Compositional Data Analysis (CoDA) Approaches to Distance in Information Retrieval

Paul Thomas
CSIRO, Australia
paul.thomas@csiro.au

David Lovell
CSIRO, Australia
david.lovell@csiro.au

ABSTRACT

Many techniques in information retrieval produce counts from a sample, and it is common to analyse these counts as proportions of the whole—term frequencies are a familiar example. Proportions carry only *relative* information and are not free to vary independently of one another: for the proportion of one term to increase, one or more others must decrease. These constraints are hallmarks of *compositional* data. While there has long been discussion in other fields of how such data should be analysed, to our knowledge, Compositional Data Analysis (CoDA) has not been considered in IR.

In this work we explore compositional data in IR through the lens of distance measures, and demonstrate that common measures, naïve to compositions, have some undesirable properties which can be avoided with composition-aware measures. As a practical example, these measures are shown to improve clustering.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval—*clustering, retrieval models*.

Keywords

Aitchison's distance; compositions; distance; similarity, ratio

1. COMPOSITIONAL DATA

Compositional data analysis (CoDA) deals with data that carry only relative information, such as proportions, percentages, and probabilities [1]. Information is carried only in the *ratios* of different components. Examples include the chemical composition of rocks, the fraction of a certain DNA sequence in a sample, nutritional content of food, or employment data by industry. In text processing, compositional data can include genres, documents, or probability distributions.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609492.

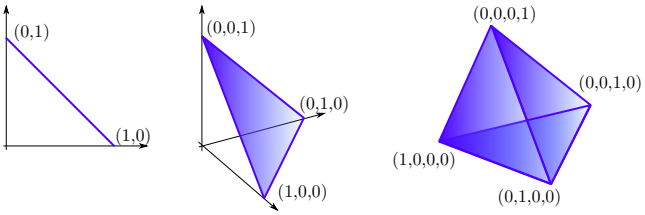

Figure 1: The set of vectors with D positive components that sum to the constant κ form the D-part simplex, denoted by $\mathcal{S}^D$ [3]. From left to right, and with $\kappa = 1$, we see $\mathcal{S}^2$, $\mathcal{S}^3$ and $\mathcal{S}^4$.

In a compositional analysis the absolute size—mass of a rock, amount of DNA in the sample, amount of food, length of document vector—is not of interest. Instead, the interest is in the *relative* information, the proportions of the whole.

We can illustrate these ideas with a simple example: the content of soils. Suppose a 50 g soil sample is found to contain 38 g of sand; 10 g of silt; and 2 g of clay. This three-component datum, $(38, 10, 2)$, is the *basis vector*. This vector is sum-constrained: if we have less sand, we must have more silt and/or clay to make up the 50 g. Now, since we are interested in the relative information but not the size of the sample, it is easy to constrain (or "close") the basis vector to the equivalent $(0.76, 0.20, 0.04)$. This is a *composition*: a vector of D parts, $(x_1, \ldots, x_D)$, that sum to a constant κ which, in this case, we choose to be 1.

Geometrically, compositions are points on a D-part *simplex* (Figure 1). CoDA theory, as pioneered by Aitchison [1] and developed by Pawlowsky-Glahn, Egozcue and others [2] is founded on logratio transformations that map the simplex $\mathcal{S}^D$ onto $\mathbb{R}^D$ where the full arsenal of statistical methods can be applied. If necessary, the results from these methods can be back-transformed to the simplex, secure in the knowledge that the constant-sum constraint will be satisfied.

Compositional data is common in IR (Section 2), but compositional data *analysis* is not. This paper aims to raise awareness of CoDA approaches, focusing especially on distances and dissimilarities common in IR and their compositional counterparts (Sections 3 and 4). Section 5 explores how these distances perform in a document clustering task.

2. COMPOSITIONS IN IR

Compositions are common in information retrieval for various reasons. IR often deals with fixed-size samples: the genres of pages on a website, for example, or the terms in

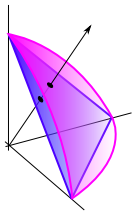

Figure 2: $\mathcal{S}^3$ and the positive orthant of the hypersphere. Cosine similarity is unchanged whether we project onto the hypersphere or the simplex.

1000 scientific abstracts, where the sum is constrained to the size of the sample. We may have proportions, when the denominator is not known but the sum is constrained to 100% (or 1). Further, even individual documents or queries can be considered compositions. This can happen directly, if we are using probabilistic language models; or as an alternative representation, if we are using vector-space (geometric) models.

Count data. Count data are often constrained in IR settings. If we count the number of in- and out-links for a web page, the sum is constrained to that page's degree. If we count the number of nouns, verbs, or specific terms, in a 10,000-word sample, the sum is constrained to 10,000.

While observational count data does not carry purely relative information[1] it is often modelled and analysed using compositional methods. (Expected counts, on the other hand, do carry only relative information but, in general, these are not known to us and have to be estimated.)

Multinomial and Dirichlet distributions. Both of these distributions play important roles in modelling discrete data. Clearly, the vector of event probabilities in a multinomial probability distribution, and the vector of concentration parameters in a Dirichlet distributions are compositional, summing to one over the set of categories they describe.

In IR, multinomial and Dirichlet distributions are often used as the basis of probabilistic document models, topic and topic/term distributions, and Markov transition probabilities.

We note that CoDA offers a more flexible alternative to the Dirichlet distribution: the *logistic normal*, which allows dependencies between components to be expressed and opens up opens up the full range of linear modelling available for the multivariate Normal distribution in $\mathbb{R}^D$ [2].

Vector-space models. Documents are often represented as points in vector-space where each dimension corresponds to a term, and the coordinate value represents the count of that term in the document. In IR, interest commonly lies in the *relative* abundance of different terms, rather than the actual counts observed since the former relates more strongly to the meaning and content of a document. For this reason, count vectors are generally normalised to a constant magnitude so that documents of different lengths become meaningfully comparable.

[1]"Two heads and one tail" in a coin toss experiment tell a different story than "two thousand heads and one thousand tails", even though the ratio of heads and tails is the same in both instances.

While this has many of the hallmarks of compositional data, IR often uses *cosine dissimilarity* as a means to compare documents, an approach based on transforming the data to the positive orthant of the D-dimensional unit hypersphere, rather than working within the D-dimensional simplex. Figure 2 illustrates that the cosine dissimilarity is unaffected by the magnitude of vectors, only their direction. However, as we shall see, distances on the hypersphere are constrained—there are limits to how different two documents can be in terms of cosine dissimilarity—whereas the logratio transformation used in CoDA enforces no such limits.

Transforming compositional data to the hypersphere is a concept that has been discussed and debated within CoDA; it appeals because it handles zero values simply but lacks the robust theoretical foundation of the logratio approach [7].

3. FIVE DISTANCES FOR IR DATA

Conceptualising documents as points on the simplex opens up similarity and distance measures that are new to IR. As well as the familiar cosine and Euclidean measures, Aitchison's (compositional) distance can be applied, as can Euclidean distance between log-transformed data and Kullback-Leibler divergence [5, 6]. In this section we will quickly review these distances and dissimilarities, then look at their behaviours and implications for information retrieval.

Distances are measured between two objects, here two documents. We will treat these as basis vectors $\mathbf{X}$ and $\mathbf{Y}$, where each element X_i and Y_i is a count of terms. We will use $\mathbf{x}$ and $\mathbf{y}$ to refer to the same two documents, as compositions: i.e. $\mathbf{x}$ is a vector such that $\sum_i x_i = 1$.

In addition, we require all x_i to be strictly positive to avoid the issue of dealing with zeros which, for approaches using logarithms or ratios, is a research topic in its own right [7]. Various language modelling approaches, including Bayesian methods, can ensure that there are no zeros in IR data [9].

Cosine dissimilarity. Vector-space models of text typically measure similarity between two documents by the cosine of the angle between them. Cosine *dis*similarity can be defined:

$$d_{\mathrm{C}}(\mathbf{x},\mathbf{y}) \quad = \quad 1 - \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} \ .$$

Euclidean distance. Euclidean distance is the length of the shortest line segment connecting our two points, and is straightforward (note that here we use squared distance):

$$d^2_{\mathrm{E}}(\mathbf{x},\mathbf{y}) \quad = \quad \sum_i (x_i - y_i)^2 \ .$$

On the simplex d_{E} is bounded by $[0, \sqrt{2}]$: two points can be at most $\sqrt{2}$ units apart and that happens when each is at a different "corner". In the case of high-dimensional data such as text, it is also clear that differences in any one component will count for little as they are "drowned out" by the large number of other components.

Euclidean distance of log-transformed data. When dealing with count data, especially counts which vary over several orders of magnitude, it is common to use a logarithmic transformation. This leads to a distance of the form:

$$d^2_{\mathrm{EL}}(\mathbf{x},\mathbf{y}) \quad = \quad \sum_i (\log x_i - \log y_i)^2 \ .$$

X	Y	Cosine	Euclidean	Euclidean log	Kullback-Leibler	Aitchison
(1, 10)	(2, 10)	0.00477	0.107	0.612	0.0525	0.490
(1, 100)	(2, 100)	5×10^{-5}	0.0137	0.682	6.73×10^{-3}	0.490
(1, 1000)	(2, 1000)	5×10^{-7}	1.41×10^{-3}	0.692	6.91×10^{-4}	0.490
(1, 10000)	(2, 10000)	5×10^{-9}	1.41×10^{-4}	0.693	6.93×10^{-5}	0.490

Table 1: Some distances and divergences. Modified from Lovell et al. [5].

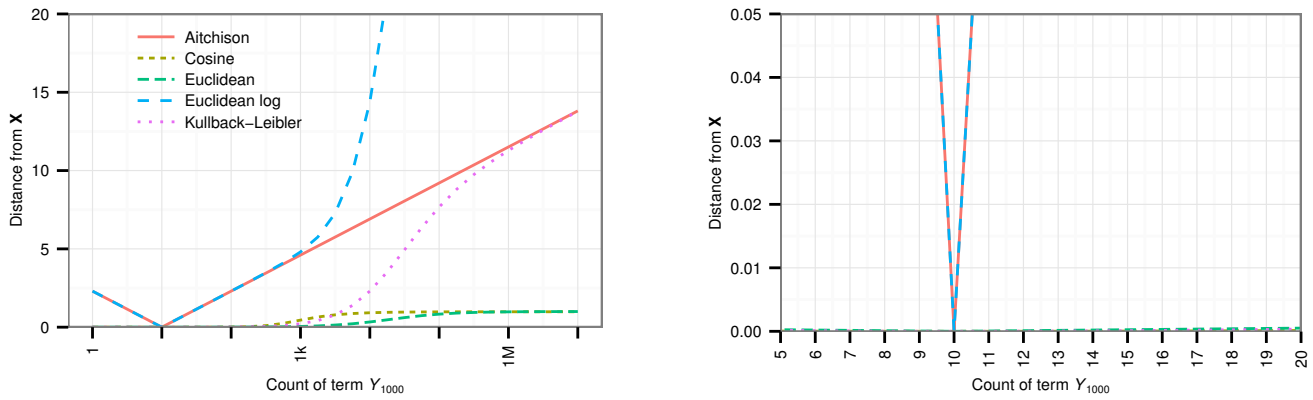

Figure 3: Distances between 1000-term synthetic documents $\mathbf{X} = (20, 20, \ldots, 10)$ and $\mathbf{Y} = (20, 20, \ldots, Y_{1000})$. The plots show how the different distances vary as Y_{1000} changes from 10^0 to 10^7. The right-hand plot gives a zoomed-in view.

Note that, unlike plain Euclidean distance, there is no upper limit to the distance between *log-transformed* components.

Kullback-Leibler divergence. Kullback-Leibler divergence can express the dissimilarity between two probability distributions [4]. It is not strictly a distance metric, but is commonly used in information retrieval to compare two sets of normalised counts—that is, to compare two compositions [8]. In its symmetric form it is written:

$$\begin{aligned} d_{\mathrm{KL}}(\mathbf{x}, \mathbf{y}) &= \sum_i \left(x_i \log \frac{x_i}{y_i} + y_i \log \frac{y_i}{x_i} \right) \\ &= \sum_i (x_i - y_i)(\log x_i - \log y_i) \; . \end{aligned}$$

Kullback-Leibler divergence depends on both the ratio of compositional components *and* their absolute values.

Aitchison's distance. This distance depends solely on the ratios of components and can be written:

$$d^2_{\mathrm{A}}(\mathbf{x}, \mathbf{y}) = \frac{1}{D} \sum_{i<j} \left(\log \frac{x_i}{x_j} - \log \frac{y_i}{y_j} \right)^2 \; .$$

Relationships between distances. If vectors $\mathbf{X}$ and $\mathbf{Y}$ are normalised to lie on the unit hypersphere (as is often done in IR applications) then $d^2_{\mathrm{E}}(\mathbf{X}, \mathbf{Y}) = 2 d_{\mathrm{C}}(\mathbf{X}, \mathbf{Y})$. This does not hold for vectors on the unit *simplex*.

By Napier's inequality, $d^2_{\mathrm{E}}(\mathbf{x}, \mathbf{y}) < d_{\mathrm{KL}}(\mathbf{x}, \mathbf{y}) < d^2_{\mathrm{EL}}(\mathbf{x}, \mathbf{y})$, and, as shown by Lovell et al. [5], $d^2_{\mathrm{A}}(\mathbf{x}, \mathbf{y}) \leq d^2_{\mathrm{EL}}(\mathbf{x}, \mathbf{y})$. Note that both Aitchison and Euclidean-log distances are invariant to re-scaling (i.e., weighting) of components, such as by the inverse document frequency of terms.

4. BEHAVIOURS OF DISTANCES

Table 1 uses vectors with two components as a simple illustration of how various distances and dissimilarities behave. Vectors $\mathbf{X}$ and $\mathbf{Y}$ differ only in their first component by a factor of 2. Aitchison's distance depends only on the ratios of components, that is $X_1 : Y_1$ and $X_2 : Y_2$: these ratios are always the same here and so d_{A} does not change. Other measures depend on both these ratios and the absolute values of each component, and vary from case to case.

Figure 3 shows further aspects of behaviour, based on the distances between two fictional documents. Here $\mathbf{X}$ has 999 "noise" terms repeated 20 times each, and one rarer "signal" term repeated ten times, that is $\mathbf{X} = (20, 20, \ldots, 10)$. $\mathbf{Y}$ has the same first 999 "noise" terms, but the final "signal" is varied from one occurrence up to ten million ($Y_{1000} \in [1, 10^7]$). As the "signal" grows large, both cosine and Euclidean distance saturate and do not change appreciably although counts vary over several orders of magnitude.

The right-hand graph illustrates another behaviour. As we halve or double the amount of the "signal" term, from 5 through 10 to 20, intuitively it seems that distances should differ. However, the cosine, Euclidean, and Kullback-Leibler measures barely change: the less-important terms "drown out" the more-important one.

We have several apparently undesirable behaviours, displayed by three of the five metrics at different times: (1) the distance between two documents which vary only in one component depends on the size of the other (nonvarying) components—but these other components are presumably not interesting; (2) the distance between documents which vary only in one component depends on the number of dimensions; (3) the distance can saturate; (4) the distance is not sensitive to even quite large relative changes in the counts

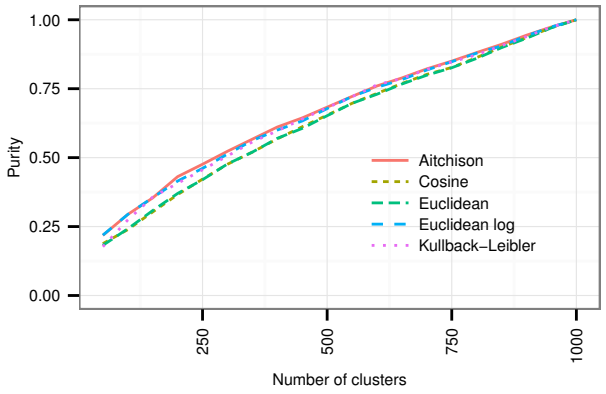

Figure 4: Cluster purity as number of clusters varies, using different distance measures to form the clusters. With this data, clusters formed using d_{EL}, d_{A}, or d_{KL} are better than those formed with d_{E} or d_{C}.

of uncommon terms, although these changes are presumably interesting.

5. DISTANCES IN PRACTICE

Thus far, we have looked at the theoretical properties of distances and dissimilarities commonly used in IR and CoDA. This section explores their use in a typical IR task—clustering—in which performance depends on both the distribution of the data to be clustered and the way in which we convert differences between multivariate data points into a univariate distance or dissimilarity. The aim here is not some kind of "performance shoot out", rather than to see whether and how these measures affect distance-based analyses.

We explore the behaviour of the distances and dissimilarities presented so far by using them to cluster Usenet articles from the 20-newsgroups collection. These experiments used 50 articles from each of 20 groups, without headers such as Newsgroups: or Path: which would identify the group, and with no stemming or stopping.[2] Heirarchical clustering used Ward's method, as implemented in R, and a tree was built using each of the distance measures above.

We report the quality of the clustering with purity. Purity for a clustering of 1000 documents is the mean proportion of documents in each cluster which are in the majority class: purity $= {}^1\!/_{1000} \sum_i \max_j |c_i \cap g_j|$, where c_i is the ith cluster and g_j is the jth newsgroup. Purity ranges up to 1, and scores highest when each cluster consists entirely of documents from the same newsgroup.

Figure 4 plots purity when the heirarchies formed with each distance measure were cut to 20, 40, ... 1000 clusters. At 1000 clusters, each cluster is a single document and purity must be 1; but at all other points, different distance measures produce clusterings of different quality.

d_{C} and d_{E} produce very similar distances, and hence very similar clusterings; d_{EL} and d_{A} are also related, and d_{KL} sits somewhere in between. Aitchison's distance produces somewhat better clusters than the commonly-used cosine and Euclidean, with purity a relative 7% better than d_{C} across all cuts and as much as 20% better when fewer clusters are asked for. Clustering with d_{A} never produces worse results than clustering with the standard measures.

Similar effects were seen with different numbers of documents from the same collection, and using the Rand index instead of purity. We also saw similar effects building clusters with complete linkage, although no effect using single linkage.

[2]This is the "bydate" collection from Jason Rennie, `http://qwone.com/~jason/20Newsgroups/`.

6. DISCUSSION AND DIRECTIONS

CoDA is relevant to IR. It provides a well-founded theoretical basis for analysing data on the relative abundance of document topics and terms, and gives a distance metric that depends solely on these relative abundances.

Our initial clustering experiment suggests logarithm- and ratio-based distances (d_{A}, d_{EL}, d_{KL}) warrant further investigation as alternatives to distances on the hypersphere (d_{E}, d_{C}).

Our hope is that this paper will encourage exploration of CoDA methods on other IR data sets and tasks, including ranking. We note that both d_{A} and d_{EL} are invariant to re-scaling of components, such as weighting by inverse document frequency. Our suspicion is that idf-weighting will not improve d_{E} and d_{C} to the point that they surpass logarithm- and ratio-based distances in clustering performance, but this too demands further experiments to verify.

7. REFERENCES

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

[2] J. Bacon-Shone. A short history of compositional data analysis. In V. Pawlowsky-Glahn and A. Buccianti, editors, *Compositional Data Analysis*, pages 1–11. John Wiley & Sons, Ltd, 2011.

[3] J. J. Egozcue and V. Pawlowsky-Glahn. Basic concepts and procedures. In V. Pawlowsky-Glahn and A. Buccianti, editors, *Compositional Data Analysis*, pages 12–28. John Wiley & Sons, Ltd, 2011.

[4] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[5] D. Lovell, W. Müller, J. Taylor, A. Zwart, and C. Helliwell. Caution! Compositions! Can constraints on omics data lead analyses astray? Technical Report EP10994, CSIRO, Mar. 2010.

[6] J. A. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn. Measures of difference for compositional data and hierarchical clustering methods. In *Proc. Int. Assoc. for Mathematical Geology*, pages 526–531, 1998.

[7] J. A. Martín-Fernández, J. Palarea-Albaladejo, and R. A. Olea. Dealing with zeros. In V. Pawlowsky-Glahn and A. Buccianti, editors, *Compositional Data Analysis*, pages 43–58. John Wiley & Sons, Ltd, 2011.

[8] J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *Proc. ACM SIGIR*, pages 254–261, 1999.

[9] C. Zhai. *Statistical Language Models for Information Retrieval: A Critical Review*, volume 2 of *Foundations and Trends in Information Retrieval*. now Publishers, Delft, 2008.