Caution! Compositions! Can constraints on omics data lead analyses astray?

Article · March 2010

CITATIONS
13

READS
630

5 authors, including:

David Roger Lovell
Queensland University of Technology
72 PUBLICATIONS **1,074** CITATIONS
SEE PROFILE

Jennifer M Taylor
The Commonwealth Scientific and Industrial Research Organisation
137 PUBLICATIONS **4,836** CITATIONS
SEE PROFILE

Chris A Helliwell
The Commonwealth Scientific and Industrial Research Organisation
112 PUBLICATIONS **7,836** CITATIONS
SEE PROFILE

www.csiro.au

Caution! Compositions!
Can constraints on omics data lead analyses astray?

David Lovell, Warren Müller, Jen Taylor, Alec Zwart and Chris Helliwell

Report Number: EP10994

20 March 2010

Enquiries should be addressed to:

David Lovell
Transformational Biology - Bioinformatics and Analytics Leader
CSIRO Mathematical and Information Sciences
GPO Box 664, ACT 2601
+61 2 6216 7042 (w) +61 2 6216 7111 (f) +61 419 167 136 (m)
email : David.Lovell@csiro.au

Distribution list

Client	(3)
Publications Officer	(1)
Stream Leader	(1)
Authors	(2)

Contents

Abstract

Some DNA or RNA sequencing methods produce data that can be considered as counts of the number of times each sequence was observed in the biological sample. Because the sum of these counts within each sample is constrained by the sequencing process and the physical amount of the sample, they constitute *compositional data* [1]. There are many other examples of compositional data in "the omics", including relative abundances of species (in metagenomics) or GO terms (in functional genomics). Compositional data need not be limited to counts: there are constraints on microarray data by virtue of being measured from a fixed amount of total RNA.

Few researchers have broached the issue of analysis of compositional data in the omics, but in the geosciences there has been debate for nearly half a century about how sum-constrained data should be analysed. One important difference between "omics" and geosciences data is that molecular biology frequently produces compositions with tens- if not hundreds-of-*thousands* of components, whereas geosciences data usually has much lower dimension (tens to hundreds).

This report aims to raise awareness of whether, and in what situations, naïve analysis of sum-constrained data could lead to incorrect inference, and explore the extent to which this might be a problem in omics applications. In particular, we compare the analysis of log-transformed data to full compositional data analysis.

1 Introduction

Compositional data analysis has its roots in the geosciences where geologists faced a challenge of how to analyse and interpret measurements of the mineral content of rocks; samples would be described in terms of percentages of different components, or in parts per million (or billion) for trace elements.

In simple terms: "The most important characteristic of compositional data is that *they carry only relative information*" [2].

> "When in any discipline we say that a problem is compositional we are recognizing that the sizes of our specimens are irrelevant[1]. For example, a geologist talking about the composition of an object, such as the major oxide composition of a rock, is admitting that interest is in a dimensionless problem. There is no concern about whether the rock specimen weighs one gm or one lb. Similarly in the study of the dietary content of cows' milk interest will focus on the dietary composition—proportions by weight of the total dietary content of the parts - protein, milk fat, carbohydrate, calcium, sodium, potassium—rather than on the size of the milk sample. Such trivial admissions have far-reaching consequences." [3]

It can take a long time for deep methodological knowledge developed in one domain to be applied elsewhere. (Even within the geosciences, there is not yet unanimous agreement that compositional data warrants special attention.) We are motivated by the concern that there is a lot of compositional data in "the omics" (genomics, transcriptomics, metagenomics, *etc.*) but little awareness of the pitfalls of ignoring compositional constraints[2], or even that compositional constraints are at play. We are concerned that molecular biology not be led astray by findings that are more to do with artifacts of the measurement process than the biological process being measured.

So how widespread is compositional data in "the omics"? Examples include

- Fixed size/volume samples of different components
  - 1g of tissue (containing different kinds of cells)
  - 1μg of total RNA (containing different kinds of RNAs)
  - 1μg of metagenomic DNA (containing DNA from different genomes)
  - 1mL of blood (containing different concentrations of metabolites)
- Constrained counts
  - The numbers of different cells in a fixed size sample
  - The numbers of different bases in a fixed length DNA sequence
  - The numbers of different proteins in a fixed length DNA sequence

[1] Note that in the biosciences, the size (sometimes referred to as the *absolute abundance*) of specimens is often highly relevant to the topic being studied, but many methods of sample preparation or measurement remove information about size, leaving only relative information behind.

[2] We note that other biosciences are more familiar with compositions [4], including ecology [5]. We also note that areas of computer science (*e.g.*, information retrieval—IR) concerned with characterising and comparing symbol frequencies in strings seem unaware of compositional data analysis [6]: Kullback-Leibler divergence appears to be the prevailing approach to comparison of unit-sum-constrained data (*i.e.*, word-frequency histograms) in IR. We shall examine that non-metric measure of difference in Section 4.4.

- Proportions
  - The proportions of different k-mers in genomes
  - The proportions of gene ontology (GO) terms in samples
  - The proportions of different reads in next-gen sequencing runs.

Compositional data is commonplace in molecular biology; evidence of principled approaches to analysing this kind of data is scarce.

For that reason, we are keen to ensure that this report be meaningful to a multidisciplinary audience that includes biologists, bioinformaticians and biostatisticians. Different parts this report will be familiar to different disciplines, and our aim is to present unfamiliar material so that it can be understood, at least at a high level.

To assist readers who dislike mathematical detail, we shall use shaded boxes like this to emphasise key messages where there are lots of equations.

We begin with a thought experiment to highlight some important aspects of compositional data in an omics setting.

1.1 The Omics Imp

Small enough to fit into a cell, yet still somehow able to wield pencil and paper, The Omics Imp is a molecular accountant *par excellence*. Without disrupting biological processes, the Imp can tally the different molecules it observes.

Figure 1.1 shows it counting messenger RNA (mRNA) as it emerges from the nucleus of a cell. The Imp can help experimentalists by counting the different types of mRNA molecules it sees in a specified time interval. Clearly, these counts are non-negative and constrained only by productivity of the nucleus in the time interval. This vector of counts is known as a *basis*.

The Imp can also work in other styles of experiment. Figure 1.2 shows it counting messenger mRNA collected in a bucket *after* emerging from the nucleus of a cell. The Imp can help experimentalists by counting the different types of mRNA molecules it sees in this *full* bucket. Once again, these counts are non-negative but, unlike the scenario in Figure 1.1, they are constrained by the (arbitrary) size of the bucket, they carry no information about the *absolute* rate of mRNA production, and they are not independent of each other—if the amount of one kind of mRNA in the full bucket increases, the amounts of one or more other kinds of mRNA must decrease. The *sum-constrained* vector of counts produced in this experiment is known as a *composition*, and this constraint has a profound impact on both the information carried by the counts, and their subsequent interpretation.

Without The Omics Imp, most omics measurement processes follow the *bucket-survey* paradigm shown in Figure 1.2, often using a series of buckets (*e.g.*, extracting a fixed mass of total RNA), with filters (*e.g.*, RNA size fractionation by gel electrophoresis) and multipliers (*e.g.*, polymerase chain reaction (PCR) amplification) between them. However, it is easy to lose sight of the sum-constraints being placed on experimental material, particularly as there is no strong tradition of concern for these issues in molecular biology and bioinformatics.

We point out these two different conceptual styles of molecular biology experiments as a preamble to showing they demand different styles of analysis and interpretation.

The current raft of nucleotide-counting sequencing technologies also give the impression that a biologist can count—or at least estimate the count of—the different types of DNA or RNA sequences produced by a sample of cells. But some thought about the sample preparation and DNA/RNA extraction process should make it clear that there are a few different buckets constraining the numbers of molecules under measurement[3] including

- commencing with a fixed weight or volume tissue sample
- extracting a fixed weight or volume of DNA/RNA
- concluding with a finite (if very large) number of sequence fragment reads.

The terms *under-* and *over-expression* are often used in gene expression analysis to refer to mRNAs that are less/more expressed in comparison to some reference situation. These mRNAs are also described as being *down-/up-regulated* by processes that control their level of expression. Figure 1.3 emphasises the perils of conflating these terms with under- and over-*production*. The bucket-survey suggests that, in comparison to Treatment A, mRNA z is over-expressed in Treatment B, even though it is being produced at exactly the same rate in both situations. Figure 1.4 highlights the situation faced by molecular biologists using conventional approaches to

[3]Which is not to preclude the possibility of methods that could help us estimate the mRNA actually being produced by cells, as described by Kanno *et al.* [7] for example.

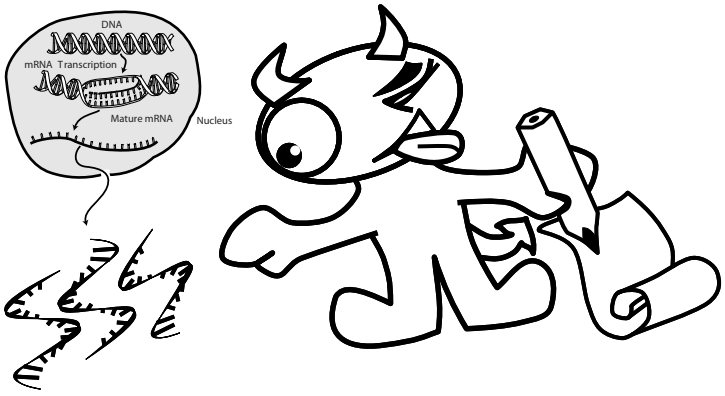

Figure 1.1: The Omics Imp tallying the different kinds of messenger RNA molecules emerging from the nucleus of a cell. (Illustration of mRNA courtesy: National Human Genome Research Institute.)

Figure 1.2: The Omics Imp tallying the different kinds of messenger RNA molecules in a full bucket collected from the nucleus of a cell. (Illustration of mRNA courtesy: National Human Genome Research Institute.)

omics data. To make comprehensive statements about gene expression, we have to know the amount of mRNA being produced (or, as Aitchison terms it, the *size* of the specimen) as well as the relative abundances of different mRNA species.

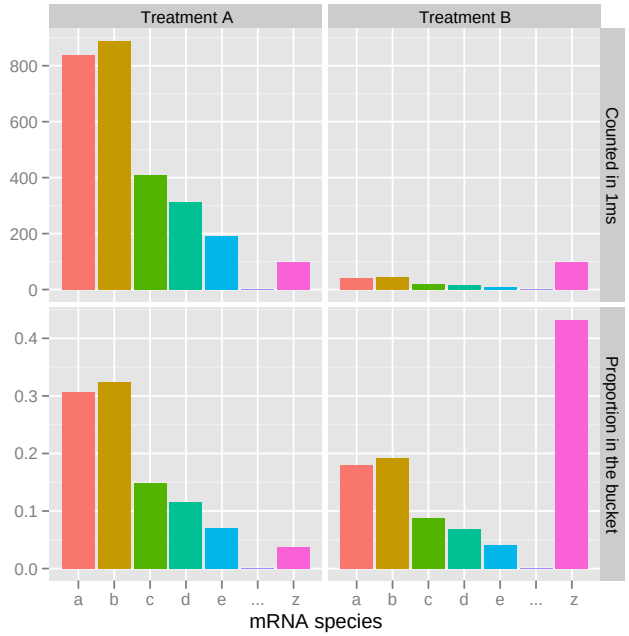

Figure 1.3: (Fictitious) data obtained by the Omics Imp on different mRNA species (a to z) from a cell nucleus under two different treatments (A and B). The top row shows counts of the mRNAs observed over 1ms. (Note that the Omics Imp counted exactly the same number of mRNA z in both treatments.) The bottom row shows the findings of the corresponding bucket-survey, expressed as the proportion of each mRNA species collected in the Imp's (full) bucket.

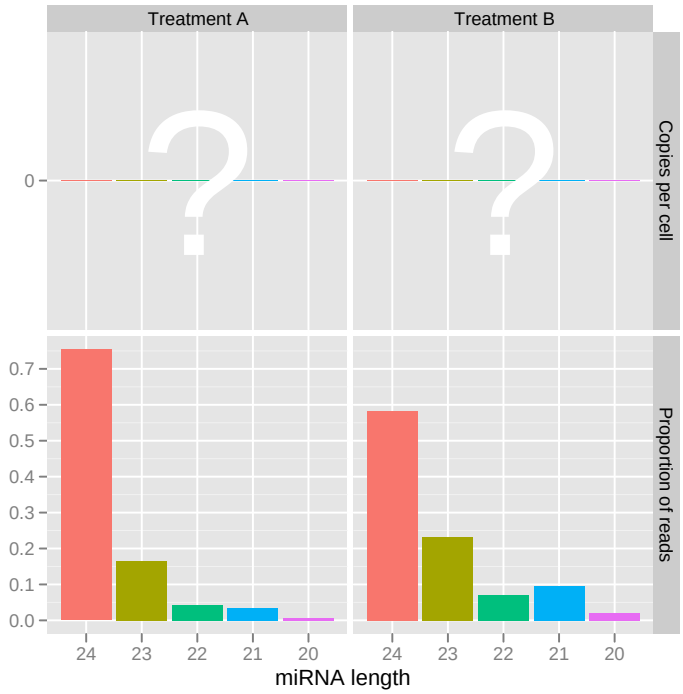

Figure 1.4: Real data obtained by Ms Ning Wang showing the number of micro RNAs (miRNAs) of lengths 24 to 20 obtained in two different sequencing runs (A and B). Without knowing the total number of miRNA molecules per cell in each treatment, we cannot translate the proportions in the bottom row into absolute abundances of the different lengths of miRNA.

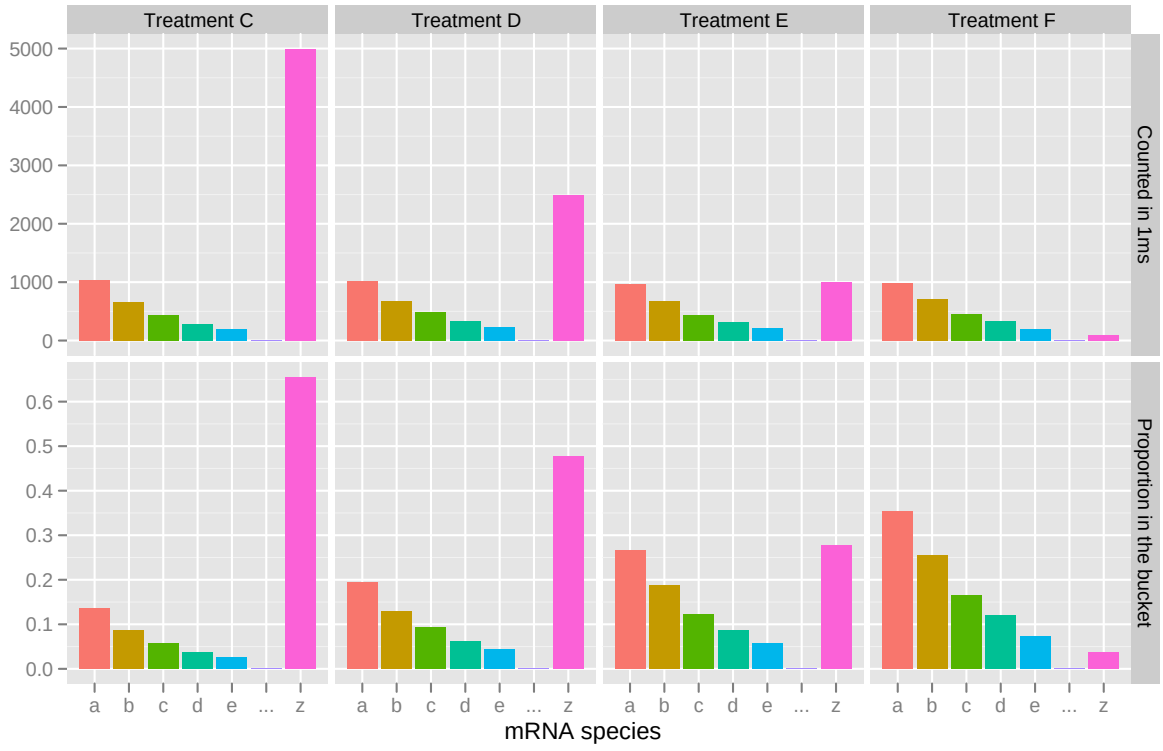

Figure 1.5: (Fictitious) data obtained by the Omics Imp on different mRNA species (a to z) from a cell nucleus under four different treatments (C to F). Again the top row shows counts of the mRNAs observed over 1ms. (Note that the Omics Imp counted approximately the same number of mRNAs a to e in all treatments.) The bottom row shows the findings of the corresponding bucket-survey, expressed as the proportion of each mRNA species collected in the Imp's (full) bucket.

Now we show the potential of relative abundance data to make statistically independent components appear correlated. In the fictitious data of Figure 1.3 the absolute amount of mRNA z remained constant in both Treatment A and B while mRNAs a to e changed dramatically. Figure 1.5 shows the opposite scenario: the absolute abundances of mRNAs a to e remain constant across Treatments C to F while mRNA z changes dramatically. A naïve interpretation of the mRNA proportions in the bucket would describe mRNAs a to e as positively correlated with each other and negatively correlated with mRNA z across the four treatments—the proportions of mRNAs a to e increase together while the proportion of mRNA z decreases. All this despite the fact that the absolute number of copies of mRNAs a to z are statistically independent in the four treatments. This is another manifestation of the sum constraint imposed by looking at the contents of a full bucket—we repeat: if the amount of one kind of mRNA in the full bucket increases, the amounts of one or more other kinds of mRNA must decrease.

Before we explore in more detail the impact that compositional constraints can have, we will define some terms to aid our exposition.

2 Definitions

These definitions are needed to put subsequent proofs and arguments on a sound mathematical footing. However, readers who prefer to skip the maths should still be able to get a *qualitative* appreciation of the potential for compositional data to lead naïve analyses astray. At the very least, readers should come to appreciate that proportions, percentages and parts per million ought not be analysed as though they were variables that are free to assume any value.

Compositional data is made up of non-negative components whose sum is constrained. For simplicity, this report deals with positive components that sum to 1.

A set of percentages is constrained to sum to 100; a set of parts per million is constrained to sum to 1,000,000.

Definition 1 (Composition) *A* composition $\mathbf{x}$ *of* D *parts is a* $D \times 1$ *vector* $(x_1, x_2, \ldots, x_D)$ *of positive components whose sum is 1.*

Figure 2.1 gives geometrical depictions of two-, three- and four-part compositions.

Compositions are often the result of scaling a vector of positive components known as a *basis*

Definition 2 (Basis) *A* basis $\mathbf{w}$ *of* D *parts is a* $D \times 1$ *vector* $(w_1, w_2, \ldots, w_D)$ *all recorded on the same measurement scale.*

The scaling process is known as *closure* or *constraining*.

Definition 3 (Closure) *The* closure operator $\mathcal{C}$ *transforms each vector* $\mathbf{w}$ *of* D *positive components into the unit-sum vector* $\mathbf{w}/\mathbf{w} \cdot \mathbf{j}$*, where* $\mathbf{j}$ *is the column vector of units.*

Turning a set of positive numbers into percentages of their total is a familiar example of closure. In RNA-seq, Mortazvi *et al.*[8] apply closure by working in terms of reads per kilobase of exon per million mapped sequence reads (RPKM).

The term $\mathbf{w} \cdot \mathbf{j}$ is known as the *size*, t, of $\mathbf{w}$.

Definition 4 (Size) *The* size *of a basis* $\mathbf{w}$ *is*

$$\begin{aligned} t &= w_1 + \cdots + w_D \\ &= \mathbf{w} \cdot \mathbf{j} \end{aligned}$$

There is a many-to-one-to-many relationship between a composition $\mathbf{x}$ and the bases from which it could have been derived.

Property 1 (Relationship of bases to composition) *The set of bases with composition* $\mathbf{x}$ *is*

$$\mathcal{B}(\mathbf{x}) = \{t\mathbf{x} : t > 0\}$$

The *geometric mean* plays a useful role in compositional data analysis.

Definition 5 (Geometric mean) *The* geometric mean *of a* $D \times 1$ *vector* $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ *of positive components is*

$$\mathrm{g_m}(\mathbf{x}) = (x_1 \ldots x_D)^{1/D}.$$

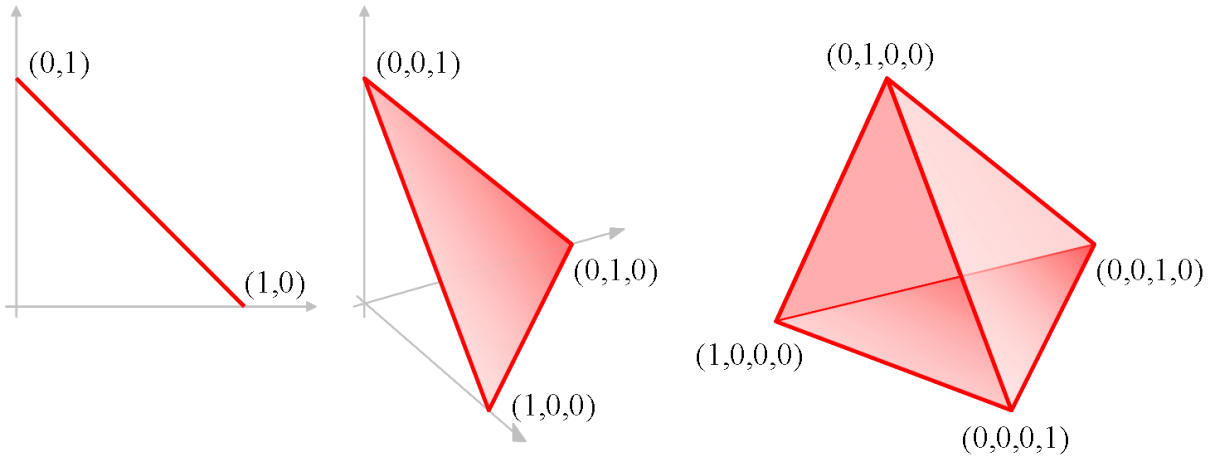

Figure 2.1: Geometric depictions of low-dimensional compositions. The left diagram shows how any two-part composition must lie on the red line corresponding to $x_1 + x_2 = 1$ with extremes (or *vertices*) $(1, 0)$ and $(0, 1)$. The middle diagram shows how any three-part composition must lie on the red triangle corresponding to $x_1 + x_2 + x_3 = 1$. The right diagram shows how any four-part composition must lie within the red tetrahedron corresponding to $x_1 + x_2 + x_3 + x_4 = 1$. Each of these figures is a *simplex* (a 1-, 2-, and 3-simplex, respectively), and because the composition is constrained to sum to 1, each figure represents a *standard* (or *unit*) simplex. (Higher dimensional simplices are hard to visualise, but Wikipedia hosts a mind-bending animation of a 5-simplex or *pentachoron* performing a double rotation about two orthogonal planes.) Note that the straight-line distance between any two compositions is limited, and that that limit is reached when the compositions are at two different vertices.

Let's relate these definition to some data: the first column of Figure 1.3. Here the basis $\mathbf{w}$ is in $D = 6$ dimensions, formed by the counts of molecules of different species of mRNA as seen by the Omics Imp in 1ms:

		a		b		c		d		e		z	
basis:	$\mathbf{w} = ($	1001		809		488		352		211		100	$)$
size:	$t =$	1001	$+$	809	$+$	488	$+$	352	$+$	211	$+$	100	$= 2961$
composition:	$\mathbf{x} = ($	$\frac{1001}{2961}$		$\frac{809}{2961}$		$\frac{488}{2961}$		$\frac{352}{2961}$		$\frac{211}{2961}$		$\frac{100}{2961}$	$)$
	$=$	(0.340		0.270		0.160		0.120		0.071		0.034)	
geometric mean:	$g_m =$	(0.340	$\times$	0.270	$\times$	0.160	$\times$	0.120	$\times$	0.071	$\times$	$0.034)^{1/6}$	$= 0.128$

With these definitions in place, we are ready to embark on our exploration of the impacts of closure.

3 Univariate impact of closure

This section is about how the statistics of single components (*i.e.*, *univariate* statistics) are affected by all components having to sum to a constant (*i.e.*, *closure*). One example of univariate statistics would be using a t-test to assess whether the values of a component under one treatment differed more than could be explained by chance alone from the values of that component under another treatment.

Consider a two-part composition $\mathbf{x}$ derived from a basis $\mathbf{w}$ as follows (see Figure 3.1):

$$\begin{aligned}
\mathbf{w} &= (w_1, w_2) \\
\mathbf{x} &= \mathcal{C}(\mathbf{w}) \\
&= \left(\frac{w_1}{w_1 + w_2}, \frac{w_2}{w_1 + w_2} \right)
\end{aligned}$$

If we fixed w_1, what would we have to do to w_2 to double x_2?

$$\begin{aligned}
\mathbf{w}' &= (w_1, w_2') \\
\mathbf{x}' &= (x_1', x_2') \\
&= (x_1', 2x_2) \\
\therefore \frac{w_2'}{w_1 + w_2'} &= 2 \cdot \frac{w_2}{w_1 + w_2} \\
w_2' &= 2w_2 \cdot \frac{w_1}{w_1 - w_2} \\
&\approx 2w_2 \qquad \text{if} \qquad w_1 \gg w_2
\end{aligned}$$

Conversely, If we fixed w_2, what would we have to do to w_1 to double x_2?

$$\begin{aligned}
\mathbf{w}' &= (w_1', w_2) \\
\mathbf{x}' &= (x_1', x_2') \\
&= (x_1', 2x_2) \\
\therefore \frac{w_2'}{w_1' + w_2} &= 2 \cdot \frac{w_2}{w_1 + w_2} \\
w_1' &= \frac{1}{2}(w_1 - w_2) \\
&\approx \frac{1}{2} w_1 \qquad \text{if} \qquad w_1 \gg w_2
\end{aligned}$$

Imagine that we are not working with a two-part composition but, instead, one that has many parts. Imagine also that x_1 is the aggregation of all components *except* x_2. The idea of this thought exercise is to help us understand the impact of closure on a single component in a composition of many parts. We can say that "$x_2 \mapsto kx_2$ implies $w_2 \mapsto kw_2$" provided

- x_2 is a relatively small component
- the rest of the basis stays the same
- $\log k$ is "small."

This means that we are reasonably safe to do univariate statistics on components of very large compositions that are not changing dramatically.

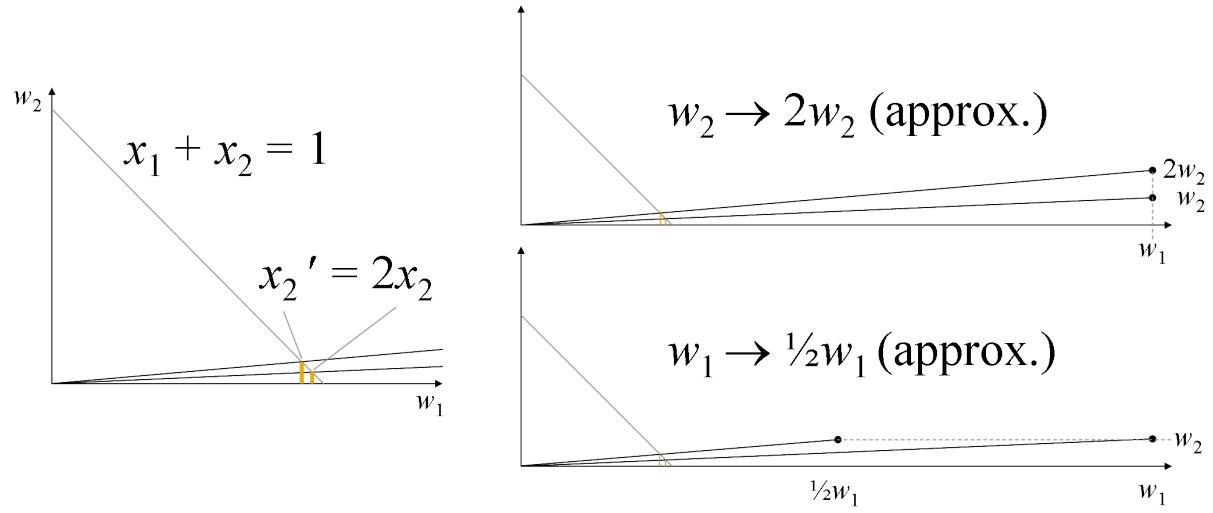

Figure 3.1: Did a two-fold change in x_2 occur because w_2 doubled? Or because w_1 halved? The left diagram shows two compositions lying on the simplex corresponding to $x_1 + x_2 = 1$ at the tops of the tiny orange bars. The taller bar has twice the amount of x_2 than the shorter. The two right hand diagrams zoom out to show two different basis vectors that could give rise to these compositions: the upper diagram uses twice the amount of w_2 to give a two-fold change in x_2; the lower diagram uses *half* the amount of w_1 to the same end. We cannot tell from the compositional data alone (the left diagram) what happened to give rise to the two-fold change in x_2. Many transcriptomic analyses assume (often implicitly) that the vast majority of mRNAs vary little in expression across different treatments, so that a two-fold change observed in a composition corresponds to a two-fold change in the basis.

In practice, this is exactly the situation with "spike-in" experiments, where a known concentration of a readily identifiable molecule (or cocktail of molecules) is added to samples of a mixture containing unknown concentrations of molecules under study, usually to assess or compare the sensitivity of different microarray technologies [9] or as a means to infer sample mRNA concentrations [10].

We hazard a guess that samples in most high-throughput gene expression experiments are much more variable than in carefully controlled spike-in evaluations. Are the amounts of mRNA produced by the cells under study similar enough in size and composition not to confound univariate statistical analysis? We do not know. But we suspect that the mRNA products of cells from different tissues (*e.g.*, brain/liver, cancer/non-cancer) or tissues at different stages of life (e.g., dormant/germinating) are quite different in both size and composition. To us, this calls into question the whole paradigm of testing for "significant differential expression" using only measures of relative abundance.

4 Impact of closure on multivariate distance metrics

Having looked briefly at the potential impact of closure on univariate statistics in large compositions, we now look at an important multivariate concept under closure: *distance*.

Distance refers to the relationship between points in a space. The space may have 1, 2, 3, or any number of dimensions. (Points defined in one dimension are called "univariate"; points defined in more than one dimension are called "multivariate".)

Bioscience is replete with multivariate data sets, including microarray data in which each sample is represented by a point in a space of as many dimensions as there are spots on the array.

Multivariate distance metrics underpin clustering methods (*e.g.*, hierarchical clustering) by telling us how "close" multivariate points are to each other.

This section explores how different approaches (a.k.a. *metrics*) to measuring distance between multivariate points are affected when the dimensions of those points have to sum to a constant (*i.e.*, *closure*).

Vêncio *et al.* [11] propose Aitchison's (compositional) distance $d_{\mathsf{A}}(\cdot,\cdot)$ as an alternative to Euclidean distance $d_{\mathsf{E}}(\cdot,\cdot)$ in clustering digital gene expression data where

$$d_{\mathsf{A}}(\mathbf{x},\mathbf{X}) \triangleq \sqrt{\frac{1}{D}\sum_{i<j}\left(\log\frac{x_i}{x_j}-\log\frac{X_i}{X_j}\right)^2} \tag{4.1}$$

$$= \sqrt{\frac{1}{2D}\sum_{i}\sum_{j}\left(\log\frac{x_i}{x_j}-\log\frac{X_i}{X_j}\right)^2} \tag{4.2}$$

$$d_{\mathsf{E}}(\mathbf{x},\mathbf{X}) \triangleq \sqrt{\sum_i (x_i-X_i)^2}. \tag{4.3}$$

Vêncio *et al.* demonstrated that d_{A} clustered simulated RNA-seq data in essentially the same way as a Euclidean clustering of data from the corresponding microarray experiment.

They also showed that Aitchison's distance clustered simulated RNA-seq data more interpretably than Euclidean clustering of that data. In this section, we look for quantitative reasons as to why this was the case, and investigate an approach that is commonly used in analysing RNA-seq data (though not always with clear rationale): working with the *logarithm* of the counts.

We will look at these questions using two basis vectors, $\mathbf{w}$ and $\mathbf{W}$, where

$$\begin{aligned} W_i &= K_i \cdot w_i, \qquad K_i > 0 \\ &= \mathrm{g_m}(\mathbf{K}) \cdot k_i \cdot w_i \end{aligned}$$

where $\mathrm{g_m}(\mathbf{K})$ is the *geometric mean* of the elements of $\mathbf{K}$. This means that $\mathrm{g_m}(\mathbf{k}) = 1$ or, alternately

$$\sum_i \log k_i = 0. \tag{4.4}$$

$\mathbf{x}$ and $\mathbf{X}$ are the corresponding closures of $\mathbf{w}$ and $\mathbf{W}$:

$$\begin{aligned} x_i &= \frac{w_i}{\sum_j w_j} \\ &= w_i/\mathbf{w}\cdot\mathbf{j} \\ &= x_i/\mathbf{x}\cdot\mathbf{j} \end{aligned} \tag{4.5}$$

$$\begin{aligned} X_i &= \frac{K_i \cdot w_i}{\sum_j K_j \cdot w_j} \\ &= \frac{\mathrm{g_m}(\mathbf{K}) \cdot k_i \cdot w_i}{\sum_j \mathrm{g_m}(\mathbf{K}) \cdot k_j \cdot w_j} \\ &= k_i w_i/\mathbf{w}\cdot\mathbf{k} \\ &= k_i x_i/\mathbf{x}\cdot\mathbf{k}. \end{aligned} \tag{4.6}$$

Aitchison refers to this kind of operation as a *perturbation* of $\mathbf{x}$, symbolised as $\mathbf{x} \oplus \mathbf{k}$ by Pawlowsky-Glahn *et al.* [2].

4.1 Aitchison's distance between compositions

For clarity, we will work with squared distances. Let us begin by simplifying $d^2_{\mathsf{A}}(\mathbf{x}, \mathbf{X})$:

$$\begin{aligned} d^2_{\mathsf{A}}(\mathbf{x}, \mathbf{X}) &= \frac{1}{D} \sum_{i<j} \left(\log \frac{x_i}{x_j} - \log \frac{k_i x_i}{k_j x_j} \right)^2 \\ &= \frac{1}{D} \sum_{i<j} \left(\log \frac{x_i}{x_j} - \log \frac{k_i}{k_j} - \log \frac{x_i}{x_j} \right)^2 \\ &= \frac{1}{D} \sum_{i<j} \left(\log \frac{k_i}{k_j} \right)^2 . \end{aligned} \tag{4.7}$$

Aitchison's distance tells us only about the perturbation that has occurred. This is not the case with $d^2_{\mathsf{E}}(\mathbf{x}, \mathbf{X})$ as we shall show shortly but, before that, let us show how Equation 4.7 can be simplified further because the perturbation vector $\mathbf{k}$ satisfies Equation 4.4.

We can rearrange the summation

$$\begin{aligned} d^2_{\mathsf{A}}(\mathbf{x}, \mathbf{X}) &= \frac{1}{D} \sum_{i<j} \left(\log \frac{k_i}{k_j} \right)^2 \\ &= \frac{1}{2D} \sum_{i} \sum_{j} \left(\log \frac{k_i}{k_j} \right)^2 \\ &= \frac{1}{2D} \sum_{i} \sum_{j} \log^2 k_i + \log^2 k_j - 2 \log k_i \log k_j \\ &= \sum_{i} \log^2 k_i \end{aligned} \tag{4.8}$$

since $\sum_i \log k_i = 0$.

Aitchison's distance metric tells us only about the *relative* differences (*i.e.*, *ratios*) of the corresponding components in compositions.

4.2 Euclidean distance between compositions

Euclidean distance is familiar to us in 1, 2 or 3 dimensions as the length of the straight line between two points.

Now, let us look at the Euclidean distance between $\mathbf{x}$ and $\mathbf{X}$.

$$\begin{aligned} d^2_{\mathsf{E}}(\mathbf{x}, \mathbf{X}) &= \sum_i (x_i - X_i)^2 \\ &= \sum_i \left(\frac{w_i}{\mathbf{w} \cdot \mathbf{j}} - \frac{k_i \cdot w_i}{\mathbf{w} \cdot \mathbf{k}} \right)^2 . \end{aligned} \tag{4.9}$$

This is not going to lend itself to simplification without some assumptions about $\mathbf{k}$, but we note that d_{E} is bounded by $\sqrt{2}$, the longest Euclidean distance on a D-dimensional simplex.

If we assume that $\mathbf{k}$ is such that $\mathbf{w}$ and $\mathbf{W}$ each sum to the same amount we can now write

$$\begin{aligned} d^2_{\mathsf{E}}(\mathbf{x}, \mathbf{X}) &= \sum_i \left(\frac{w_i - k_i \cdot w_i}{\mathbf{w} \cdot \mathbf{j}} \right)^2 \\ &= \sum_i (x_i - k_i \cdot x_i)^2 \\ &= \sum_i (1 - k_i)^2 x_i^2 \end{aligned} \tag{4.10}$$

which is really not much of an improvement over Equation 4.9. The main points to draw from this are that

1. d_{E} is bounded by $\sqrt{2}$
2. d_{E} depends on both the perturbation by $\mathbf{k}$ and the composition $\mathbf{x}$ that was perturbed.

We note also that the maximum Euclidean distance from the barycentre of a D-dimensional simplex (*i.e.*, the point $(1/D, 1/D, \ldots, 1/D)$ is $\sqrt{(D-1)/D}$ which approaches 1 as D increases.

In summary, because compositions are constrained to lie on a D-dimensional simplex, the Euclidean distance to $\mathbf{x} \oplus \mathbf{k}$ depends on both $\mathbf{x}$ and $\mathbf{k}$ and is bounded above by $\sqrt{2}$, whereas Aitchison's distance depends on $\mathbf{k}$ alone and has no upper bound.

The Euclidean distance metric tells us the "straight-line" distance between compositions. Because compositions are constrained to lie on a simplex, the Euclidean distance between compositions is limited: the furthest distance two points can be apart on the line, triangle, or tetrahedron in Figure 2.1 is $\sqrt{2}$. Aitchison's distance depends only on the *ratios* of corresponding components and has no such upper limit.

Imagine very high dimensional compositions, *e.g.*, RNA-seq counts of thousands of different mRNA species. By necessity, each component will be a very small proportion of the total number of mRNA sequence reads. Consequently, the Euclidean distance between any two such compositions will be very small also, even though they may have components that are many-fold different in relative abundance. Aitchison's distance, with its focus on the ratio of corresponding components, will emphasise these differences in relative abundance much more effectively.

4.3 Euclidean distance between logged compositions

A common approach to analysing count data is to adopt a log-linear Poisson model [12]. It is also common in the biosciences to analyse and present strictly positive data using a logarithmic transformation, without necessarily referring to an underlying probabilistic model. "Logged data" is the commonplace in microarray analysis and now, RNA-seq data analysis [13, 14].

With this in mind, let us look at the Euclidean distance between $\log \mathbf{x}$ and $\log \mathbf{X}$.

$$\begin{aligned}
d_{\mathsf{E}}^2(\log \mathbf{x}, \log \mathbf{X}) &= \sum_i (\log x_i - \log X_i)^2 \\
&= \sum_i \left(\log x_i - \log \frac{k_i \cdot x_i}{\mathbf{x} \cdot \mathbf{k}} \right)^2 \\
&= \sum_i (\log \mathbf{k} \cdot \mathbf{x} - \log k_i)^2 \\
&= \sum_i \log^2 \mathbf{k} \cdot \mathbf{x} - 2 \log k_i \log \mathbf{k} \cdot \mathbf{x} + \log^2 k_i \\
&= D \log^2 \mathbf{k} \cdot \mathbf{x} + \sum_i \log^2 k_i \qquad (4.11) \\
&= D \log^2 \mathbf{k} \cdot \mathbf{x} + d_{\mathsf{A}}^2(\mathbf{x}, \mathbf{X}) \qquad (4.12) \\
&\geq d_{\mathsf{A}}^2(\mathbf{x}, \mathbf{X}).
\end{aligned}$$

So the Euclidean distance between logged compositions is closely related to Aitchison's distance but, like d_{E}, still depends on both the perturbation by $\mathbf{k}$ and the composition $\mathbf{x}$ that was initially perturbed. Furthermore, the Euclidean distance between logged compositions also depends explicitly on D, the dimensionality of the composition.

The Euclidean distance between log-transformed compositions is equal to Aitchison's distance—a function of the ratios of composition components only—*plus* an amount that depends on the absolute values of the compositions and their dimensionality.

Because the values of the log-transformed components can take on any real value from $-\infty$ to 0, log-transformed compositions are no longer constrained to lie on unit simplices (such as those in Figure 2.1). This means that, unlike the Euclidean distance between untransformed compositions in Section 4.2, the "straight-line" distance between *log*-transformed compositions has no upper limit.

Microarray fluorescence data are conventionally dealt with on a $\log_2$ scale. So even though compositional constraints may be at play in these data (because they are derived from fixed weights of total RNA), distance-based analyses (*e.g.*, the familiar hierarchically-clustered heatmap) are likely to be quite similar to what we would get with a purely compositional approach—more of which in Section 4.6

4.4 Kullback-Leibler divergence between compositions

We noted earlier that the Kullback-Leibler (K-L) divergence [15], though not a distance metric, is used often in information-retrieval (as well as its home turf: probability and information theory) as a means to compare unit-sum-constrained data. The term "K-L divergence" covers two forms of expression: *directed* and *symmetric* divergence. The Kullback-Liebler *directed* divergence from probability distribution P to probability distribution Q on some discrete random variable, indexed by i, is

$$D_{\mathsf{KL}}(P|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

whereas the *symmetric* divergence is

$$D_{\mathsf{KL}}(P,Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} + Q(i) \log \frac{Q(i)}{P(i)}.$$

Switching to the notation set out at the start of in Section 4, we can write the symmetric K-L divergence between unit-sum constrained compositions $\mathbf{x}$ and $\mathbf{X}$ as

$$\begin{aligned} D_{\mathsf{KL}}(\mathbf{x},\mathbf{X}) &= \sum_i x_i \log \frac{x_i}{X_i} + X_i \log \frac{X_i}{x_i} \\ &= \sum_i (x_i - X_i) \log \frac{x_i}{X_i} \\ &= \sum_i (x_i - X_i) \log \frac{\mathbf{k}\cdot\mathbf{x}}{k_i} \\ &= \sum_i (x_i - X_i)(\log \mathbf{k}\cdot\mathbf{x} - \log k_i) \\ &= D \log \mathbf{k}\cdot\mathbf{x} \sum_i (x_i - X_i) + \sum_i (X_i - x_i) \log k_i \end{aligned} \tag{4.13}$$

which, at a pinch, bears some similarity to the terms of Equation 4.11. Unlike Equation 4.12, there is no way to partition symmetric K-L divergence into a term that depends solely on the perturbation by $\mathbf{k}$, and a term that depends also on the composition $\mathbf{x}$ initially perturbed.

Kullback-Leibler divergence is often used to measure the difference between two probability distributions. One property of (discrete) probability distributions is that they sum to 1, just as compositions do. So, even though K-L divergence is not a distance metric (because it does not satisfy the *triangle inequality*), it is still worth entertaining as a means to compare compositions. (For examples of its use in microarray analysis see [16]).

The K-L divergence between compositions depends on the ratios of composition components, their absolute values of components, and the dimensionality of the compositions.

4.5 Four distance measures on some two-part compositions

To provide additional insight into the behaviour of Euclidean, Euclidean-log, and Aitchison's distances and Kullback-Leibler divergence, we apply these measures to the simplest possible compositions: those with only two parts.

We start with two basis vectors $\mathbf{w} = (w_1, w_2)$ and $\mathbf{W} = (W_1, W_2)$, and compare distances between their corresponding compositions. We fix $\mathbf{W}$ and w_1 while sweeping through a range of w_2 values, *e.g.*,

$\mathbf{w}$		$\mathbf{W}$	
w_1	w_2	W_1	W_2
1024	1	1024	1
1024	2	1024	1
1024	4	1024	1
⋮	⋮	⋮	⋮
1024	1024	1024	1
1024	2048	1024	1
⋮	⋮	⋮	⋮

We then observe the distances between $\mathbf{x} = \mathcal{C}(\mathbf{w})$ and $\mathbf{X} = \mathcal{C}(\mathbf{W})$ as a function of w_2. Figure 4.1 shows $d_{\mathsf{A}}(\mathbf{x}, \mathbf{X})$, $d_{\mathsf{E}}(\mathbf{x}, \mathbf{X})$, and $d_{\mathsf{E}}(\log \mathbf{x}, \log \mathbf{X})$.

As we are working with only two components, we can go a little further in understanding the difference between $d_{\mathsf{E}}(\log \mathbf{x}, \log \mathbf{X})$ and $d_{\mathsf{A}}(\mathbf{x}, \mathbf{X})$, i.e., $D \log^2 \mathbf{k} \cdot \mathbf{x}$ (Equation 4.12). This will be zero when $\mathbf{k} \cdot \mathbf{x} = 1$.

Since $\mathbf{x}$ is a two-part composition, $x_2 = 1 - x_1$. Also, because $\mathbf{k}$ satisfies Equation 4.4, we know that $k_2 = 1/k_1$. This allows us to write

$$\begin{aligned} \mathbf{k} \cdot \mathbf{x} &= k_1 x_1 + k_2 x_2 && (4.14) \\ &= k_1 x_1 + (1 - x_1)/k_1 && (4.15) \end{aligned}$$

which is equal to 1 in two situations: when $k_1 = 1$ or when $k_1 = (1 - x_1)/x_1$. In the left panel of Figure 4.1 we see these situations at $\mathbf{x} = \mathcal{C}(1024, 1)$ and $\mathbf{x} = \mathcal{C}(1024, 1024^2)$.

This section graphically demonstrates the behaviour of the four distance measures discussed in the preceding sections. Each panel of Figure 4.1 shows the distances between a fixed composition, and other possible two-part compositions.

Aitchison's distance (in red) depends only on the (log) ratio of corresponding components. Euclidean distance between log-components (in blue) has an additional positive amount that depends on the absolute values of components.

However the main point to observe is that Euclidean distance on raw compositional data (in green) has a limited range. *Euclidean distance does not reflect relative changes in components very well.*

Kullback-Liebler divergence (in purple) does not exhibit the pathologies of Euclidean distance on raw compositional data, nor does it show a clear relationship to the other distances.

Here is another exposition the distance metrics using more commonplace terms. Consider a

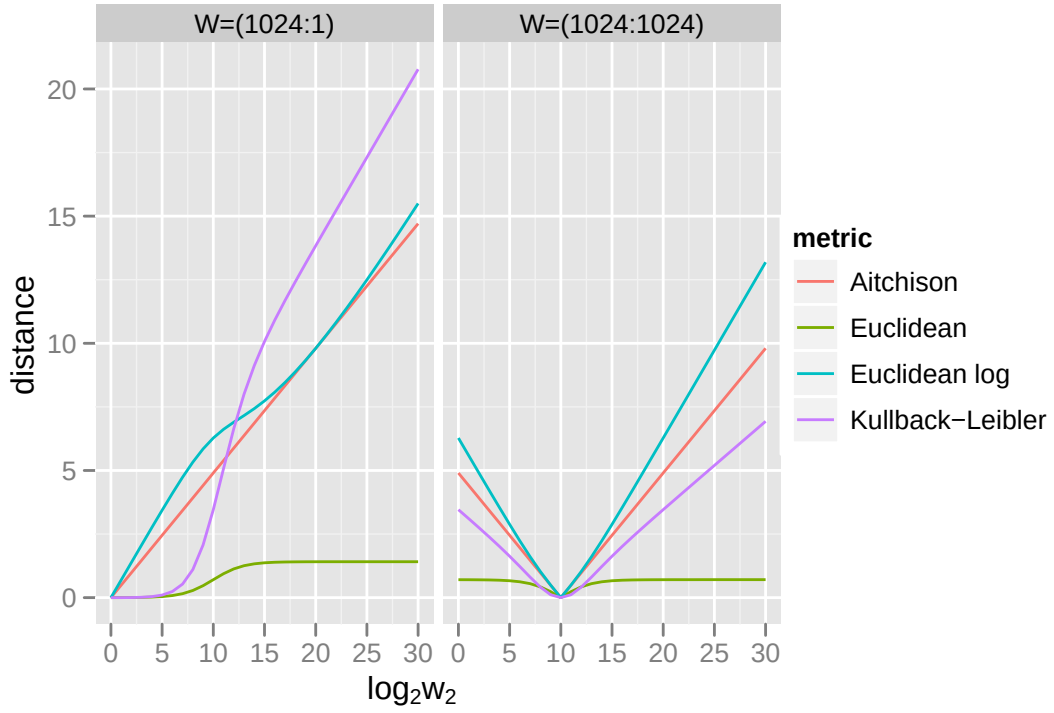

Figure 4.1: Plots of four different distance measures between composition $\mathcal{C}(1024, w_2)$ and $\mathcal{C}(\mathbf{W})$ for $\mathbf{W} = (1024, 1)$ in the left panel, and $\mathbf{W} = (1024, 1024)$ in the right. Note that the Euclidean distance to $\mathcal{C}(1024, 1)$ approaches $\sqrt{2}$, the length of the edge of a D-dimensional simplex. The Euclidean distance to $\mathcal{C}(1024, 1024)$, the barycentre of the simplex, approaches $\sqrt{(D-1)/D} = 1/\sqrt{2}$.

two-part mixture made up of ingredients a and b. Using the everyday notation a:b to denote how many parts of a we mix with parts of b let's look at the three different distances when we *double* the parts of a:

a:b ⟶	a:b	Aitchison	Euclidean log	Kullback-Leibler	Euclidean
1:10	2:10	0.4901	0.6123	2.0880	0.1071
1:100	2:100	0.4901	0.6834	0.2496	0.01373
1:1000	2:1000	0.4901	0.6921	0.1384	0.001410
1:10000	2:10000	0.4901	0.6930	0.1333	0.0001414
1:100000	2:100000	0.4901	0.6931	0.1326	0.00001414
1:1000000	2:1000000	0.4901	0.6931	0.1325	0.000001414

Aitchison's distance clearly depends only on the relative abundance of corresponding ingredients.

4.6 Four distance measures on some high-dimensional compositions

While we cannot think of a way to systematically visualise the behaviour of different distance measures on higher dimensional compositions, we can look at a particular data set that has already been used to exemplify differences between d_{A} and d_{E}.

Vêncio *et al.* [11] proposed d_{A} as an alternative to d_{E} in clustering digital gene expression data. They showed that Aitchison's distance clustered simulated RNA-seq data[4] in essentially the same way as a Euclidean clustering of data from the corresponding microarray experiment.

They also showed that Aitchison's distance clustered simulated RNA-seq data *more interpretably* than Euclidean clustering of that same count data.

Figure 4.2 (a) and (b) replicate the results obtained by Vêncio *et al.*, demonstrating that the Euclidean distance between vectors of RNA counts (a) fails to separate the experimental groups as neatly as Aitchison's distance and the Euclidean distance between the *logarithm* of the RNA counts (b). The Euclidean distance between the *logarithm* of the RNA counts gave essentially identical results to those using Aitchison's distance. (This was the observation that led us to look more closely at the relationship between d_{A} and d_{E} on log-transformed data (Equation 4.12).)

Like Euclidean distance, symmetric Kullback-Liebler divergence (Figure 4.2 (c)) results in a clustering that fails to separate the experimental groups.

As we mentioned in Section 4.3, we would expect a distance-based analysis of microarray fluorescence data on a $\log_2$ scale to be quite similar to a purely compositional approach. Vêncio *et al.* [11] advocate Aitchison's distance as a metric for RNA-seq and other forms of enumeration-based gene expression data on the grounds that they are compositional data. As we shall see in the next section, we think issues to do with correlation and covariance provide even stronger empirical reasons for using compositional methods.

[4] Counts of the number of reads of a given mRNA were drawn from a Poisson distribution whose rate parameter was equal to the total number of reads times the relative fluorescence of the oligonucleotide probe for that mRNA, as observed in an actual microarray experiment.

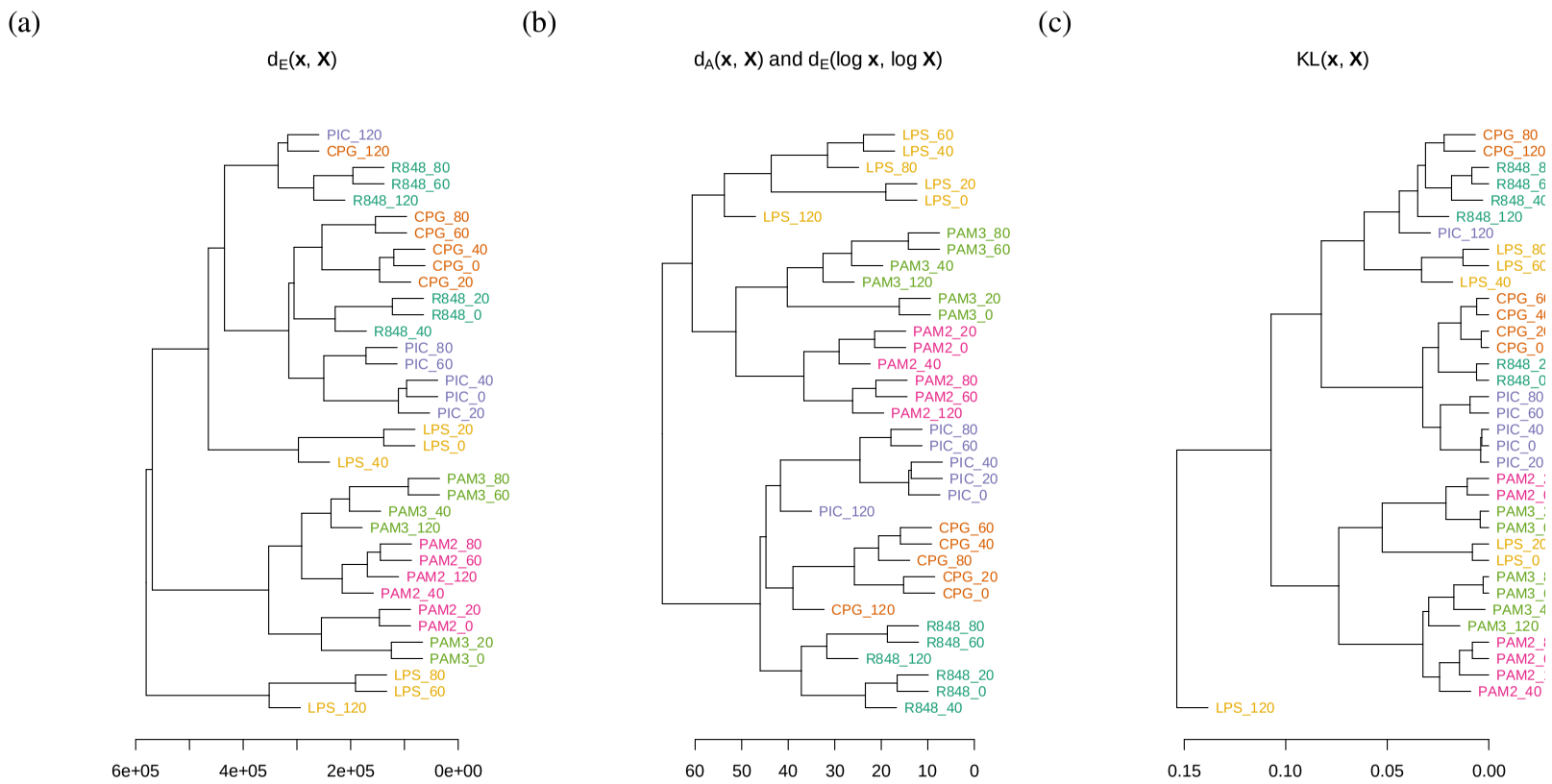

Figure 4.2: Hierarchical clustering of the simulated RNA-seq data from Vêncio *et al.* [11] using four different distance measures: (a) Euclidean (b) Aitchison's and Euclidean of log-transformed data (results were identical). Each group of subjects is coloured differently. The main points to note are that tree (b) clusters subjects into their actual groups whereas trees (a) and (c) produce more mixed results.

5 Impact of closure on correlation and covariance

Compositional constraints are notorious for their impacts on on the covariance and correlation structures of data [1, Section 3.3]. A major motivation for the investigations described in this technical report is to understand better how compositional constraints in omics data might affect our estimates of covariance and, in doing so, lead our inferences astray. We have two particular omics scenarios in mind:

Transcriptomics: this refers to the measurement and analysis of the complement of mRNA transcripts present in a given tissue at a given time. This is one of the most popular areas to infer "networks of associations" between different RNA species often, but not always on the basis of correlations observed between mRNA species across different experimental conditions. Typically, the number of mRNA species is in the thousands to hundreds of thousands, though there can be experiments that investigate the abundance of a much smaller number of classes of mRNA such as small RNAs classified by their length ($\leq 20, 20, 21, \ldots, \geq 30$ bases).

Metagenomic profiling: this refers to the measurement and analysis of the complement of (usually) microbial species present in an environmental sample. The 16S region of ribosomal RNA is often used as an indicator of microbial "species", so one approach to profiling microbial community structure is to tally the different 16S sequences observed[5]. Microbial community function is about the kinds of genes that are being expressed across a community at a given time. Function is inferred from the different kinds of mRNA observed within the sample. The numbers of microbial "species" within environmental samples (the community complexity) ranges from fewer than ten (in environments with strong selection pressure) to many thousands[6].

The number of different functions observed within environmental samples depends on the resolution of functional descriptors being used. For example, one broad functional class might be "nitrogen-cycling" which can then be broken down into more specific aspects of that function. Function can be described using anywhere from around ten to over one hundred classes.

A similar approach can be applied in comparative genomics, using *gene ontology* (GO) terms as a hierarchical labeling system for the function of an organism's genes. At this point in time, the function of many genes is uncertain or unknown and we are mindful of the impact that this may have in comparing organisms that differ widely in that respect. (Imagine if component z in Figure 1.5 corresponded to the count of "function unknown" genes.)

Current popular methods of analysing multivariate transcriptomic or metagenomic structure/function data include

Visual inspection of pie charts or stacked percentage charts

Rarefaction/accumulation analysis which relates the number of species observed to the number of individuals or samples measured [17]

[5] We note that 16S rRNA gene studies to determine the species composition of a sample are traditionally considered to be separate from metagenomics (in which shotgun sequencing is used to look at all genes). However, this distinction is becoming blurred over time.

[6] ... acknowledging debates as to the extent to which some of this apparent complexity can be attributed to artifacts of the measurement process (especially sequencing errors).

Clustering, often two-way hierarchical clustering with some kind of "heat map"

Network inference, through correlation-based, sparse regression or Bayesian network methods

Lower-dimensional projections, including principal component analysis (PCA) and linear discriminant analysis (LDA).

With regard to compositional data, we will leave aside visual inspection methods as they are more about qualitative rather than quantitative insights. Rarefaction/accumulation analysis focuses on estimating the number of species rather than the relative abundance of species within samples, so we will leave that aside as well. We have already explored clustering by exploring distance metrics that can be sensibly applied to compositions. In the following sections, we focus on the covariance and correlation aspects of compositional data analysis as applied to the omics.

5.1 A three-part approach to understanding closure: algebra

To help us understand the impact of constraints on correlation and covariance where there are large numbers of components D, we will work in terms of three parts: components 1 and 2 are measurements whose pairwise correlation is of interest, and component 3 represents "the rest", *i.e.*, the other $D-2$ measurements aggregated together.

We'll begin with *covariance*. Let's consider how $\text{cov}(\log x_1, \log x_2)$ — the covariance between the two components of interest in the composition — relates to $\text{cov}(\log w_1, \log w_2)$ — the covariance between the two components of interest in the composition's *basis*. To do this, we will use the result that

$$\text{cov}(A+C, B+C) \;=\; \text{cov}(A,B) + \text{cov}(A,C) + \text{cov}(B,C) + \text{var}(C). \tag{5.1}$$

As we are working with logarithms, the covariance between the two components of interest in the composition can be written

$$\begin{aligned} &\text{cov}(\log x_1, \log x_2) \\ &= \text{cov}\left(\log(w_1/t), \log(w_2/t)\right) \\ &= \text{cov}(\log w_1 - \log t, \log w_2 - \log t) \\ &= \text{cov}(\log w_1, \log w_2) - \text{cov}(\log w_1, \log t) - \text{cov}(\log w_2, \log t) + \text{var}(\log t) \end{aligned} \tag{5.2}$$

The covariance between the log of the two components of interest in a composition is equal to

- the covariance between the log of those components in the basis
- plus or minus some terms related to variation in the *size* of the basis.

If these latter terms become large, the covariances observed in the composition could be very different to what is happening in the basis.

This highlights the effect of variation in the *size* t of the D-part basis and we note that

> "...when interest is in the relationship between the compositional and size aspects of the basis it is more appropriate to consider the basis covariance structure defined in terms of the pattern of variability in $\mathbf{x}$ and t. This is possible since $\mathbf{x}$ and t together determine $\mathbf{w}$ as $t\mathbf{x}$." [1, Section 9.2 – Covariance relationships]

Equation 5.2 shows explicitly that $\text{cov}(\log x_1, \log x_2) = \text{cov}(\log w_1, \log w_2)$ when t is constant, in other words, *when* $\mathbf{w}$ *is already a composition* (but one constrained to sum to t rather than 1 as $\mathbf{x}$ is).

Turning now to *correlation* between the (log) components of interest, by definition we have that

$$\text{corr}(A+C, B+C) \;=\; \frac{\text{cov}(A+C, B+C)}{\sqrt{\text{var}(A+C)}\sqrt{\text{var}(B+C)}} \tag{5.3}$$

and

$$\text{var}(A+C) \;=\; \text{var}(A) + \text{var}(C) - 2\text{cov}(A,C). \tag{5.4}$$

Unfortunately, substituting $\log x_1$ *etc.* into these equations does not lead to as neat an expression as Equation 5.2 and we will not present that expansion here. However, we note that $\text{corr}(\log x_1, \log x_2) = \text{corr}(\log w_1, \log w_2)$ when t is constant.

There is no straightforward expression to clearly describe the relationship between correlation observed in the composition, and correlation observed in the basis. However, knowing that correlation is a function of variance and covariance tells us that the correlations we see in compositional data could also be very different to those observed in the basis.

We found this algebraic understanding of the relationship between $\log \mathbf{x}$ and $\log \mathbf{w}$ useful, but we also felt a need to simulate and visualise possible relationships between bases and compositions to explore the extent to which $\text{cov}(\log x_1, \log x_2)$ could give us an accurate or misleading impression of what's going on in the underlying basis.

5.2 A three-part approach to understanding closure: simulation... take 2

Simulation has been used previously to explore compositional data (see [19, 20] for example) but, as far as we know, these explorations have not been conducted to investigate the properties of log-transformed compositions.

Simulation gives us complete control over the statistical properties of the data, at the expense of losing connection to real experimental data. Unfortunately, at this time, we know of no one who has actual experimental data from both bases and their corresponding compositions in a molecular biology setting.

When it came to simulating data from a three-part basis, we were confronted by having to assume a distribution for $\mathbf{w}$. Clearly, w_1, w_2, w_3 must all be positive. For simplicity, we decided to create a simulation where w_1, w_2 could have some straightforward statistical dependence while remaining statistically independent of w_3. So far, we think these assumptions are "one step along" from the simplest scenario of completely independent parts.

But what probability distribution should we adopt for $\mathbf{w}$? For the sake of exposition, we shall first present the approach that we actually took on our *second* attempt. (Later we shall describe what we learned from our first try.)

We think that a trivariate log-normal distribution for $\mathbf{w}$ is the simplest but most general way to create a three-part basis that can be used to explore the impact of closure on the pair-wise relationship between w_1, w_2. In this scenario

$$\begin{aligned} \mathbf{w} &\sim \Lambda_3(\mu, \mathbf{\Sigma}) \\ \log \mathbf{w} &\sim \mathcal{N}_3(\mu, \mathbf{\Sigma}) \end{aligned} \tag{5.5}$$

where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

and

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & 0 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}$$

Note also that a random variable w distributed according to

$$w \sim \Lambda(\mu, \sigma^2) \tag{5.6}$$

has mean $e^{\mu+\frac{1}{2}\sigma^2}$ and variance $e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$.

Initially, we tried to visualise and explore the properties of $\mathbf{w}$ and $\mathbf{x} = \mathcal{C}(\mathbf{w})$ by simulating and plotting $\log(\mathbf{w})$ and $\log(\mathbf{x})$ across a range of μ and $\mathbf{\Sigma}$ values, but with seven parameters ($\mu_{1,2,3}$, $\sigma_{1,2,3}$, ρ_{12}) this proved cumbersome and unrevealing.

What we wanted to understand was the impact that parameter changes had in different parts of this 7-dimensional space. This led us to develop interactive plotting software in Java, screen-shots of which can be seen in Figures 5.1 and 5.2, and of which further details are given in Appendix A.

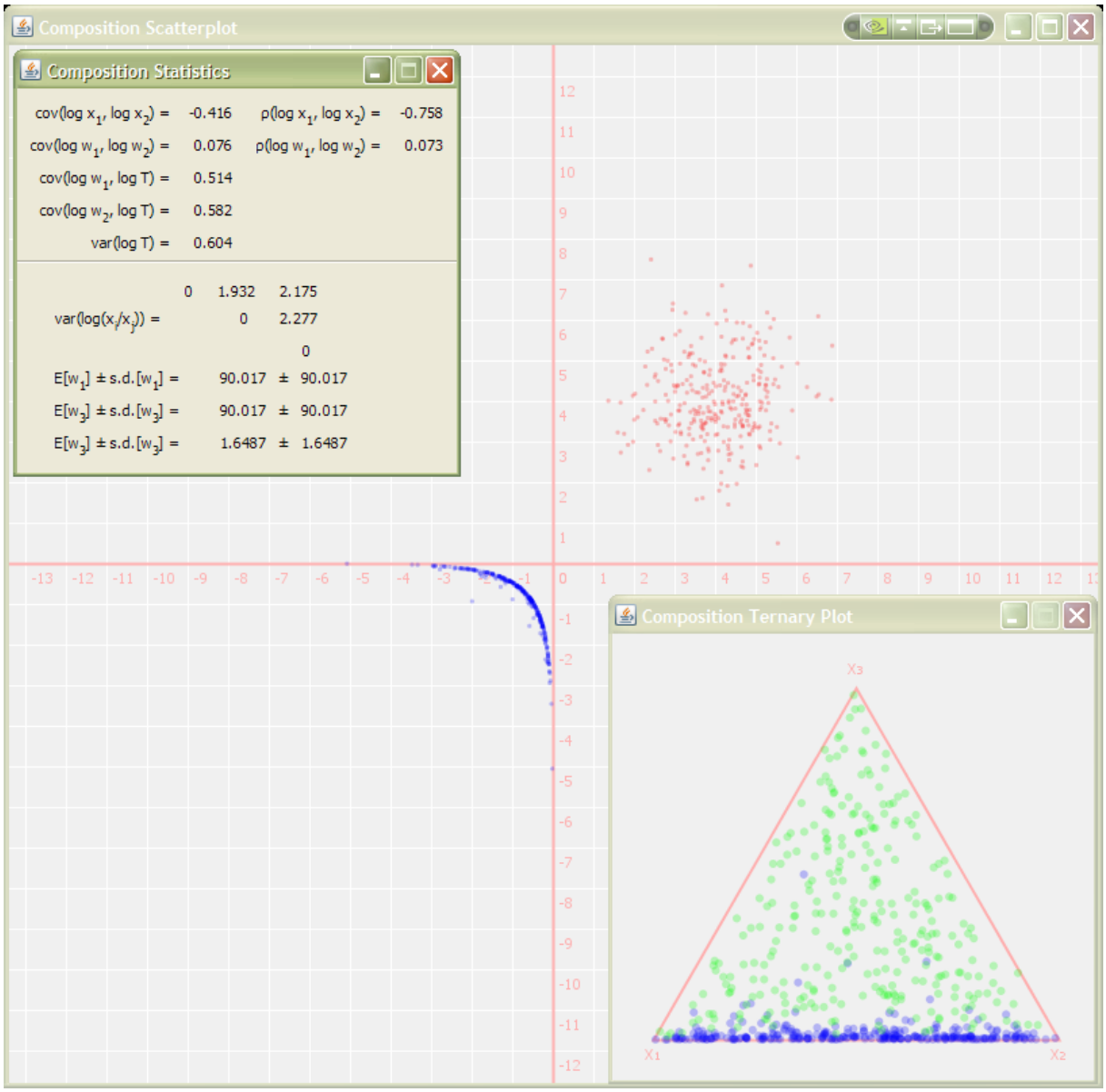

Figure 5.1: A screenshot of interactive simulation software (see Appendix A) showing 300 samples from a trivariate log-normal basis in red, and the corresponding composition in blue. Only components 1 and 2 are shown, and axes are drawn on a log scale. In this example, the basis variables w_1 and w_2 are greater than 1, so the red samples of $\log w_1$ and $\log w_2$ appear in the upper right quadrant. By definition, the positive components x_1 and x_2 are less than 1, and therefore the blue samples of $\log x_1$ and $\log x_2$ *always* appear in the lower left quadrant.
The upper left quadrant shows statistical summaries while the lower right displays a ternary plot of the data [18].
Here $\log \mu_{1,2,3} = (4, 4, 0)$ and $\log \sigma_{1,2,3} = (0, 0, 0)$. While the sample correlation between $\log w_1$ and $\log w_2$ is zero, the correlation observed between $\log x_1$ and $\log x_2$ is around -0.75.

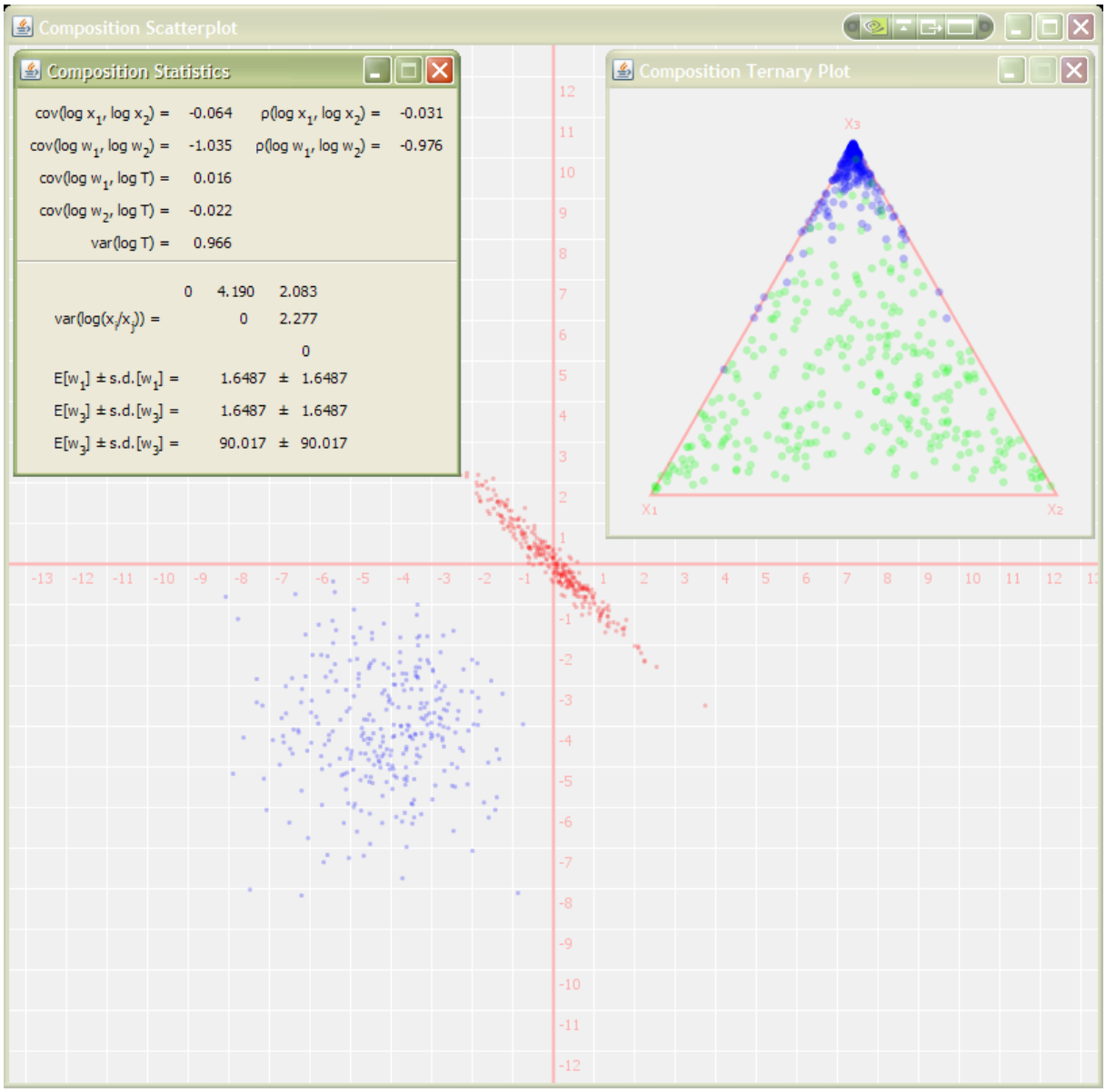

Figure 5.2: Another screenshot, this time of samples where $\log \mu_{1,2,3} = (0, 0, 4)$ and $\log \sigma_{1,2,3} = (0, 0, 0)$.While the sample correlation between $\log w_1$ and $\log w_2$ is nearly -0.98, the correlation observed between $\log x_1$ and $\log x_2$ is around 0.

We have used this software to find two extreme (but plausible) situations that characterise how the analysis of log-transformed compositional data could lead to incorrect inferences about the relationship between the components of interest, x_1 and x_2:

$w_1, w_2 \gg w_3$: When the basis is dominated by the components of interest, $\log x_1$ and $\log x_2$ tend to move towards their upper limits, *i.e.*, the boundary defined by $\log x_2 = \log(1-x_1)$ for $x_1 \in (0,1)$. This imposes a negative bias on the corr$(\log x_1, \log x_2)$ in comparison to corr$(\log w_1, \log w_2)$.

Try exploring this by sweeping ρ_{12} through its range, with $\log \mu_{1,2,3} = (4,4,0)$ and $\log \sigma_{1,2,3} = (0,0,0)$ (see Figure 5.1).

This situation is easy to detect in compositional data because $x_1 + x_2$ will be close to 1. However, nothing can be done to infer the relationship between the basis variables w_1 and w_2 using the compositional data alone.

$w_3 \gg w_1, w_2$: When the basis is not dominated by the components of interest, the degree of correspondence between corr$(\log x_1, \log x_2)$ and corr$(\log w_1, \log w_2)$ depends on the variance of $\log w_3$. As σ_3 increases, corr$(\log x_1, \log x_2)$ tends to be positively biased in comparison to corr$(\log w_1, \log w_2)$.

Try exploring this by sweeping $\log \sigma_3$ through its range, with $\log \mu = (0,0,4)$, $\log \sigma_{1,2} = (0,0)$ and $\rho_{12} = -0.98$ (see Figure 5.2).

While it is again easy to detect this situation in compositional data (this time $x_1 + x_2$ will be close to 0), there is nothing in that data to tell us about the variance of $\log w_3$.

In summary, the correlation between log-transformed components of interest in the composition will be approximately the same as that of their counterparts in the basis only when $w_3 \gg w_1, w_2$ and var$(\log w_3)$ is small — in other words, when w_1 and w_2 are small parts of a relatively constant total — *but we are unable to tell when that is the case using the compositional data alone*.

This leads us to share what we learned on our first attempt at simulating a three-part composition.

5.3 A three-part approach to understanding closure: simulation. . . take 1

This section highlights that having complete control over the statistical properties of data (through simulation) carries with it the burden of deciding what those properties should be. Our initial attempt at using simulation to explore the impact of closure led us to believe there was no problem as long as one worked with log-transformed data. However, our choice of data distribution masked the potential for naïve analysis of compositional data to be misleading.

Our first approach to simulating a three-part composition was partly inspired by some of the distributions used by Vêncio *et al.* [11] in simulating digital gene expression data. They generated tag counts using Poisson distributions, but we also needed a way to create a statistical dependence between the two components of interest. To do this, we chose

$$\begin{aligned} \mathbf{w} &\sim \Lambda_2(\mu, \mathbf{\Sigma}) \\ \log \mathbf{w} &\sim \mathcal{N}_2(\mu, \mathbf{\Sigma}) \end{aligned}$$

where

$$\mu = \begin{pmatrix} \log m \\ \log m \end{pmatrix}$$

and

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

For the third component representing "the rest" of our measurements we chose

$$w_3 \sim \text{Poisson}(m \cdot (D-2))$$

to correspond to the size of the rest of our $D-2$ other measurements.

We then set about simulating the impact of increasing the apparent dimension of the composition on $\text{corr}(\log x_1, \log x_2)$. As Figure 5.3 shows, the larger D, the better this correlation approximates $\text{corr}(\log w_1, \log w_2)$, an apparent victory for those who wish to work with log-transformed counts of digital gene expression without fear of being misled.

Then, somehow, the penny dropped, probably because we started thinking about what was going on in this scenario in terms of Equation 5.2. As D grows, the size t of the basis becomes like a Poisson random variable with a larger and larger mean. As this happens, $\text{var}(\log t)$ — and, by necessity, $\text{cov}(\log w_1, \log t)$ and $\text{cov}(\log w_2, \log t)$ — become smaller and smaller[7].

In other words, as D grows in this scenario, we steadily approach the happy situation (discussed in the previous section) where w_1 and w_2 are small parts of a relatively constant total.

[7]The variance of the log of a Poisson distributed variable can be approximated using a Taylor series expansion. The variance of a smooth function $f(\cdot)$ of a random variable X is approximately

$$\text{var}(f(X)) \approx (f'(\mathbb{E}(X)))^2 \text{var}(X).$$

Using this approximation, and knowing that the variance of a Poisson distributed variable equals its expectation λ, the variance of the *logarithm* of a Poisson distributed variable X can be written

$$\begin{aligned} \text{var}(\log(X)) &\approx \frac{1}{\mathbb{E}^2(X)}\mathbb{E}(X) \\ &= \frac{1}{\lambda}. \end{aligned}$$

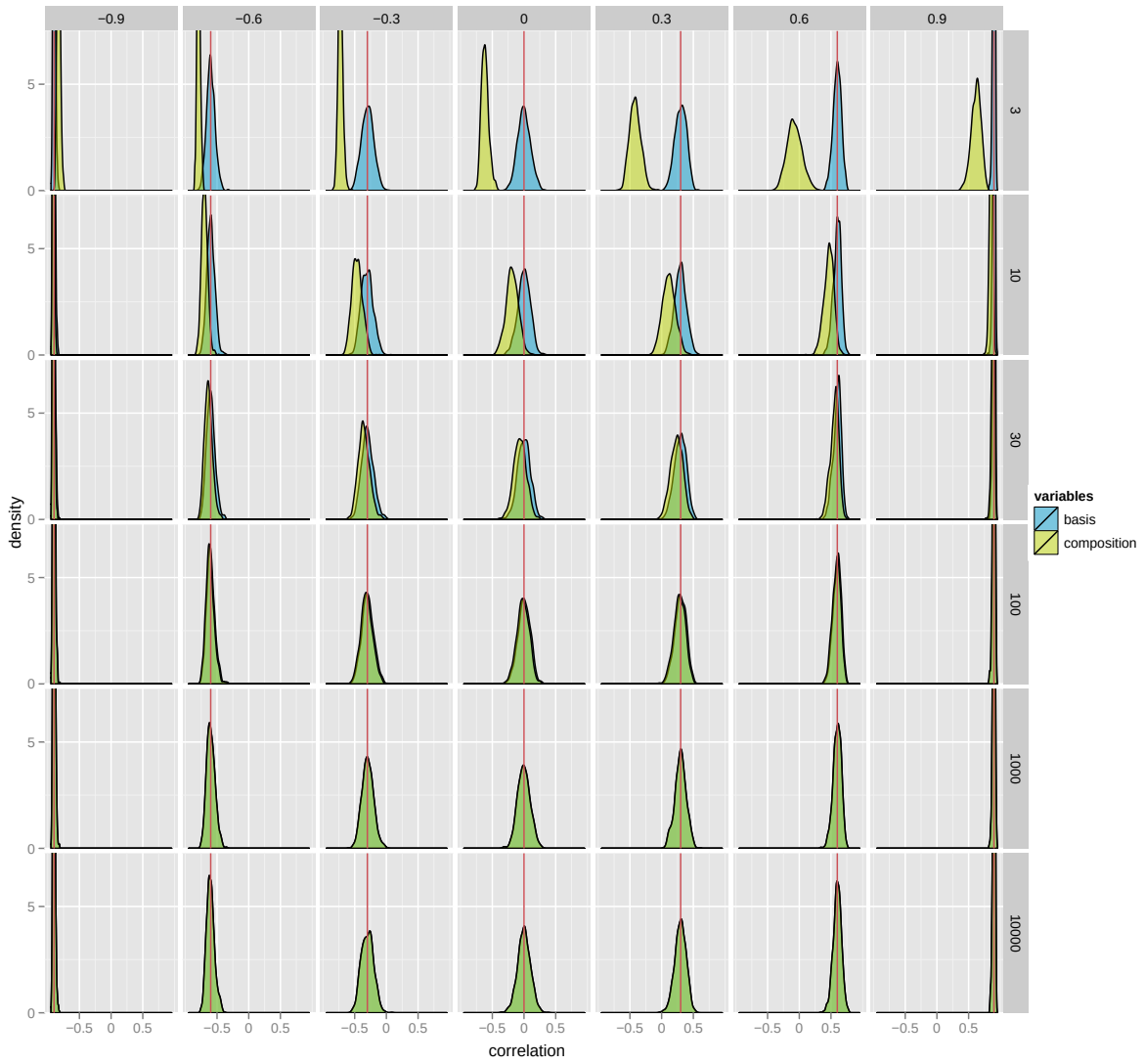

Figure 5.3: Correlation coefficients observed in a simulation of two variables of interest (w_1, w_2) that are statistically independent of a third (w_3) which represents the sum of $D - 2$ other measurements, each similar in size to w_1, w_2. Each column of panels (and red vertical line) corresponds to a different true correlation between $\log w_1$ and $\log w_2$. Each row corresponds to a different value of D, from 3 through to 10^4. The blue probability densities show the sample correlations observed between $\log w_1$ and $\log w_2$ in simulation; the green probability densities show the corresponding correlations observed between $\log x_1$ and $\log x_2$, the elements in $\mathcal{C}(\mathbf{w})$. Each correlation coefficient was calculated from 100 simulated data points; each panel shows the distributions of 1000 correlation coefficients.

This highlighted that the data simulated by Vêncio *et al.* [11] was already pretty close to compositional, and that today's "next generation" sequencers place sum constraints on data because they report a large, but finite and relatively constant total number of sequence reads.

6 Implications

This report has explored the potential for sum-constrained data to lead analyses of omics data astray. We have seen that *provided the components of interest are relatively small parts of mixture samples that remain relatively constant in size and composition*, univariate statistics, distances on log-transformed components, and correlations between log-transformed components will not lead us to draw radically different conclusions to analyses on unconstrained data.

The main problem is: we can't tell when that proviso holds using compositional data alone.

We think this has two main implications:

1. That, wherever possible, experimentalists should gather additional information that allows the absolute abundance of the components under study to be inferred.
2. That, when only relative abundance information exists, data are analysed using appropriate methods.

Like any good implications, these two have consequences...

6.1 Gathering information to infer absolute abundance

In the footnote to our introductory quotation of Aitchison's on page 4 we stressed that absolute abundance of specimens (*e.g.*, mRNAs, organisms, *etc.*) is often very important in the biosciences. We introduced The Omics Imp as a means to show how different experimental paradigms can determine whether absolute abundance can be inferred, and how relative abundance alone does not tell us about how many copies of an mRNA are being produced.

Miura *et al.* [21] and Kanno *et al.* [7] describe methods to measure mRNA absolute abundance, stating eloquently that

> "An ideal format to describe transcriptome would be its composition[8] measured on the scale of absolute numbers of individual mRNAs per cell. It would help not only to precisely grasp the structure of the transcriptome but also to accelerate data exchange and integration." [21]

and

> "Transcriptome data from quantitative PCR (Q-PCR) and DNA microarrays are typically obtained from a fixed amount of RNA collected per sample. Therefore, variations in tissue cellularity and RNA yield across samples in an experimental series compromise accurate determination of the absolute level of each mRNA species per cell in any sample. Since mRNAs are copied from genomic DNA, the simplest way to express mRNA level would be as copy number per template DNA, or more practically, as copy per cell." [7]

To the best of our knowledge, application of these methods is not yet commonplace, but we hope that this technical report will serve as an argument for these, and other absolute abundance techniques, to be employed more often in the pursuit of understanding biological systems.

We note with interest the recent pre-production publication of Robinson and Oshlack [22] who

[8] Note that Miura *et al.* are using the term *composition* in its general sense, rather than as per Definition 1. In other words, they suggest to describe a cell's transcriptome in terms of the counts of each different kind of transcript present in that cell.

propose "an empirical strategy [for estimating the relative RNA production levels in two samples] that equates the overall expression levels of genes between samples under the assumption that the majority of them are not [differentially expressed]." We look forward to the full publication with supplemental material as this approach appears to offer a relatively simple means to compare RNA-seq measurements when this assumption holds.

6.2 Analysing compositional omics data appropriately

There are circumstances where omics data are truly relative (*e.g.*, metabolite concentrations within the bloodstream), or when interest genuinely centres on comparing relative amounts (*e.g.*, the nucleotide or codon composition of samples of genomic DNA). There are also many circumstances where measurements have been made in a ways that ensure that data carry only relative information (*e.g.*, RNA-seq or microarray data obtained from fixed volumes of total RNA). In their seminal paper on RNA-seq, Mortazvi *et al.*[8] explicitly render their data compositional by working in terms of reads per kilobase of exon per million mapped sequence reads (RPKM). (By working with fixed weight aliquots of mRNA and using a sequencing platform that has limits (albeit very large ones) on the number of sequences that can be read, the data were already constrained to be compositional.)

In these situations, we think much more needs to be done to apply compositional data analysis methods instead of analysis techniques that assume data are unconstrained. This technical report has shown that simply log-transforming the compositional data is not a panacea—we need to be sure that the components of interest are relatively small parts of mixture samples that remain relatively constant in size and composition, and this cannot be determined using compositional data alone.

Aitchison [1] has pioneered development of methods for compositional data analysis, founded upon *logratios* of components[9]. We conjecture that bringing these methods into play with omics data would mean, for example

- working with (log) ratios of fluorescence intensities *between* spots within a microarray. This would be an explicitly multivariate treatment of the data rather than, say, the conventional approach of multiple univariate analyses that seek to test for significant differential expression. (One of the beliefs that has to be abandoned in working with compositional data is the idea that a single component means anything in isolation—it is only meaningful *relative* to other components.) We wonder also whether adopting this approach would obviate or simplify the process of microarray normalisation that seeks to render arrays comparable within and across experiments.
- working with (log) ratios of mRNA counts within RNA-seq runs.
- adopting Aitchison's distance as a metric for compositional comparison. Given the relationship between Aitchison's distance and Euclidean distance with log-transformed data (Equation 4.12), and the fact that omics data is often log-transformed before hierarchical clustering or other distance-based methods are applied, this may not lead to dramatically different results across the board. However, in areas that use Euclidean distance on (un-

[9]Note that the interactive plotting software described in Section 5.2 displays the *variation matrix* of the three-part composition: $\mathrm{var}(\log(x_i/x_j))$. As Aitchison states "The covariance structure of a composition is completely determined by knowledge of the covariance structure, that is the logratio variances, of all of its 2-part compositions" [1, p.74].

transformed) compositional data, we expect the application of Aitchison's distance to provide more meaningful insights.

We can see that omics data poses challenges to compositional data analysis methods. Datasets often contain zero measurements—either because a component was not present, or because it was present but not sampled, or because some measurement error occurred. The problem of zeros becomes more pernicious the less that samples have in common, e.g., metagenomic samples drawn from very different environments. Of course, this is not so much a defect of compositional data analysis methods as a sharp reminder that comparing samples with different attributes is an ill-posed problem.

A second challenge posed by omics data to compositional (indeed *any*) analysis methods is the paucity of independent samples (n) in comparison to the abundance of measurements (p). Modern bioscience data is as notoriously high-dimensional as modern bioscience data collection is underfunded, and $p \gg n$ datasets are commonplace. The industrialization of biology has us, at present, in a situation where it is feasible to make millions of measurements on a few individuals, but not *vice versa*. Funding levels and measurement costs favour small experiments (say a comparison of two treatments using three biological replicates per treatment) which, through omics technologies, generate thousands to millions of measurements.

We acknowledge these challenges, both methodological and financial. However, our primary aim is to ensure that bioscientists are not lead astray by artifacts of the measurement process, and we hope through this, and subsequent publications that awareness will be raised about the need to handle compositional data appropriately.

7 Acknowledgments

We thank Ms Ning Wang for allowing us to use the microRNA data of Figure 1.4. We gratefully acknowledge colleagues who provided us with feedback on the ideas presented here. In particular, we thank

- Rob Knight (University of Colorado, Boulder) for suggesting the exposition depicted in Figure 1.5 and for other insights that have made this document more accessible
- Paul Thomas (CSIRO) for insights into distance measures used in information retrieval
- Ian Saunders (CSIRO) for meticulous review and errol-correction.

References

[1] J. Aitchison, *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability, Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press), 1986. 416 p.

[2] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana Delgado, "Lecture notes on compositional data analysis." http://dugi-doc.udg.edu//handle/10256/297, May 2007.

[3] J. Aitchison, "The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies," in Daunis-i Estadella and Martin-Fernandez [23]. CD-ROM.

[4] V. Liebscher, "Compositions in life science data," in Daunis-i Estadella and Martin-Fernandez [23]. CD-ROM.

[5] D. A. Jackson, "Compositional data in community ecology: The paradigm or peril of proportions?," *Ecology*, vol. 78, pp. 245–263, Sept. 2008.

[6] S. Chapman, "String similarity metrics for information integration." http://www.dcs.shef.ac.uk/˜sam/stringmetrics.html.

[7] J. Kanno, K. Aisaki, K. Igarashi, N. Nakatsu, A. Ono, Y. Kodama, and T. Nagao, ""Per cell" normalization method for mRNA measurement by quantitative PCR and microarrays," *BMC Genomics*, vol. 7, p. 64, 2006. PMID: 16571132.

[8] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nat Meth*, vol. 5, pp. 621–628, July 2008.

[9] M. N. McCall and R. A. Irizarry, "Consolidated strategy for the analysis of microarray spike-in data," *Nucleic Acids Research*, vol. 36, p. e108, Oct. 2008. PMID: 18676452.

[10] U. Bissels, S. Wild, S. Tomiuk, A. Holste, M. Hafner, T. Tuschl, and A. Bosio, "Absolute quantification of microRNAs by using a universal reference," *RNA (New York, N.Y.)*, vol. 15, pp. 2375–2384, Dec. 2009. PMID: 19861428.

[11] R. Vêncio, L. Varuzza, C. de B Pereira, H. Brentani, and I. Shmulevich, "Simcluster: clustering enumeration gene expression data on the simplex space," *BMC Bioinformatics*, vol. 8, no. 1, p. 246, 2007.

[12] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. London: Chapman and Hall, 2nd ed ed., 1989.

[13] M. D. Robinson and G. K. Smyth, "Moderated statistical tests for assessing differences in tag abundance," *Bioinformatics*, vol. 23, pp. 2881–2887, Nov. 2007.

[14] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, vol. 18, pp. 1509–1517, Sept. 2008.

[15] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[16] R. Gentleman, B. Ding, S. Dudoit, and J. Ibrahim, "Distance measures in DNA microarray data analysis," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pp. 189–208, Springer-Verlag, 2005.

[17] N. J. Gotelli and R. K. Colwell, "Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness," *Ecology Letters*, vol. 4, no. 4, pp. 379–391, 2001.

[18] R. J. Howarth, "Sources for a history of the ternary diagram," *The British Journal for the History of Science*, vol. 29, pp. 337–356, Sept. 1996. ArticleType: primary_article / Full publication date: Sep., 1996 / Copyright 1996 The British Society for the History of Science.

[19] W. Skala, "A mathematical model to investigate distortions of correlation coefficients in closed arrays," *Mathematical Geology*, vol. 9, no. 5, pp. 519–528, 1977.

[20] J. Brehm, S. Gates, and B. Gomez, "A Monte Carlo comparison of methods for compositional data analysis," in *1998 annual meeting of the Society for Political Methodology*, July 1998.

[21] F. Miura, N. Kawaguchi, M. Yoshida, C. Uematsu, K. Kito, Y. Sakaki, and T. Ito, "Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs," *BMC Genomics*, vol. 9, p. 574, 2008. PMID: 19040753.

[22] M. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biology*, vol. 11, no. 3, p. R25, 2010.

[23] J. Daunis-i Estadella and J. E. Martin-Fernandez, eds., *Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop*, University of Girona, Girona (Spain), May 27-30 2008. CD-ROM.

A Using the interactive composition software

We have developed interactive plotting software in Java to help explore the impact of closure on log-transformed basis and compositional data. The underlying data model is described in Section 5.2. Currently, this software is up to version 0.7 and is available by contacting the first author. The software is packaged as a Java archive (`.jar`) file and needs Java to be installed to run (see `www.java.com`). It be executed either by double-clicking on the file's icon (in a windowing environment) or from a command line by typing

```
java -jar "Compositions v07.jar"
```

This will open five windows:

Composition Parameters: contains sliders that control the parameters of the trivariate log-normal distribution from which samples are drawn (Equation 5.5). The Resample button draws a new set of points from this distribution. Closing this window exits the software.

Composition Statistics: shows *sample* covariances, variance, and correlations of (log) components 1 and 2, and the log size of the basis (T). It also shows the sample *variation matrix* var$(\log(x_i/x_j))$, and the *theoretical* mean and standard deviation of w_1, w_2, w_3 (see Equation 5.6).

Composition Scatterplot shows 300 samples from a trivariate log-normal basis in red, and the corresponding composition in blue. Only components 1 and 2 are shown, and axes are drawn on a log scale.

Composition Ternary Plot shows the compositional samples in blue and the corresponding barycentred data points in green [18].

About compositions... provides help and version information.

Contact Us

Phone: 1300 363 400
+61 3 9545 2176

Email: enquiries@csiro.au

Web: www.csiro.au

Your CSIRO

Australia is founding its future on science and innovation. Its national science agency, CSIRO, is a powerhouse of ideas, technologies and skills for building prosperity, growth, health and sustainability. It serves governments, industries, business and communities across the nation.