On Bicompositional Correlation

Bergman, Jakob

2010

Introduction

1 Compositions

An essential part of statistics is analysing measurements of various entities. Normally these values make perfect sense; we may be interested in the number of cars, the velocity of each car, or the weight of each car. There are however situations when we are not interested in the absolute values of our measurements, but the relative ones; the absolute values may not even be available to us. The absolute amount of a certain oxide in a rock sample or the absolute number of respondents who would vote for a certain party in a party preference survey are seldom of interest, whereas the relative amount of a certain oxide and the relative number of respondents are usually more interesting. We often refer to these relative values as proportions. The proportions of all the different outcomes must of course sum to 1 (or 100 %). A vector of these proportions is known as a *composition*, or put more mathematically: a composition is a vector of positive components summing to a constant, usually taken to be 1. As indicated above, compositions arise in many different areas; the geochemical compositions of different rock specimens, the proportion of expenditures on different commodity groups in household budgets, and the party preferences in a party preference survey are all examples of compositions from three different scientific areas.

The sample space of a composition is the simplex. Without loss of gener-

ality we will always take the summation constant to be 1, and we define the D-dimensional simplex $\mathscr{S}^D$ as

$$\mathscr{S}^D = \left\{(x_1, \ldots, x_D)^\top \in \mathscr{R}_+^D : \sum_{j=1}^{D} x_j = 1\right\},$$

where $\mathscr{R}_+$ is the positive real space.

In this thesis we will refer to compositions with two components (or parts), i.e. $D = 2$, as *bicomponent*, with three components, i.e. $D = 3$, as *tricomponent*, and with more than two components, i.e. $D > 2$, as *multicomponent*. Please note the difference between *bicompositional* referring to two compositions and *bicomponent* referring to a composition with two components. The two notions will be used together as in "a bicomponent bicompositional distribution," i.e. a joint distribution of two compositions each with two components.

2 A short historical review

Compositions have been studied almost as long as the subject of modern statistics has existed. Pearson (1897) was the first to realize that if you divide two independent random variates with a third random variate, independent of the first two, the two quotients will be correlated. Pearson called this "spurious correlation" and warned researchers for this phenomenon. This "spurious correlation" of course applies to compositions, since compositions are usually made up of a number of measurements divided by their sum; in fact for compositions the denominator is not even independent of the measurements. Since then it should have been known that compositions have to be treated with care. During the following 60 years this was however usually not the case.

In 1986 Aitchison published his pivotal book *The Statistical Analysis of Compositional Data* (reprinted 2003). In this book he argues for the concept of logratio transformations as a way to resolve the problems caused by

the compositional summation constraint. Aitchison presented two logratio transformations: the additive logratio transformation (ALR) and the centred logratio transformation (CLR). Later Egozcue et al. (2003) introduced the isometric logratio transformation (ILR). The ALR transformation consists of the logarithms of the components, omitting one, divided by the omitted reference component; the CLR transformation consists of the logarithms of the components divided by the geometric mean of the components. The ILR transformation is a much more complex transformation. If for example $\mathbf{x} = (x_1, x_2, x_3, x_4)^{\mathrm{T}} \in \mathscr{S}^4$, then the resulting vectors of the different transformations are the following:

$$\begin{aligned}
\mathrm{alr}(\mathbf{x}) &= \left(\log \frac{x_1}{x_4},\ \log \frac{x_2}{x_4},\ \log \frac{x_3}{x_4} \right)^{\mathrm{T}} \\
\mathrm{clr}(\mathbf{x}) &= \left(\log \frac{x_1}{g(\mathbf{x})},\ \log \frac{x_2}{g(\mathbf{x})},\ \log \frac{x_3}{g(\mathbf{x})},\ \log \frac{x_4}{g(\mathbf{x})} \right)^{\mathrm{T}} \\
\mathrm{ilr}(\mathbf{x}) &= \left(\frac{1}{\sqrt{2}} \log \frac{x_1}{x_2},\ \frac{1}{\sqrt{6}} \log \frac{x_1 x_2}{x_3^2},\ \frac{1}{\sqrt{12}} \log \frac{x_1 x_2 x_3}{x_4^3} \right)^{\mathrm{T}}
\end{aligned}$$

where $g(\mathbf{x}) = (x_1 \cdots x_D)^{1/D}$, i.e. the geometric mean. The three transformations are related, see for instance Barceló-Vidal et al. (2007).

Aitchison and Egozcue (2005) distinguish four phases in the evolution of compositional analysis, the first one being the phase until 1960s when the complications with compositional data were ignored, and the second being the phase from the 1960s until the 1980s when different ideas were tried to resolve the problems of the multivariate methods not working for compositional data. The third phase is that when the logratio methodology gains acceptance. The fourth phase started some ten years ago, with the realization that the simplex is a Hilbert space (see e.g. Pawlowsky-Glahn and Egozcue, 2001, 2002). This has given rise to a "stay-in-the-simplex" approach. This approach basically provides a way of modelling the operations done on the logratio transformed data, then usually referred to as *coordinates*, in the simplex.

3 Compositional time series

The interest for bicompositional correlation resulting in this thesis originally began as an interest in compositional time series (CTS), i.e. time series of compositions. Compositional time series arise in many different situations, for instance party preference surveys, labour force surveys or pollution measurements.

Even though there have only been relatively few papers published on CTS, there have been several approaches to CTS; these have been reviewed by Larrosa (2005) and Aguilar Zuil et al. (2007).

The first to discuss and use an ALR approach to CTS seem to be Aitchison (1986) and Brunsdon (1987), which were followed by Smith and Brunsdon (1989) and Brunsdon and Smith (1998). In that approach the CTS is transformed with an ALR, and the transformed series is then analysed with standard models, e.g. VAR or VARMA. Bergman (2008) and Aguilar and Barceló-Vidal (2008) have also used ILR to model the data. The choice of logratio transformation is of course arbitrary.

There have also been some ideas on how to model the time series on the simplex. Apart from Aitchison and Brunsdon, Billheimer and Guttorp (1995) and Billheimer et al. (1997) have used autoregressive and conditional autoregressive models. Barceló-Vidal et al. (2007) introduced a compositional ARIMA model, defined using the "stay-in-the-simplex" approach.

As an illustration of CTS we present a figure from Bergman (2008), where a time series from the Swedish labour force survey (AKU) was modelled. Figure 1 gives three views of the analysed time series; the top plot shows the time series in a ternary time series plot (sometimes referred to as a "Toblerone plot"), the middle plot shows the three components of the time series in a standard time series plot, and the bottom plot shows a standard time series plot of the ILR-transformed time series. In all three plots the structural change in the series due to the Swedish fiscal crisis during the early 1990s is clearly visible, as well as a seasonal pattern.

Figure 1 (Next page) Three views of a compositional time series. The top plot shows the time series in a ternary time series plot, where the top corner of the Simplex represents 100 % Unemployment, the bottom left corner 100 % Employment, and the bottom right corner that 100 % of the population are Not belonging to the labour force. The middle plot shows the three components of the time series in a standard time series plot. (Note that the vertical axis has been cut and has different scales in the different parts.) The bottom plot shows the ILR-transformed series. (The second component of the transformed series is plotted with a dotted line.) In all three plots the structural change in the series during the early 1990s is clearly visible, as well as the seasonal pattern.

Source: Statistics Sweden

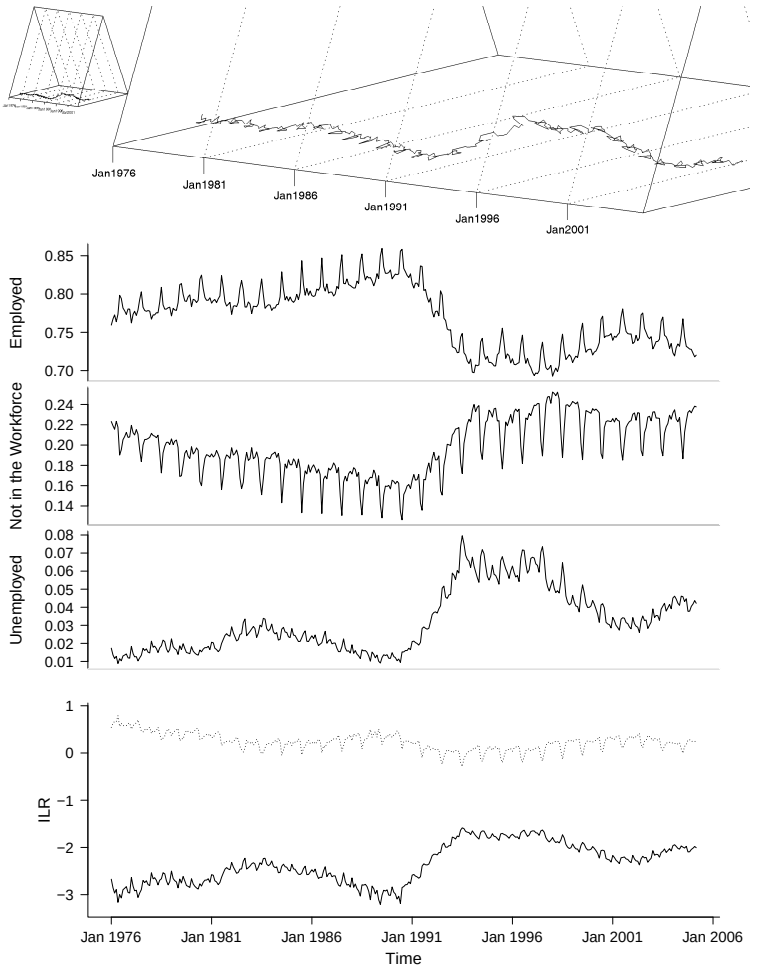
4 Correlation

Unlike the observations in cross-sectional data, the observations in time series are usually not independent. A not entirely unintuitive starting point for describing this dependence is to consider the concept of correlation. This thesis tries to target the question: "How do we model, measure and compare similarity or dissimilarity between two compositions?"

When hearing the word "correlation" most people would probability think of the product moment correlation coefficient

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}},$$

which measures the linear relationship between two variables. This is also how correlation is defined in *Encyclopedia of Statistical Sciences* (Rodriguez, 1982). However, correlation does not have be restricted to linear relationships or univariate variables. Dodge (2003) for instance states that it can be "used broadly to mean some kind of statistical relation between variables." This wider approach includes correlation coefficients that need not measure linear relationships, for instance the rank correlation coefficient Spearman's ρ_S. It is this wider approach we will utilize. We thus consider correlation as a measure


Jan1976
Jan1981
Jan1986
Jan1991
Jan1996
Jan2001
Employed
0.85
0.80
0.75
0.70
Not in the Workforce
0.24
0.22
0.20
0.18
0.16
0.14
Unemployed
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
ILR
1
0
−1
−2
−3
Jan 1976
Jan 1981
Jan 1986
Jan 1991
Jan 1996
Jan 2001
Jan 2006
Time

of similarity.

A good measure of correlation (or similarity) should also be able to compare not just two observations of the same composition at different time points, but also of two different compositions at the same time point. These two compositions might not even have equal numbers of components. We could for instance consider the correlation between some composition of the labour force and some composition of the gross domestic product. In this thesis we will however restrict our analysis to the correlation between two observations of the same composition, but with the introduction of suitable distributions, the result of this thesis is easily generalized to the above situations.

5 Bicompositions

In order to parametrically quantify the correlation between two compositions one needs to consider the joint distribution of the compositions. As stated above, the sample space of a D-component composition is the simplex $\mathscr{S}^D$. The sample space of two compositions $\mathbf{X}$, $\mathbf{Y}$, defined on $\mathscr{S}^D$, is consequently the Cartesian product $\mathscr{S}^D \times \mathscr{S}^D$. This is however not a simplex, but a manifold with two constraints, a *bisimplex*. We note that whereas the Cartesian product of two random vectors on the real space $\mathscr{R}^p$ will form a new random vector on the real space $\mathscr{R}^{p+p}$, this does not hold for two simplices: $\mathscr{S}^D \times \mathscr{S}^D \neq \mathscr{S}^{D+D}$.

The Cartesian product of two D-component compositions could have been denoted

$$\mathbf{Z} = (Z_1, \ldots, Z_D, Z_{D+1}, \ldots, Z_{D+D})^{\mathrm{T}},$$

where $\sum_{j=1}^{D} Z_j = \sum_{j=D+1}^{D+D} Z_j = 1$. However, throughout this thesis we choose to denote it

$$(\mathbf{X}, \mathbf{Y}) = (X_1, \ldots, X_D, Y_1, \ldots, Y_D)^{\mathrm{T}},$$

to stress the fact that we regard it primarily as two compositions and not as one *bicomposition*.

We will in this thesis base our modelling of correlation on an extension of the Dirichlet distribution. Following Aitchison (1986), we define the Dirichlet probability density function with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D) \in \mathscr{R}_+^D$ as

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_D)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_D)} x_1^{\alpha_1 - 1} \cdots x_D^{\alpha_D - 1},$$

where $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}} \in \mathscr{S}^D$ and $\Gamma(\cdot)$ is the Gamma function. We will present a bicompositional generalization of the Dirichlet distribution, defined on the Cartesian product of two simplices, i.e. a bisimplex. The notation $(\mathbf{X}, \mathbf{Y})$ will also allow us to emphasize the relationship between the new distribution and the product of two Dirichlet distributions.

In accordance with the Dirichlet integral, the new distribution is defined with respect to the Lebesgue measure. It remains as future work to reformulate it using the Aitchison (or simplicial) measure (Pawlowsky-Glahn, 2003) along the lines of Mateu-Figueras and Pawlowsky-Glahn (2005).

6 Outline of the thesis

This thesis is based on four papers concerning bicompositions and modelling the correlation between compositions. The contents of the papers are presented briefly below.

6.1 Paper I

We search the literature for distributions defined on the Cartesian product $\mathscr{S}^D \times \mathscr{S}^D$, and find a few bivariate Beta distributions for the bicomponent case, but no distributions defined on $\mathscr{S}^D \times \mathscr{S}^D$ when $D > 2$.

We introduce a bicompositional Dirichlet distribution. The distribution is defined on the Cartesian product $\mathscr{S}^D \times \mathscr{S}^D$ and is based on the product

of two Dirichlet distributions. The probability density function is

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = A\left(\prod_{j=1}^{D} x_j^{\alpha_j-1} y_j^{\beta_j-1}\right)\left(\mathbf{x}^{\mathrm{T}}\mathbf{y}\right)^{\gamma},$$

where $\mathbf{x} = (x_1,\ldots,x_D)^{\mathrm{T}} \in \mathscr{S}^D$, $\mathbf{y} = (y_1,\ldots,y_D)^{\mathrm{T}} \in \mathscr{S}^D$, and $\alpha_j, \beta_j \in \mathscr{R}_+$ $(j = 1,\ldots,D)$. The parameter space of γ depends on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$; however, all non-negative values are always included. The parameter γ models the degree of covariation between $\mathbf{X}$ and $\mathbf{Y}$. When $\gamma = 0$, the distribution is the product of two independent Dirichlet distributions.

We prove that the distribution exists in the bicomponent case if and only if $\gamma > -\min(\alpha_1+\beta_2, \alpha_2+\beta_1)$ and at least for $\gamma \geq 0$ in the multicomponent case. We also give expressions for the normalization constant A for all γ in the bicomponent case and for integers γ in multicomponent case.

In the bicomponent case we present expressions for the cumulative distribution function and the product moment. In both the bicomponent and the multicomponent case, we derive expressions for the marginal probability density functions and the marginal moments, and for the conditional probability density distribution and conditional moments.

6.2 Paper II

We consider two families of parametric models $\{f(\mathbf{x},\mathbf{y};\theta), \theta \in \Theta_i\}$ $(i = 0, 1)$ with $\Theta_0 \subset \Theta_1$ when modelling $(\mathbf{X},\mathbf{Y})$ and assume that the true joint density function is $g(\mathbf{x},\mathbf{y})$. Kent (1983) defines the Fraser information as

$$F(\theta) = \int \log f(\mathbf{x},\mathbf{y};\theta) g(\mathbf{x},\mathbf{y}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}$$

and the information gain as

$$I(\theta_1 : \theta_0) = 2\{F(\theta_1) - F(\theta_0)\},$$

where θ_i is the parameter value that maximizes $F(\theta)$ under the parameter space Θ_i $(i = 0, 1)$. Using $I(\theta_1 : \theta_0)$, Kent (1983) proposes a general measure of correlation, or joint correlation coefficient, between $(\mathbf{X}, \mathbf{Y})$ defined as

$$\rho_J^2 = 1 - \exp\{-I(\theta_1 : \theta_0)\},$$

where $\mathbf{X}$ and $\mathbf{Y}$ are modelled as independent quantities under Θ_0.

We use the bicompositional Dirichlet distribution presented in Paper I to model two compositions $\mathbf{X}$ and $\mathbf{Y}$. We let $\theta = (\alpha, \beta, \gamma)$ and $\Theta_0 = \{\theta : \gamma = 0\}$, while Θ_1 is the unrestricted parameter space.

The joint correlation coefficient is calculated, utilizing that the bicompositional Dirichlet distribution constitutes an exponential family of distributions, and it is presented graphically for a large number of bicomponent bicompositional models. We note that ρ_J^2 as a function of γ is not symmetric around 0.

We also calculate the joint correlation coefficient for nine tricomponent bicompositional models.

In the Appendices we present and examine expressions for the first derivative of the binomial coefficient

$$\frac{\mathrm{d}}{\mathrm{d}r}\binom{r}{n},$$

and we also give a suggestion for numerical integration over $\mathscr{S}^3 \times \mathscr{S}^3$.

6.3 Paper III

We use the rejection method to generate random variates with a bicompositional Dirichlet density f. Given a dominating density g and a constant $c \geq 1$ such that $f(\mathbf{x}, \mathbf{y}) \leq cg(\mathbf{x}, \mathbf{y})$, and a random number U uniformly distributed on the unit interval, a generated variate $(\mathbf{x}, \mathbf{y})$ is accepted if

$$U \leq \frac{f(\mathbf{x}, \mathbf{y})}{cg(\mathbf{x}, \mathbf{y})},$$

otherwise it is rejected and new $(\mathbf{x}, \mathbf{y})$ and U are generated until acceptance. We hence need to find dominating densities g and constants c. We examine three cases.

First we look at the (trivial) case when $\gamma = 0$, i.e. the product of two independent Dirichlet distributions. Dirichlet distributed variates are easily generated using Gamma distributed variates, and thus we need not use the rejection method.

Secondly we examine the case when $\gamma > 0$. We use a bicompositional Dirichlet distribution with $\gamma = 0$, i.e. the product of two independent Dirichlet distribution, as dominating density. We find that the random variate is accepted if $U \leq (\mathbf{x}^{\mathrm{T}}\mathbf{y})^{\gamma}$. Evidently, we need not calculate the normalization constant $A(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$, and hence we can generate random numbers from bicompositional Dirichlet distributions whose probability density functions we cannot calculate. When γ is very large, the method will be slow, as the acceptance probability $\Pr\{U \leq (\mathbf{x}^{\mathrm{T}}\mathbf{y})^{\gamma}\} = (\mathbf{x}^{\mathrm{T}}\mathbf{y})^{\gamma}$ will be very low. We note that we can always use a uniform density as g, with $c = \max_{\mathbf{x},\mathbf{y}} f(\mathbf{x}, \mathbf{y})$. This is though only applicable for non-negative integers γ, since it is necessary to calculate $A(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$.

Thirdly we examine the bicomponent case when $\gamma < 0$. We partition the sample space into four quadrants Q1-Q4, and choose a quadrant Qk ($k = 1, 2, 3, 4$) randomly with probability

$$\iint_{\mathrm{Q}k} f(x, y)\mathrm{d}x\mathrm{d}y \qquad (k = 1, 2, 3, 4),$$

where $f(x, y)$ is the bicomponent bicompositional Dirichlet probability density function viewed as a function of x and y. For each of the quadrants we find a dominating density based on the product of two Dirichlet distributions and a constant c, and generate a random variate using the rejection method. A slight problem with the method is to find effective ways of generating random Dirichlet distributed variates that are restricted to a particular quadrant.

We compare the efficiencies of the two suggestions for dominating densities, Dirichlet and uniform, with a Monte Carlo study.

6.4 Paper IV

We present maximum likelihood estimates of the parameter $\theta = (\alpha, \beta, \gamma)$ of the bicompositional Dirichlet distribution presented in Paper I. Following Kent (1983) we also present an estimator of the general measure of correlation, or joint correlation coefficient, presented in Paper II, assuming that the data follow a bicompositional Dirichlet distribution,

$$\hat{\rho}_J^2 = 1 - \exp\{-\widehat{I}(\hat{\theta}_1 : \hat{\theta}_0)\},$$

where $\widehat{I}(\hat{\theta}_1 : \hat{\theta}_0)$ is an estimator of the information gain when allowing for dependence,

$$\widehat{I}(\hat{\theta}_1 : \hat{\theta}_0) = \frac{2}{n}\left(\sum_{k=1}^{n} \log f(\mathbf{x}_k, \mathbf{y}_k; \hat{\theta}_1) - \sum_{k=1}^{n} \log f(\mathbf{x}_k, \mathbf{y}_k; \hat{\theta}_0)\right),$$

and $\hat{\theta}_1$ and $\hat{\theta}_0$ are the maximum likelihood estimates under the parameter spaces Θ_1 and Θ_0, respectively.

We also present two confidence intervals for the joint correlation coefficient: one when $I(\theta_1 : \theta_0)$ is large,

$$\left[1 - \exp\left\{-\widehat{I}(\hat{\theta}_1 : \hat{\theta}_0) + \sqrt{s^2\chi^2_{1;\alpha}/n}\right\},\right.$$
$$\left. 1 - \exp\left\{-\widehat{I}(\hat{\theta}_1 : \hat{\theta}_0) - \sqrt{s^2\chi^2_{1;\alpha}/n}\right\}\right],$$

where s^2 is the sample variance of $2\log\{f(\mathbf{x}_j, \mathbf{y}_j; \hat{\theta}_1)/f(\mathbf{x}_j, \mathbf{y}_j; \hat{\theta}_0)\}$ and $\chi^2_{1;\alpha}$ is the upper α quantile of the χ^2_1 distribution; and one when $I(\theta_1 : \theta_0)$ is small,

$$\left[1 - \exp\left\{-\frac{\varkappa_{1;\alpha/2}(\hat{a})}{n}\right\}, 1 - \exp\left\{-\frac{\delta_{1;\alpha/2}(\hat{a})}{n}\right\}\right],$$

where $\varkappa_{1;\alpha/2}$ and $\delta_{1;\alpha/2}$ are non-centrality parameters of certain χ^2_1 distributions and $\hat{a} = n\widehat{\Gamma}(\hat{\theta}_1 : \hat{\theta}_0)$.

Using a Monte Carlo study, we compare the empirical confidence coefficients of the two intervals for a number of models. The random variates are generated by means of the method described in Paper III. It is apparent for the models that we have examined that the "small" confidence interval (based on non-central χ^2-distributions) will produce the smaller intervals, yielding an empirical confidence coefficient for almost all models of approximately 95 %, when the nominal confidence coefficient is 95 %. The "large" confidence intervals will in general be wider.

We also examine a bias correction, suggested by Kent (1983), of the information gain estimator. This correction involves the second derivative of the binomial coefficient

$$\frac{\mathrm{d}^2}{\mathrm{d}r^2}\binom{r}{n},$$

and an expression for this is given in the appendix of that paper. In our examples, however, the suggested correction actually yields estimates that are more biased than the uncorrected ones. We believe that this might be due to numerical issues, as the correction involves a large number of infinite sums. Due to this lack of improvement we have not used this bias correction in our estimations.

As an example we have also estimated the general measure of correlation for GDP data from the 50 U.S. states and District of Columbia. The estimate of the general measure of correlation is

$$\hat{\rho}^2_J = 0.3027,$$

with a "small" confidence interval of

$$(0.0993, 0.5371)$$

thus indicating that composition of the government GDP in 1967 is correlated with that in 1997.

References

Aguilar, L. and C. Barceló-Vidal (2008, May). Multivariate ARIMA compositional time series analysis. In J. Daunis i Estadella and J. Martín-Fernández (Eds.), *Proceedings of CoDaWork'08, The 3rd Compositional Data Analysis Workshop, CD-ROM*. Univeristy of Girona, Girona (Spain).

Aguilar Zuil, L., C. Barceló-Vidal, and J. M. Larrosa (2007). Compositional time series analysis: A review. In *Proceedings of the 56th Session of the ISI (ISI 2007)*, Lisboa, August 22-29.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. London: Chapman and Hall.

Aitchison, J. (2003). *The Statistical Analysis of Compositional Data*. Caldwell, NJ: The Blackburn Press.

Aitchison, J. and J. J. Egozcue (2005). Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology 37*(7), 829–850.

Barceló-Vidal, C., L. Aguilar, and J. Martín-Fernández (2007). Time series of compostional data: A first approach. In *Proceedings of the 22nd International Workshop of Statistical Modelling (IWSM 2007), Barcelona, July 2-6*, pp. 81–86.

Bergman, J. (2008, May). Compositional time series: An application. In J. Daunis i Estadella and J. Martín-Fernández (Eds.), *Proceedings of CoDaWork'08, The 3rd Compositional Data Analysis Workshop, CD-ROM*. Univeristy of Girona, Girona (Spain).

Billheimer, D. and P. Guttorp (1995, Nov). Spatial models for discrete compositional data. Technical report, Dept. of Statistics, University of Washington, Seattle.

Billheimer, D., P. Guttorp, and W. F. Fagan (1997). Statistical analysis and interpretation of discrete compositional data. Technical Report Series 11, NRSCE.

Brunsdon, T. M. (1987). *Time series analysis of compositional data*. Ph. D. thesis, Dept. of Mathematics, University of Southampton.

Brunsdon, T. M. and T. M. F. Smith (1998). The time series analysis of compositional data. *Journal of Official Statistics 14*(3), 237–253.

Dodge, Y. (Ed.) (2003). *The Oxford Dictionary of Statistical Terms* (6th ed.). Oxford: Oxford University Press.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*(3), 279–300.

Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika 70*(1), 163–173.

Larrosa, J. M. (2005, Oct). Compositional time series: Past and present. Technical Report Econometrics 0510002, EconWPA.

Mateu-Figueras, G. and V. Pawlowsky-Glahn (2005, October). The Dirichlet distribution with respect to the Aitchison measure on the simplex - a first approach. In G. Mateu-Figueras and C. Barceló-Vidal (Eds.), *Proceedings of CoDaWork'05, The 2nd Compositional Data Analysis Workshop*. Universitat de Girona.

Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. In S. Thió-Henestrosa and J. A. Martín-Fernández (Eds.), *Proceedings of CoDaWork'03, Compositional Data Analysis Workshop*. Universitat de Girona.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment 15*(5), 384–398.

Pawlowsky-Glahn, V. and J. J. Egozcue (2002). BLU estimators and compositional data. *Mathematical Geology 34*(3), 259–274.

Pearson, K. (1897). 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*, 489–498.

Rodriguez, R. N. (1982). Correlation. In S. Kotz and N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, Volume 2. New York: John Wiley & Sons.

Smith, T. M. F. and T. M. Brunsdon (1989). The time series analysis of compositional data. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 26–32.