, an approach based on the bootstrap may be more successful than the analytical methods described above.

2. Some Distributions on the Simplex

One of the most popular models for random compositions is the Dirichlet distribution. This model is obtained by closing a basis of independent equally-scaled Gamma random variables (r.v.' s) $W_i \sim Ga(\alpha_i, 1)$. We denote the Dirichlet distribution with parameter vector $\underline{\alpha} = (\alpha_1, \ldots, \alpha_D) \in \mathbf{R}_+^D$ by $\mathcal{D}^D(\underline{\alpha})$.

A useful generalization of the Dirichlet model is the flexible Dirichlet distribution [13], which is a finite mixture of Dirichlet distributions. It is generated by closing a basis $\underline{Y}$ of positive, but dependent r.v.'s as follows: $Y_i = W_i + Z_i U$, where $U \sim Ga(\tau, 1)$ and $\underline{Z} = (Z_1, \ldots, Z_D) \sim M_D(1; p_1, \ldots, p_D)$ has a multinomial distribution, and the W_i's, U, and $\underline{Z}$ are independent r.v.'s and $\underline{\alpha} = (\alpha_1, \ldots, \alpha_D) \in \mathbf{R}_+^D$, $\underline{p} = (p_1, \ldots, p_D) \in S^D$, and $\tau > 0$. We denote this flexible Dirichlet distribution by $\mathcal{FD}^D(\underline{\alpha}, \underline{p}, \tau)$. Any component of $\underline{p}$ is called *mixing proportion*.

3. Assessing Correlation for the Basis

Our goal is to approximate the null distribution of the correlation coefficient between two components of the composition, i.e., the distribution when the corresponding basis vectors are independent. The bootstrap principle allows to simulate this null distribution, provided we have an estimate of the distribution of the basis. Since, we need to resample under the null hypothesis of independence, it is enough to have estimates of the marginal basis distributions. But obtaining suitable estimates is not straightforward: we observe only realizations of the composition, not of the basis, and there is no unique way to reconstruct the basis distribution from that of the composition. For example, each scale change of the basis will lead to the same composition.

We construct estimates with a two-step procedure as follows. First, we construct a rough estimate of the basis distributions by assuming that the composition $\underline{X}$ follows a Dirichlet distribution, for which we can estimate the parameters with maximum likelihood. Then, we refine this estimate by computing its expected value given the observed data $\underline{X}$. In more detail:

1. Assuming that $\underline{X} \sim \mathcal{D}^D(\alpha_1, \dots, \alpha_D)$, we estimate the parameters α_i from the sample with the method of maximum likelihood, using the algorithm of Ronning [15]:

First, we use the method of moments to obtain starting values for the algorithm:

$$\hat{\alpha}_i^{(0)} = \frac{(\bar{x}_{11} - \bar{x}_{21})\bar{x}_{1i}}{\bar{x}_{21} - (\bar{x}_{11})^2}, \quad \hat{\alpha}_D^{(0)} = \frac{(\bar{x}_{11} - \bar{x}_{21})\left(1 - \sum_{i=1}^{D-1} \bar{x}_{1i}\right)}{\bar{x}_{21} - (\bar{x}_{11})^2},$$

where $\bar{x}_{1i} = \frac{1}{n}\sum_{j=1}^{n} x_{ji}$ and $\bar{x}_{2i} = \frac{1}{n}\sum_{j=1}^{n} x_{ji}^2$, for $i = 1, \dots, D-1$.

Then, we use the Newton-Raphson algorithm to obtain the maximum likelihood estimates $\hat{\alpha}_i$:

$$\underline{\hat{\alpha}}^{(k)} = \underline{\hat{\alpha}}^{(k-1)} - \mathbf{H}^{-1}\underline{g},$$

where $\mathbf{H}$ is the Hessian matrix of the log-likelihood In $L(\underline{\alpha}|\underline{x}_j)$ and $g_i = \partial \ln L(\underline{\alpha}|\underline{x}_j)/\partial\alpha_i$ is the gradient. The stopping rule for the algorithm is:

$$\left|L(\underline{\alpha}^{(k+1)}) - L(\underline{\alpha}^{(k)})\right| \leq \varepsilon,$$

where $\varepsilon = 1e - 8$, the absolute convergence tolerance, i.e., a tolerance for reaching zero.

2. Next, we update our estimates by conditioning on the data $\underline{X}$:

$$\begin{aligned}\mathbb{P}(W_i \leq t) &= \mathbb{E}_{X_i}\mathbb{P}(W_i \leq t|X_i)\\ &= \mathbb{E}_{X_i}\mathbb{P}\left(\sum_{j=1}^{D} W_j \leq \frac{t}{x_i}|X_i\right)\\ &\approx \int \mathbb{P}\left(\sum_{j=1}^{D} W_j \leq \frac{t}{x}\right) f_{X_i}(x)\,dx\\ &\approx \sum_{k=1}^{n}\mathbb{P}\left(\sum_{j=1}^{D} W_j \leq \frac{t}{x_{ki}}\right)\frac{1}{n}.\end{aligned}$$

The first approximation above uses approximate independence of X_i and $\sum_{j=1}^{D} W_j$, as is exactly true in the gamma (Dirichlet) case. The second approximation uses the law of large numbers. For evaluating the last expression, we use our estimate obtained via maximum likelihood above, so $\sum_{j=1}^{D} W_j \sim Ga(\sum_{j=1}^{D}\hat{\alpha}_j)$. Hence, $\mathbb{P}(\sum_{j=1}^{D} W_j \leq \frac{t}{x_{ki}})$ can be evaluated quickly with standard statistical software. We evaluate $F_i(t) = \mathbb{P}(W_i \leq t)$ on a grid of 1,000 points.

Now, we can use the nonparametric bootstrap to sample W_i^* from the updated marginal estimate F_i, $i = 1, \ldots, D$. This can be done in a well known way via the probability integral transformation: we generate

$\mathcal{U} \sim U[0, 1]$ and set $W_i^* = \min\{t : F_i(t) \geq \mathcal{U}\}$, where the minimum is taken over the grid. We generate $B = 1{,}000$ bootstrap samples $\underline{W}^*$ with independent components $W_i^* \sim F_i$, $i = 1, \ldots, D$.

Our test for correlation proceeds then as follows:

Given the data matrix $\mathbb{X}_{n \times D}$, we compute the sample correlation matrix:

$$R_{obs} = [r_{ik}], \quad i, k = 1, \ldots, D,$$

where $r_{ik} = \mathrm{cor}(\underline{x}_i, \underline{x}_k)$, and $\underline{x}_i$ and $\underline{x}_k$ are the column vectors corresponding to the observations on the i-th and k-th component. Then, we compute Fisher's transformation for each correlation coefficient r_{ik}:

$$z_{ik} = \frac{1}{2} \ln\left(\frac{1 + r_{ik}}{1 - r_{ik}}\right).$$

This transformation is well known to improve confidence intervals, see, e.g., Efron and Tibshirani [9]. Next, we compute the bootstrap null distribution and assess the significance of z_{ik} against this null distribution:

1. Generate $B = 1{,}000$ bootstrap samples $\underline{W}^*$ as described above and compute their closures $\underline{X}^*$.

2. Compute the sample correlation matrix for each bootstrap closure: $R_b = [r_{ik}^{*(b)}]$, where $i, k = 1, \ldots, D$ and $b = 1, \ldots, B$. Then use the Fisher's transformation to obtain $z_{ik}^{*(b)}$.

3. For each of the $D(D-1)/2$ tuples (i, k) compute $\left[z_{ik,0.025}^*, z_{ik,0.975}^*\right]$, where $z_{ik,0.025}^*$ and $z_{ik,0.975}^*$ are the 0.025-quantile and 0.975-quantile, respectively, of the empirical distribution of the $z_{ik}^{*(b)}$.

4. If $z_{ik} \notin [z_{ik,0.025}^*, z_{ik,0.975}^*]$, then reject the null hypothesis of independence between W_i and W_k.

4. Simulations, Application and Conclusion

We performed a simulation study to test our procedure. We considered four cases:

1. $\underline{X} \sim \mathcal{D}^3(\underline{\alpha} = (1, 2, 3))$.

2. $\underline{X}$ is generated by taking the closure of 3 independent $W_i \sim |N(0, 1)|$, $i = 1, \ldots, 3$.

3. $\underline{X} \sim \mathcal{FD}^3(\underline{\alpha} = (4, 3, 7), \underline{p} = (0.35, 0.45, 0.2), \tau = 3)$.

4. $\underline{X}$ is generated by taking the closure of the absolute values of the components of a multivariate normal distribution with mean vector (3, 4, 7), and the following covariance matrix:

$$\Sigma = \begin{bmatrix} 12 & -5.2 & -2 \\ -5.2 & 8 & 3 \\ -2 & 3 & 2 \end{bmatrix}.$$

Cases 1 and 2 pertain to the null hypothesis of independence, whereas the bases in Cases 3 and 4 are not independent. In particular, the Case 1 starts from a usual distribution for random compositions defined in the unit simplex. In Case 2, we start from an independent basis of non-negative random variables that are non-common in the compositional framework. In Case 3, we generate samples from the flexible Dirichlet distribution, which is a useful distribution for a random composition. In Case 4, we consider samples from a dependent basis of non-negative random variables that are non-common in the compositional framework.

We simulated the probabilities of rejection of our bootstrap test with a Monte Carlo simulation study by using 1,000 Monte Carlo samples of size $n = 500$ in each case. The result of the Monte Carlo study is given in Table 1.

Table 1. Proportion of rejections at the 5% level

(i, k)	Case 1	Case 2	Case 3	Case 4
(1, 2)	0.026	0.021	0.75	1
(1, 3)	0.012	0.011	0.24	0.34
(2, 3)	0.001	0.01	0.10	1

The simulation results show that our bootstrap-test works well for each model considered. We did not obtain satisfactory results in null cases, where the sizes of the basis variables were very different. Thus, our conclusion is that, the bootstrap test with updating works when the components are of approximately similar size.

We applied the bootstrap test to the Kola data set. These data contain the concentrations of more than 50 chemical elements in about 600 soil samples taken at the Kola Peninsula. At each sample site, four different layers of soil have been analyzed. The complete data set is available in the R library Stat DA [8].

Here, we use the analytical results from the O-horizon, in particular, we consider the compositions of Gold (Au), Palladium (Pd), and Zinc (Zn).

We plotted the histograms of the bootstrap null distributions for the three Fisher's transformations in Figure 1. The dotted vertical lines represent the 95% bootstrap confidence limits. The solid vertical line represents the observed values of z_{13}. The other two observed values are outside the range of the histogram. We conclude that, we cannot reject the hypothesis of no correlation between Gold (Au) and Zinc (Zn), but we can reject the null hypothesis for the other two correlations.

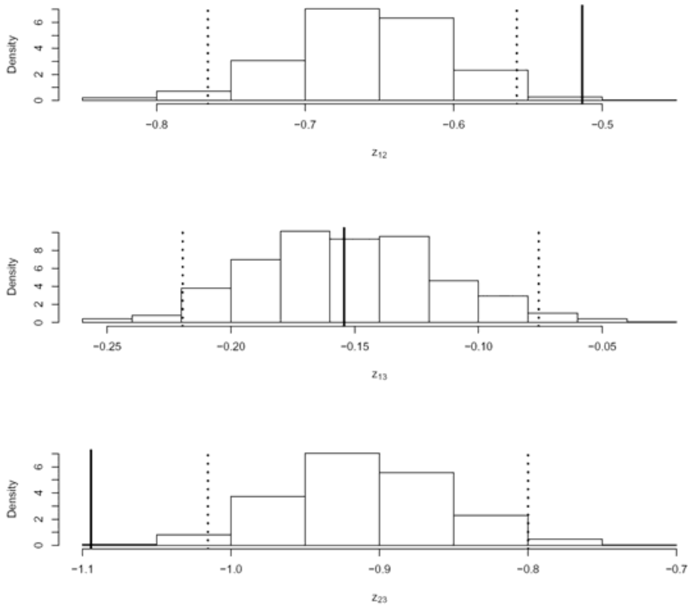

Figure 1. Empirical distributions of the test statistic.

References

[1] J. Aitchison, The Statistical Analysis of Compositional Data, Chapman and Hall Ltd., London, 1986.

[2] F. Chayes, On correlation between variables of constant sum, Journal of Geophysical Research 65 (1960), 4185-4193.

[3] F. Chayes and W. Kruskal, An approximate statistical test for correlations between proportions, Journal of Geology 74 (1966), 692-702.

[4] R. J. Connor and J. E. Mosimann, Concept of independence for proportions with a generalization of the Dirichlet distributions, J. Amer. Statist. Assoc. 64(325) (1969), 194-206.

[5] J. N. Darroch, Null correlation for proportions, Mathematical Geology 1(2) (1969), 221-227.

[6] J. N. Darroch and D. Ratcliff, Null correlation for proportions II, Journal of Mathematical Geology 2 (1970), 307-312.

[7] J. N. Darroch and D. Ratcliff, No-association of proportions, Mathematical Geology 10(4) (1978), 307-312.

[8] R Development Core Team, R: A Language and Environment for Statistical Computing, Vienna, 2008. http://www.r-project.org

[9] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.

[10] J. O. Kork, Examination of the Chayes–Kruskal procedure for testing correlations between proportions, Mathematical Geology 9(6) (1977), 543-562.

[11] G. S. Monti, Analysis of Correlation Among Components of a Flexible Dirichlet Distribution, Proceedings of S.Co.09, Complex Data Modelling and Computationally Intensive Statistical Methods for Estimation and Prediction, Politecnico di Milano, Milan (Italy), 2009.

[12] A. Narayanan, Algorithm AS266: maximum likelihood estimation of parameters of the Dirichlet distribution, Journal of the Royal Statistical Society, Series C (Applied Statistics) 40(2) (1991), 365-374.

[13] A. Ongaro, S. Migliorati and G. S. Monti, A New Distribution on the Simplex Containing the Dirichlet Family, Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop, University of Girona, Girona (Spain), CD-ROM (ISBN: 978-84-8458-272-4), 2008.

[14] K. Pearson, Mathematical contributions to the theory of evolution, On a form of spurious correlation which may arise when indices are used in the measurement of organs, Proceedings of the Royal Society of London LX (1897), 489-502.

[15] G. Ronning, Maximum likelihood estimation of Dirichlet distributions, J Stat. Comput. Simul. 32(5) (1989), 215-221.

[16] A. B. Vistelius, Discussion on paper by F. Chayes, on correlation between variables of constant sum, Journal of Geophysical Research 6(5) (1961), 1601.

[17] E. L. Zodrow, Empirical behavior of Chayes null model, Mathematical Geology 8(1) (1976), 37-42.

■