International Journal of Business Analytics & Intelligence
9 (1 & 2) 2021, 01-05
http://publishingindia.com/ijbai/

Beyond the Pie Diagram : Analysing Compositional Data

Arnab Kumar Laha*

Compositional data abound in our lives! If you are a newspaper reader, it would not be surprising that you encountered this kind of data while reading the newspaper today! You may read about the proportions of various ingredients of a food item, or that of a medicine, or about the proportional representation of different societal groups in a legislature, or the proportion of value added by different sectors to the economy etc. The following are the characteristic features of compositional data - (a) there are finitely many components (i.e. categories) (b) each component is a positive number and (c) the total sum of the components is a constant (typically 1 or 100 or 1 million). A widely used graphical representation of compositional data is the Pie Diagram.

There are a variety of subject areas which needs to analyse compositional data for deriving insights. Some examples are Geology, Economics, Medicine, Food industry, Chemistry, Ecology, Agriculture, Sociology, Environmental sciences, Sports, Education, Public policy and Transportation. The well-known statistical methods cannot be used without modification for analysing compositional data. This is primarily due to the constraint that the total sum of the components is a constant. Specifically, suppose there are k components which are represented as $(X_1, X_2, ..., X_k)$ with the constraint that $X_1 + X_2 + ... + X_k = c$. Since,

$$Cov(X_1, X_1 + ... + X_k) = Cov(X_1, c) = 0$$

we get

$$Cov(X_1, X_2) + ... + Cov(X_1, X_k) = -Var(X_1).$$

This simple result warns us that usual statistical analysis with compositional data may lead us to misleading results because of the spurious correlation induced by the constraint. The emerging field of Compositional Data Analysis (CoDA) provides sound statistical methods for analysis of such data.

A key observation regarding compositional data is that these observations are points within a simplex. Without loss of generality, we take c = 1. When k=2, the simplex is $x_1 + x_2 = 1, x_1 > 0, x_2 > 0$, $x_1 + x_2 = 1, x_1 > 0, x_2 > 0$, which is a line segment joining the points (1,0) and (0,1) in the 2-dimensional coordinate plane (R^2) with the end-points excluded. Thus when k=2, every compositional observation would be a point on this line segment. When k=3, the simplex is $x_1 + x_2 + x_3 = 1, x_1, x_2, x_3 > 0$. This can be visualized as the equilateral triangle in R^3 with vertices at (1,0,0), (0,1,0) and (0,0,1) with the 3 sides of the triangle excluded. In general, when there are k components the compositional observations are points belonging to the set (called the standard (k-1)-simplex)

$$\Delta^{(k-1)} = \{(x_1,...,x_k) : \sum_{i=1}^{k} x_i = 1, x_i > 0 \text{ for } i = 1,...,k\}$$

It was not until the 1980s, that theoretically sound methods for analyzing compositional data were developed. John Aitchison pioneered the development of these methods based on log-ratios. Let $y_i = \ln\left(\frac{x_i}{x_k}\right), i = 1,...,(k-1)$. Note that $(x_1,...,x_k) \to (y_1,...,y_{k-1})$ is an one-one transformation and the x_i's can be obtained from the y_i's using the transformation $T = \sqrt{(x_{P_0} - x_{P_1})^2 + (y_{P_0} - y_{P_1})^2}$, for i = 1, ..., $(k - 1)$ and $x_k = \frac{1}{1 + \sum_{i=1}^{k-1} e^{y_i}}$. Further, note that $-\infty < y_i < \infty$ for all i = 1, ..., $(k - 1)$. This allows us an opportunity to use the well-known techniques of multivariate statistical analysis on the transformed data $(y_1, ..., y_{k-1})$.

* Indian Institute of Management Ahmedabad, Gujarat, India. Email: arnab@iima.ac.in

We next discuss the Aitchison geometry of compositional data. Towards this we begin by defining the notion of "compositional equivalence". Let $x = (x_1, ..., x_k)$ and $y = (y_1, ..., y_k)$ be 2 vectors with all positive components. We say x and y are compositionally equivalent if $x = \lambda y$ for some positive constant λ. Note that if we define $x \sim y$ if x and y are compositionally equivalent then ~ is an equivalence relation. Thus, ~ decomposes the positive orthant of R^k into a set of equivalence classes and every composition v can be thought to be the equivalence class containing v. Another important notion is that of "closure". The closure of x to θ > 0 is defined as $C(x) = (\frac{\theta x_1}{\sum_{i=1}^k x_i}, \ldots, \frac{\theta x_k}{\sum_{i=1}^k x_i})$. Two common choices of θ are θ = 1 and θ = 100.

We now define two fundamental operations on compositional data namely, Perturbation and Powering. If $x = (x_1, ..., x_k)$ and $y = (y_1, ..., y_k)$ and are 2 compositions then we define the Perturbation of x by y as $x \oplus y$ as the closure (to θ = 1) of $(x_1y_1, ..., x_ky_k)$ i.e.

$$x \oplus y = C((x_1y_1, ..., x_ky_k)).$$

Note that when $y = \left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$ we get $x \oplus y = x = y \oplus x$ i.e. $y = \left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$ i.e. is the identity perturbation. If we define $(-x) = C\left(\left(\frac{1}{x_1}, \ldots, \frac{1}{x_k}\right)\right)$ then $x \oplus (-x) = (\frac{1}{k}, \ldots, \frac{1}{k})$. Thus (-x) is the inverse perturbation of x.

The Powering of x by a constant is defined as

$$\alpha \odot x = C((x_1^\alpha, \ldots, x_k^\alpha)).$$

Note that $(-x) = (-1) \odot x$. Now, it is not difficult to check that $(\Delta^{(k-1)}, \oplus, \odot)$, is a vector space over the reals.

The Aitchison distance between two compositions x and y is defined as

$$d_a(x, y) = \sqrt{\frac{1}{2k} \sum_{i=1}^{k} \sum_{j=1}^{k} \left(ln \frac{x_i}{x_j} - ln \frac{y_i}{y_j} \right)^2}$$

and the Aitchison inner product of x and y is

$$< x, y >_a = \frac{1}{2k} \sum_{i=1}^{k} \sum_{i=1}^{k} ln \frac{x_i}{x_j} ln \frac{y_i}{y_j}.$$

Since $\sqrt{< x, x >_a} = ||x||_a = \sqrt{\frac{1}{2k} \sum_{i=1}^{k} \sum_{j=1}^{k} \left(ln \frac{x_i}{x_j} \right)^2}$, we can define the angle between the compositions x and y as

$$\xi = cos^{-1} \frac{< x, y >_a}{||x||_a ||y||_a}$$

It is not hard to check that with the above defined inner-product $(\Delta^{(k-1)}, \oplus, \odot)$ is a real Hilbert space.

Now, consider a dataset of n compositional observations, where each observation consists of k components $(x_{1,i}, x_{2,i} \ldots x_{k,i})$, $x_{1,i} + x_{2,i} + \ldots + x_{k,i} > 0$ for all $j = 1, ..., k$ and $i = 1, ..., n$. Let and i = 1,…,n. As noted above the transformed observations $(y_{1,i}, \ldots, y_{k-1,i}) \in R^{k-1}$. A well-known measure of central tendency in multivariate statistics is the multivariate mean $\bar{y} = (\overline{y_1}, \ldots, \overline{y_{k-1}})$ where $\overline{y_j} = \frac{1}{n} \sum_{i=1}^{n} y_{j,i}$, , j = 1,…, (k-1). The mean of the compositional data is defined $\bar{x} = (\overline{x_1}, \ldots, \overline{x_k})$ as where

$$\overline{x_j} = \frac{e^{\overline{y_j}}}{1 + \sum_{l=1}^{k-1} e^{\overline{y_l}}}, \text{ j = 1,…,(k-1) and } \overline{x_k} = \frac{1}{1 + \sum_{l=1}^{k-1} e^{\overline{y_l}}}$$

Since

$$\overline{y_j} = \frac{1}{n} \sum_{i=1}^{n} y_{j,i} = \frac{1}{n} \sum_{i=1}^{n} ln \left(\frac{x_{j,i}}{x_{k,i}} \right) = ln \left(\left(\frac{\prod_{i=1}^{n} x_{j,i}}{\prod_{i=1}^{n} x_{k,i}} \right)^{1/n} \right) = ln \left(\frac{GM(x_j)}{GM(x_k)} \right)$$

where $GM(x_1)$ denotes the geometric mean of the observations $x_{l,1}, \ldots, x_{l,n}$. Using this fact, we can simplify the expression of $\overline{x_j}$ as $\overline{x_j} = \frac{GM(x_j)}{\sum_{l=1}^{k} GM(x_l)}$. In other words,

$$\bar{x} = C(GM(x_1), \ldots, GM(x_k))$$

The Sample Total Variance (Totvar) of a dataset of n compositional observations is a measure of dispersion which is defined as

$$\text{Totvar} = \frac{1}{n} \sum_{k=1}^{n} d_a^2(x_k, \bar{x})$$

Clearly, Totvar = 0 if and only if all the observations are equal i.e. $x_1 = x_2 = \ldots = x_n$.

We now discuss 2 real-life examples of the use of compositional data analysis. Table 1 gives data regarding the proportion of reported COVID-19 cases for the 5 quarters Q1, 2020-21 to Q1, 2021-22 for various states/

UTs of India (Data source: www.covid19india.org). For example, for the state of Maharashtra only 3.1% of the total number of COVID-19 cases during this 5 quarter period happened during the Q1, 2020-21 whereas a whopping 54.5% of the cases happened during Q1, 2021-22. The intensity and pan-India nature of the second wave of COVID-19 can be seen from the fact that for almost all the states/UTs more than 50% of the total COVID-19 cases happened during Q1, 2021-22. The mean of this compositional dataset () is

	Q1/20-21	Q2/20-21	Q3/20-21	Q4/20-21	Q1/21-22
$\bar{x}$	0.013	0.192	0.138	0.027	0.629

This confirms the severity of the second wave of COVID-19 with the states on average experiencing more than 3 times the number of cases during the second wave than what they experienced during the first wave. The Sample Total variance (Totvar) is 1.154.

As a second example consider the data on the religious compositions of 60 different countries of the Asia-Pacific obtained from www.pewforum.org. Table 2 gives the data[1]. As we see there is significant variation in the religious compositions of the countries in this region. For summarising this information we use the compositional mean which is

Buddhists	Christians	Hindus	Muslims	Unaffiliated/Other Religions
4.53	62.69	1.53	9.23	22.02

The Sample Total Variance, Totvar= 4.708 which indicates that this data far more dispersed than the COVID-19 datadiscussed in the earlier example. In fact, a quick look at the data indicates that thisdatais not homogeneous. Most countries in this region has a dominant religion and the data can be viewed as aggregation of clusters where each cluster is determined by a dominant religion. Thus, instead of reporting a single compositional mean it would be more appropriate to first cluster the data using the Aitchison distance and report the summary statistics of each cluster separately.

Table 1: Proportion of Reported COVID-19 Cases for the Period Q1, 2020-21 to Q1, 2021-22 for Different States/ UTs of India

States/UT	*Q1/20-21*	*Q2/20-21*	*Q3/20-21*	*Q4/20-21*	*Q1/21-22*
Maharashtra	0.031	0.198	0.090	0.136	0.545
Karnataka	0.006	0.206	0.111	0.025	0.651
Andhra Pradesh	0.008	0.360	0.100	0.010	0.521
Tamilnadu	0.040	0.202	0.089	0.027	0.643
Uttar Pradesh	0.014	0.219	0.108	0.018	0.639
Kerala	0.002	0.066	0.192	0.128	0.613
Delhi	0.064	0.131	0.240	0.025	0.540
West Bengal	0.013	0.159	0.196	0.024	0.608
Odisha	0.008	0.235	0.122	0.013	0.622
Telengana	0.030	0.279	0.152	0.034	0.506
Bihar	0.015	0.239	0.096	0.018	0.632
Rajasthan	0.019	0.123	0.180	0.026	0.652
Assam	0.019	0.342	0.070	0.004	0.564
Chattisgarh	0.003	0.111	0.165	0.064	0.656

[1] Values are in percentages. <0.1% and >99.0% in the original data are set to 0.1% and 99.1%respectively. Subsequently the closures are computed to arrive at figures reported in the Table 2.

States/UT	Q1/20-21	Q2/20-21	Q3/20-21	Q4/20-21	Q1/21-22
Haryana	0.020	0.147	0.173	0.035	0.624
Gujarat	0.041	0.126	0.129	0.072	0.632
Madhya Pradesh	0.018	0.144	0.142	0.064	0.632
Punjab	0.010	0.182	0.087	0.116	0.606
Jharkhand	0.007	0.235	0.090	0.024	0.644
Jammu & Kashmir	0.025	0.214	0.144	0.031	0.586
Uttarakhand	0.009	0.135	0.121	0.029	0.706
Goa	0.009	0.192	0.105	0.041	0.653
Puducherry	0.007	0.229	0.090	0.027	0.647
Tripura	0.022	0.376	0.116	0.004	0.482
Himachal Pradesh	0.005	0.069	0.198	0.040	0.688
Manipur	0.019	0.143	0.252	0.019	0.568
Arunachal Pradesh	0.007	0.278	0.201	0.004	0.510
Chandigarh	0.007	0.186	0.125	0.111	0.571
Meghalaya	0.001	0.115	0.160	0.013	0.710
Nagaland	0.021	0.225	0.231	0.013	0.510
Ladakh	0.049	0.164	0.259	0.029	0.499
Anadaman & Nicober	0.013	0.500	0.148	0.016	0.323
Sikkim	0.005	0.141	0.146	0.018	0.690
Dadra & Nagar Haveli & Daman & Diu	0.024	0.265	0.030	0.022	0.659
Mizoram	0.008	0.094	0.114	0.014	0.769

Table 2: Religious Compositions of the 60 Countries in Asia-Pacific Region

Country	Buddhists	Christians	Hindus	Muslims	Unaffiliated/Other Religions
Afghanistan	0.1	0.1	0.1	99.1	0.6
American Samoa	0.1	98.3	0.1	0.1	1.4
Armenia	0.1	98.3	0.1	0.1	1.4
Australia	2.9	61.7	1.7	3.0	30.7
Azerbaijan	0.1	2.6	0.1	97.1	0.1
Bangladesh	0.1	0.1	8.2	90.8	0.8
Bhutan	74.7	0.1	22.5	0.1	2.6
Brunei	8.6	9.4	0.1	75.1	6.8
Cambodia	96.8	0.1	0.1	2.0	1.0
China	18.3	5.2	0.1	2.0	74.4
Cook Islands	0.1	96.0	0.1	0.1	3.7
Cyprus	0.1	72.3	0.1	25.0	2.5
Fiji	0.1	64.4	27.9	6.3	1.3
French Polynesia	0.1	94.0	0.1	0.1	5.7
Guam	1.1	94.2	0.1	0.1	4.5
Hong Kong	13.3	14.9	0.1	2.1	69.6
India	0.1	2.4	78.9	15.4	3.2
Indonesia	0.1	10.2	1.6	87.0	1.1
Iran	0.1	0.1	0.1	99.1	0.6
Japan	33.2	1.8	0.1	0.1	64.8

Country	*Buddhists*	*Christians*	*Hindus*	*Muslims*	*Unaffiliated/Other Religions*
Kazakhstan	0.1	23.1	0.1	72.0	4.7
Kiribati	0.1	97.0	0.1	0.1	2.7
Kyrgyzstan	0.1	10.0	0.1	89.4	0.4
Laos	64.0	1.5	0.1	0.1	34.3
Macau	17.3	7.2	0.1	0.1	75.3
Malaysia	15.7	9.4	5.8	66.1	3.0
Maldives	0.1	0.1	0.1	98.4	1.3
Marshall Islands	0.1	97.5	0.1	0.1	2.2
Federated States of Micronesia	0.1	95.3	0.1	0.1	4.4
Mongolia	54.4	2.3	0.1	3.4	39.8
Burma (Myanmar)	79.8	7.8	1.7	4.2	6.5
Nauru	1.1	79.0	0.1	0.1	19.7
Nepal	10.0	0.1	80.6	5.0	4.3
New Caledonia	0.1	85.2	0.1	2.8	11.8
New Zealand	1.9	52.9	2.5	1.6	41.1
Niue	0.1	96.4	0.1	0.1	3.3
North Korea	1.5	2.0	0.1	0.1	96.3
Northern Mariana Islands	10.6	81.3	0.1	0.1	7.9
Pakistan	0.1	1.6	1.9	96.3	0.1
Palau	0.1	86.7	0.1	0.1	13.0
Papua New Guinea	0.1	99.1	0.1	0.1	0.6
Philippines	0.1	92.4	0.1	5.7	1.7
Samoa	0.1	96.9	0.1	0.1	2.8
Singapore	32.2	17.7	6.5	16.1	27.5
Solomon Islands	0.1	97.4	0.1	0.1	2.3
South Korea	21.9	30.1	0.1	0.1	47.8
Sri Lanka	68.6	7.2	13.7	10.4	0.1
Taiwan	21.2	5.8	0.1	0.1	72.8
Tajikistan	0.1	1.8	0.1	96.4	1.6
Thailand	92.6	0.1	0.1	6.0	1.2
Timor-Leste	0.1	99.1	0.1	0.1	0.6
Tokelau	0.1	99.1	0.1	0.1	0.6
Tonga	0.1	98.9	0.1	0.1	0.8
Turkey	0.1	0.1	0.1	98.0	1.7
Turkmenistan	0.1	6.4	0.1	93.0	0.4
Tuvalu	0.1	96.7	0.1	0.1	3.0
Uzbekistan	0.1	2.0	0.1	97.1	0.7
Vanuatu	0.1	93.5	0.1	0.1	6.2
Vietnam	16.2	8.4	0.1	0.1	75.2
Wallis and Futuna	0.1	97.4	0.1	0.1	2.3

For readers who are interested in learning more about the fast growing field of compositional data analysis the references (Filzmoser, Hron & Templ, 2018) and (Pawlowsky-Glahn, Egozcue & Tolosana-Delgado, 2015) may be useful.

References

Filzmoser, P., Hron, K., & Templ, M. (2018). *Applied compositional data analysis with worked examples in R.* Springer.

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data.* Wiley.