An Introduction to Metagenomics and Microbial Ecology

Ilia Korvigo

This document is a supplement to an introductory course on microbial ecology and metagenomics taught at the Saint Petersburg State University and the ITMO University. It comprises a concise summary of lecture materials.

Contents

1 Introduction **1**

1.1 Microbial ecology and metagenomics 1

1.2 Amplicon sequencing 1

1.3 The 16S rRNA gene 2

2 A primer on linear algebra **3**

2.1 Disclaimer 3

2.2 Equivalence classes 3

2.3 Algebraic structures 3

2.4 Rings and fields 4

2.5 Vector spaces 4

2.5.1 Algebraic structure 4

2.5.2 Real normed spaces 4

2.5.3 Real inner product spaces 5

2.5.4 Metric spaces 6

2.6 Coordinate systems 6

2.6.1 Linear combination 6

2.6.2 Linear independence 6

2.6.3 Spanning (generating) sets 7

2.6.4 Basis 7

2.7 Linear maps 8

2.7.1 General properties 8

2.7.2 Matrices 8

2.7.3 Rank and nullity 9

2.7.4 Eigenvectors and eigenvalues 9

2.7.5 Diagonalisation 9

2.7.6 Elements of the spectral theorem 10

2.8 Choice of basis and covariance 11

2.8.1 Covariance 11

2.8.2 Principle component analysis 12

3 A primer on compositional data analysis (CoDA) **13**

3.1 Disclaimer 13

3.2 Compositions as equivalence classes 13

3.3 Vector space structure of $\mathcal{S}^D$ 13

3.4 Inner product space structure 14

3.5 Principles of compositional data analysis 14

3.5.1 Permutation invariance 14

3.5.2 Scale invariance 15
3.5.3 Subcompositional coherence (dominance) 15
3.6 Generating systems of $\mathcal{S}^D$ 15
3.7 Centred log-ratio transformation and coefficients 16
3.8 Isometric (orthonormal) log-ratio transformation and coordinates 18
3.8.1 General construction 18
3.8.2 Sequential binary partition 18
3.9 Additive log-ratio transformation 20

4 Amplicon Data Preprocessing 22
4.1 Problem statement 22
4.2 Clustering-based preprocessing 22
4.3 Denoising-based preprocessing 23
4.4 Chimeric sequences 23
4.5 Copy-number variations 23

5 Taxonomic annotation 25
5.1 Problem statement 25
5.2 Reference datasets 25
5.3 Alignment-based annotation 25
5.4 Machine-learning-based annotation 26
5.5 Phylogenetic placement 26

6 Statistical analysis: introduction 27
6.1 Problem statement 27
6.2 Zeros in compositional data 28
6.2.1 Types of zeros 28
6.2.2 Zero replacement 28

1 Introduction

1.1 Microbial ecology and metagenomics

Microbial ecology is the study of microbiomes, that is consortia of arbitrarily diverse microbial populations interacting with each other and the environment they inhabit. Metagenomics is an emergent field within computational biology comprising a loose grouping of experimental, computational and statistical methods dedicated to generating and analysing highly heterogeneous genetic information (i.e. mixtures of genomes originating from distinct organisms). While we should not confuse microbial ecology with metagenomics (as is often done in contemporary parlance), metagenomics has become the de facto standard way to study microbial communities, inasmuch as genetic information (be it DNA or RNA) is the only type of information we can extract from said communities efficiently enough. In other words, modern microbiome research is deeply grounded in metagenomics, though the latter has other applications. Metagenomic data come in two flavours: shotgun (aka "complete") metagenomes and amplicon libraries. As their names suggest, the difference boils down to the amount (or type) of genetic information we sample from an environment. While shotgun metagenomics is becoming increasingly accessible and viable (though not enough to become truly practical in most cases), amplicon libraries remain the most common type of metagenomic data (whilst not being metagenomes *sensu stricto*) due to an unparalleled combination of cost efficiency, community coverage and analytical convenience, which is why this course only covers amplicon metagenomics.

1.2 Amplicon sequencing

Unlike shotgun metagenomes, amplicon libraries comprise different instances of a single target DNA sequence. Any sequence is a valid target as long as we can design a pair of universal primers compatible with all its instances within a selected phylogenetic group (be it large or small). Targeting a specific piece of DNA instead of entire genomes greatly reduces sequencing costs, downstream computational requirements and improves diversity coverage (especially in low-entropy communities) at the expense of direct functional information (unless the target itself determines the function of interest). Whether the target is a sequence pertaining to some functional activity (e.g. an enzyme or a receptor) or a universal marker-gene with no immediate functional implications (e.g. 16S rRNA), it contains phylogenetic information about the underlying phylogeny of captured community components. Students are highly encouraged to read [1] and [2] to get a better grasp of phylogenetic information as a concept, though for now we can make do with a simple rule of thumb: a target sequence should be as long as possible and neither too conserved, nor too variable.

1.3 The 16S rRNA gene

The 16S rRNA gene is by far the most common target for amplicon metagenomics. It is not only truly omnipresent, but also conserved enough to enable universal primers compatible with all known prokaryotic sequences (as well as some mitochondrial and plastid sequences and even 18S eukaryotic genes [3]). The gene consists of short conserved patches interspersed with 9 longer hyper-variable regions. The most prominent of these are hyper-variable regions V3 ($\sim$ 180 bp) and V4 ($\sim$ 290 bp) Most of the older microbiome studies targeted either of the two, while newer studies are increasingly targeting both regions (i.e. the V3-V4 fragment) thanks to widespread adoption and improved quality of 250 bp and 300 bp Illumina paired-end sequencing.

2 A primer on linear algebra

2.1 Disclaimer

2.2 Equivalence classes

Given a set S and a binary relation $\sim$ that is

- reflexive, i.e. $a \sim a$,
- symmetric, i.e. $a \sim b \Rightarrow b \sim a$,
- transitive, i.e. $a \sim b, b \sim c \Rightarrow a \sim c$,

for any $a, b, c \in S$, the equivalence class of an element $a \in S$ is given by $\{x \in S \mid x \sim a\}$.

2.3 Algebraic structures

	Closure	Associativity	Identity	Invertability	Commutativity
Monoid	🟩	🟩	🟩	🟥	🟥
Group	🟩	🟩	🟩	🟩	🟥
Abelian group[1]	🟩	🟩	🟩	🟩	🟩

Table 1: Algebraic properties of monoids, groups and abelian groups. Colours show whether a structure requires a property or not: green – requires, red — does not require.

Given a set S endowed with a binary operator f:

- S is closed under $f \iff f : S^2 \to S$;
- f is associative $\iff f(f(a,b),c) = f(a, f(b,c))$ for all $a, b, c \in S$;
- there is an identity (a neutral element) under $f \iff \exists\ 0 \in S : f(a, 0) = a$ for all $a \in S$;
- there is an inverse under $f \iff \forall\ a \in S\ \exists -a : f(a, -a) = 0$;
- f is commutative $\iff f(a,b) = f(b,a)$ for all $a, b \in S$.

[1] Abelian groups are also called commutative groups

2.4 Rings and fields

A set R endowed with two binary operators, $(+)$ and $(\cdot)$, is a ring if

- R forms an abelian group under $(+)$ (R is an additive abelian group);
- R forms a monoid under $(\cdot)$ (R is a multiplicative monoid);
- $(\cdot)$ is (left and right) distributive with respect to $(+)$, i.e. $a \cdot (b + c) = a \cdot b + a \cdot c = (b + c) \cdot a$ for any $a, b, c \in R$.

A ring F is a field if it forms an abelian group under $(\cdot)$.

2.5 Vector spaces

2.5.1 Algebraic structure

A vector space V over a field F is an additive abelian group endowed with a binary operator $(\cdot)$ (called scalar multiplication), such that

- V is closed under $(\cdot)$, i.e. $(\cdot) : F \times V \to V$;
- $1 \cdot \mathbf{v} = \mathbf{v}$ for all $\mathbf{v} \in V$, where $1 \in F$ is the field multiplicative identity, i.e. $(\cdot)$ respects identity of field multiplication;
- $(\alpha + \beta) \cdot \mathbf{v} = \alpha \cdot \mathbf{v} + \beta \cdot \mathbf{v}$ and $\alpha \cdot (\mathbf{v} + \mathbf{w}) = \alpha \cdot \mathbf{v} + \alpha \cdot \mathbf{w}$ for all $\mathbf{v}, \mathbf{w} \in V$ and $\alpha, \beta \in F$, i.e. $(\cdot)$ is distributive with respect to field addition and vector addition (here symbol $+$ corresponds to both operators as appropriate);
- $(\alpha \cdot \beta) \cdot \mathbf{v} = \alpha \cdot (\beta \cdot \mathbf{v})$ for all $\alpha, \beta \in F, \mathbf{v} \in V$, i.e. $(\cdot)$ respects associativity of field multiplication (here symbol $\cdot$ denotes both the field multiplication and the scalar multiplication as appropriate).

Vector spaces over $\mathbb{R}$ (i.e. the field of real numbers) are also called real vector spaces.

2.5.2 Real normed spaces

Given a real vector space V, a unary operator $\|\cdot\| : V \to \mathbb{R}$ is called a norm if

- it is subadditive: $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in V$;
- it is absolutely scalable: $\|\alpha \cdot \mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$ for any $\alpha \in \mathbb{R}$ and $\mathbf{x} \in V$;

- it is positive-definite: $\|\mathbf{v}\| \geq 0$ for any $\mathbf{v} \in V \setminus \{\mathbf{0}\}$, that is $\|\mathbf{v}\| = 0 \iff \mathbf{v} = \mathbf{0}$.

If $\|\cdot\|$ is only positive-semidefinite (i.e. if it can assign zero to non-zero vectors), it is called a seminorm or a pseudonorm. A real vector space endowed with a norm is called a real normed space.
Norms generalise Euclidean notions of vector magnitude (length) and distances (see subsection 2.5.4). If $\|\mathbf{v}\| = 1$, $\mathbf{v}$ is called a unit vector. Any vector in a normed space can be transformed into a unit vector via normalisation:

$$\hat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|}$$

2.5.3 Real inner product spaces

Given a real vector space V, a binary operator $\langle \cdot, \cdot \rangle : V^2 \to \mathbb{R}$ is called an inner product if

- it is symmetric: $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ for any $\mathbf{x}, \mathbf{y} \in V$;
- it is bilinear: $\langle \alpha \cdot \mathbf{x} + \beta \cdot \mathbf{y}, \mathbf{z} \rangle = \alpha \cdot \langle \mathbf{x}, \mathbf{z} \rangle + \beta \cdot \langle \mathbf{y}, \mathbf{z} \rangle$ for any $\alpha, \beta \in \mathbb{R}$ and any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$;
- it is positive-definite: $\langle \mathbf{x}, \mathbf{x} \rangle > 0$ for all $\mathbf{x} \in V \setminus \{\mathbf{0}\}$, that is $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \iff \mathbf{x} = \mathbf{0}$

If the operator is only positive-semidefinite, it is called a semi-inner product or a scalar product, though — rather confusingly — this name also refers to the real dot product, which is the Euclidean inner product. A real vector space endowed with an inner product is called a real inner product space. Any inner product defined on a real vector space induces a real norm $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ (likewise semi-inner products induce seminorms).
In addition to the concepts associated with the norms they induce, inner products generalise Euclidean notions of angles between vectors and orthogonal projections. In particular, orthogonal projection of any vector $\mathbf{x}$ on a non-zero vector $\mathbf{y}$ can be defined as:

$$\operatorname{Proj}(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle} \cdot \mathbf{y} = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|} \cdot \frac{\mathbf{y}}{\|\mathbf{y}\|} = \langle \mathbf{x}, \hat{\mathbf{y}} \rangle \cdot \hat{\mathbf{y}}$$

The angle between two non-zero vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\cos\left(\theta_{\mathbf{x},\mathbf{y}}\right) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}$$

Consequently, we can also define the orthogonal projection of $\mathbf{x}$ on $\mathbf{y}$ in terms of the angle between these vectors:

$$\langle \mathbf{x}, \hat{\mathbf{y}} \rangle = \cos\left(\theta_{\mathbf{x},\mathbf{y}}\right) \cdot \|\mathbf{x}\| \Rightarrow \operatorname{Proj}\left(\mathbf{x}, \mathbf{y}\right) = \cos\left(\theta_{\mathbf{x},\mathbf{y}}\right) \cdot \|\mathbf{x}\| \cdot \hat{\mathbf{y}}$$

Two non-zero vectors $\mathbf{x}, \mathbf{y} \in V$ are called orthogonal if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.

2.5.4 Metric spaces

Given a set X, a binary function $d(a, b) : X^2 \to \mathbb{R}$ is called a metric, if

- it is symmetric, i.e. $d(a, b) = d(b, a)$ for any $a, b \in X$
- it satisfies the triangle inequality, i.e. $d(a, b) \leq d(a, c) + d(b, c)$ for any $a, b, c \in X$
- all elements in X are separable under d, i.e. $d(a, b) = 0 \iff a = b$ for any $a, b \in X$

If the last property does not hold, then d is called a semimetric or a pseudometric. A set X endowed with a metric is called a metric space.
Although not every metric space is a vector space (let alone an inner product space), any real norm induces a metric $d(a, b) = \|a - b\|$. Consequently, any real inner product space (or, more generally, any real normed space) is a metric space. A real vector space endowed with the Euclidean inner product, norm and metric is called a Euclidean vector space.

2.6 Coordinate systems

2.6.1 Linear combination

Let V be a vector space over a field F. Then the expression of form

$$\sum_{i=1}^{n} \alpha_i \cdot \mathbf{v}_i,\ \alpha_i \in F, \mathbf{v}_i \in V$$

is called a linear combination of vectors $S = (\mathbf{v}_1, \dots, \mathbf{v}_n)$ with coefficients $(\alpha_1, \dots, \alpha_n)$. The set of all possible linear combinations of vectors S is called the span of S.

2.6.2 Linear independence

A linear combination with coefficients $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$ is called trivial. S is called linearly independent if only its trivial linear combination yields the zero vector in V. Alternatively, S is called linearly independent if there are no coefficients $(\alpha_1, \dots, \alpha_{n-1})$ such that $s_n = \sum_{i=1}^{n-1} \alpha_i \cdot \mathbf{v}_i$.

2.6.3 Spanning (generating) sets

Given a vector space V over a field F, an ordered set of n non-zero vectors $B = (\mathbf{v}_1, \ldots, \mathbf{v}_n) \subset V$ is called a spanning set (or a generator[2]) of V if for any vector $\mathbf{x} \in V$ there is an ordered sequence of scalars (coefficients) $(\alpha_1, \ldots, \alpha_n)$ such that $\mathbf{x} = \sum_{i=1}^{n} \alpha_i \cdot \mathbf{v}_i$. It is often convenient to represent these coefficients as $n \times 1$ and $1 \times n$ matrices called column and row coefficient vectors respectively:

$$\mathbf{x} := \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_n \end{bmatrix}^T \Rightarrow \mathbf{x} = \sum_{i=1}^{m} \alpha_i \cdot \mathbf{v}_i$$

2.6.4 Basis

Coefficients of a vector $\mathbf{x}$ with respect to a spanning set are unique (non-degenerate) if and only if the set is linearly independent. Such a spanning set is called a minimal spanning set or a basis of V. All minimal spanning sets of a vector space V have the same cardinality that corresponds to the dimensionality of V. In other words, a vector space V is n-dimensional if the cardinality of any of its minimal spanning sets is equal to n. Consequently, a set of $m \leq n$ linearly independent vectors in V spans an m-dimensional subspace of V. Coefficients with respect to a basis are called coordinates (likewise, their matrix representations are called coordinate vectors[3]). If V is an inner product space with a basis B composed entirely of mutually orthogonal unit vectors[4], then B is an orthonormal basis. In the context of real vector spaces $\mathbb{R}^n$, there is a notion of the standard (or canonical) orthonormal basis:

$$(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n) = \left(\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \right)$$

[2] In the context of vector spaces over fields, the concept of spanning sets is equivalent to an otherwise more general concept of generators or generating systems

[3] In practice, coefficient vectors and coordinate vectors are rarely explicitly disambiguated.

[4] That is basis $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ of an n-dimensional inner product space is called orthonormal if $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = \delta_{ij}$, where δ stands for the Kronecker-delta.

2.7 Linear maps

2.7.1 General properties

Let V and W be vector spaces over a field F, then a linear map (transformation) $f : V \to W$ is a function satisfying the following properties:

- $f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$ for any $\mathbf{x}, \mathbf{y} \in V$;
- $f(\alpha \cdot \mathbf{x}) = \alpha \cdot f(\mathbf{x})$ for any $\alpha \in F$ and $x \in V$

Linear maps of type $f : V \to V$ (i.e. vector space endomorphisms) are oft-called linear operators.

2.7.2 Matrices

If V and W are finite-dimensional vector spaces, any linear map $f : V \to W$ can be defined in terms of the image of basis vectors $(\mathbf{v}_1, \cdots, \mathbf{v}_n) \subset V$ under f. Given a vector $\mathbf{x} = \begin{bmatrix}\alpha_1 & \dots & \alpha_n\end{bmatrix}^T \in V$

$$\begin{aligned}
f(\mathbf{x}) &= f(\begin{bmatrix}\mathbf{v}_1 & \dots & \mathbf{v}_n\end{bmatrix} \times \mathbf{x}) \\
&= f\left(\sum_{i=1}^{n} \alpha_i \cdot \mathbf{v}_i\right) = \sum_{i=1}^{n} \alpha_i \cdot f(\mathbf{v}_i) \\
&= \begin{bmatrix}f(\mathbf{v}_1) & \dots & f(\mathbf{v}_n)\end{bmatrix} \times \mathbf{x}
\end{aligned}$$

where × stands for matrix multiplication. Likewise, we can represent the image of any vector in $(\mathbf{v}_1, \cdots, \mathbf{v}_n) \subset V$ under f in terms of coordinates with respect to a basis $(\mathbf{w}_1, \cdots, \mathbf{w}_m) \subset W$

$$f(\mathbf{v}_j) = \begin{bmatrix}\mathbf{w}_1 & \dots & \mathbf{w}_m\end{bmatrix} \times \begin{bmatrix}\beta_{1j} \\ \vdots \\ \beta_{mj}\end{bmatrix}$$

Consequently, we can express $f(\mathbf{x})$ using an $m \times n$ matrix A

$$f(\mathbf{x}) \coloneqq A \times \mathbf{x} = \begin{bmatrix}\beta_{11} & \beta_{12} & \dots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{m1} & \beta_{m2} & \dots & \beta_{mn}\end{bmatrix} \times \begin{bmatrix}\alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n\end{bmatrix}$$

In other words, matrix A defines $f : V \to W$ with respect to a basis in V and a basis in W.

2.7.3 Rank and nullity

Let V and W be finite-dimensional vector spaces over a field F. Given an $m \times n$ matrix A representing a linear map $f : V \rightarrow W$ (with respect to some choice of basis in V and W), the set of all possible linear combinations (i.e. the span) of its column vectors is called the column space of A, that is the image of V under f. The dimensionality of the column space of A is called the rank of A. Equivalently, the rank of A is the number of basis vectors spanning its column space. The kernel (or the null-space) of A is the set of vectors $\{\mathbf{x} \in V : A \times \mathbf{x} = \mathbf{0}_W\}$. The dimensionality of the kernel (i.e. the number of basis vectors spanning the kernel) is called the nullity of A. Observe, that $\text{rank}(A) + \text{nullity}(A) = n$, where n is the dimensionality of V Given an $n \times n$ matrix M representing a linear operator $f : V \rightarrow V$ (with respect to some choice of basis in V), M is invertible if and only if it has full rank, that is $\text{rank}(M) = n$. Equivalently, M is invertible if and only if the kernel of M corresponds to the empty subspace of V (i.e. the subspace spanned by the zero-vector in V). Non-invertible square matrices are also called singular or degenerate.

2.7.4 Eigenvectors and eigenvalues

Let V be an n-dimensional vector space over a field F. Given a square $n \times n$ matrix A, representing a linear operator on V, a non-zero vector $\mathbf{v} \in V$ is called an eigenvector of A if there is a corresponding eigenvalue $\lambda \in F$ such that $A \times \mathbf{v} = \lambda \cdot \mathbf{v}$. Consequently, eigenvectors corresponding to different eigenvalues are always orthogonal. A basis composed entirely of eigenvectors of A is called the eigenbasis of A. Observe, that if any eigenvalue is zero, A is a singular matrix.

2.7.5 Diagonalisation

Let V be an n-dimensional vector space over a field F. A square $n \times n$ matrix A is called diagonalisable if it can be transformed into a diagonal[5] matrix D with respect to an ordered set of eigenvectors $(\mathbf{v}_1, \ldots, \mathbf{v}_n) \subset V$

$$D = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}$$

where diagonal entries $(\lambda_1, \ldots, \lambda_n)$ are the corresponding eigenvalues. If all eigenvalues are unique, vectors $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ constitute the eigenbasis of A.

2.7.6 Elements of the spectral theorem

Theorem 1. *Let A be a symmetric*[6] *matrix defined on an n-dimensional real inner product space. Then, A is orthogonally diagonalisable and all its eigenvalues are real. In other words, there exist real scalars $\lambda_1, \ldots, \lambda_n$ (the eigenvalues) and orthogonal non-zero real vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ (the eigenvectors) such that $A \times \mathbf{v}_i = \lambda_i \cdot \mathbf{v}_i$ for any $i \in (1, \ldots, n)$.*

This is a significantly simplified rendition of the spectral theorem for finite-dimensional real-vector spaces.

Proposition 1. *Given any $m \times n$ matrix A defined on a finite-dimensional real inner-product space, the matrices $A \times A^T$ and $A^T \times A$ share the same non-zero eigenvalue*[7]*.*

Proof. Let $\mathbf{v}$ be a non-zero eigenvector of $A^T \times A$ associated with a non-zero eigenvalue λ. Then,

$$\begin{gathered} \left(A^T \times A\right) \times \mathbf{v} = \lambda \cdot \mathbf{v} \\ \left(A \times A^T\right) \times (A \times \mathbf{v}) = \lambda \, (A \times \mathbf{v}) \end{gathered}$$

Consequently, vector $\mathbf{w} = A \times \mathbf{v}$ is an eigenvector of matrix $A \times A^T$ associated with eigenvalue λ. If $\mathbf{w} = \mathbf{0}$, then so is $\lambda \cdot \mathbf{v}$. This is a contradiction, hence $\mathbf{w} \neq \mathbf{0}$. □

This proof also shows that eigenvectors $\mathbf{v}$ and $\mathbf{w}$ of matrices $A^T \times A$ and $A \times A^T$ associated with the same eigenvalue are connected by the following equations:

$$\begin{aligned} \mathbf{w} &= A \times \mathbf{v} \\ \mathbf{v} &= A^T \times \mathbf{w} \end{aligned}$$

Proposition 2. *The eigenvalues of $A^T \times A$ and $A \times A^T$ are non-negative.*

[5] A square matrix is called diagonal, if all its off-diagonal entries are equal to 0

[6] A matrix A is called symmetric if $A = A^T$

[7] Observe that both matrices are symmetric, because $(A \times B)^T = B^T \times A^T$.

Proof. Let $\mathbf{v}$ be an eigenvector of $A^T \times A$ associated with an eigenvalue λ. Then,

$$\begin{aligned}\|A \times \mathbf{v}\|^2 &= \langle A \times \mathbf{v}, A \times \mathbf{v} \rangle \\ &= (A \times \mathbf{v})^T \times (A \times \mathbf{v}) \\ &= \mathbf{v}^T \times \left(A^T \times A\right) \times \mathbf{v} \\ &= \lambda \cdot (\mathbf{v}^T \times \mathbf{v}) \\ &= \lambda \cdot \|\mathbf{v}\|^2\end{aligned}$$

Since both $\|A \times \mathbf{v}\|^2$ and $\|\mathbf{v}\|^2$ are non-negative, so is λ. □

2.8 Choice of basis and covariance

In this section we are going to discuss, how the choice of basis affects covariance (which is a cornerstone of applied statistics) without laying any groundwork that is necessary to establish, why and when random variables can be treated as vector space objects[8].

2.8.1 Covariance

Let R be the set of all real random variables with finite second moment defined on the same probability space. Then, covariance and variance can be defined as

$$\begin{aligned}\operatorname{cov}(X, Y) &= \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right] \\ &= \mathrm{E}\left[XY - 2\mathrm{E}[X]\,\mathrm{E}[Y] + \mathrm{E}[X]\,\mathrm{E}[Y]\right] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] \\ \sigma^2(X) &= \operatorname{cov}(X, X)\end{aligned}$$

where $X, Y \in R$. Observe, that covariance satisfies all properties of a semi-inner product[9], making the squared root of variance (a.k.a the standard deviation) a seminorm induced by covariance. This observation allows us

[8] Inquisitive students can find introductions and/or references to all relevant concepts at `https://www.randomservices.org/random/expect/Spaces.html`, though I implore you to start from the very first chapter at `https://www.randomservices.org/random/index.html`. Take note, the course assumes familiarity with linear algebra and calculus.

[9] Covariance is not an inner product on the entirety of R, because for any random variable $X \in R : \sigma^2(X) = 0$, $\operatorname{cov}(X, X) = 0$, that is covariance is not positive-definite. This issue can be addressed by restricting covariance to a quotient of R, making covariance (and, by extension, variance) positive-definite. For further details, you can read about quotient spaces (in the context of linear algebra).

to interpret variance and covariance geometrically (i.e. in terms of vector magnitudes and orthogonal projections).

2.8.2 Principle component analysis

In general, for n jointly-distributed random variables with finite second moment we can define a covariance matrix Σ, that contains pair-wise covariances between these random variables. Diagonal entries in this matrix correspond to variances of individual random variables. Since Σ is a symmetric positive-semidefinite matrix, it is orthogonally diagonalisable (see subsection 2.7.6) and all its eigenvalues are non-negative. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ be the eigenvalues of Σ with corresponding orthonormal eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_n$. These orthonormal eigenvectors are called principle components of Σ. The sum of eigenvalues is equal to $\operatorname{tr}(\Sigma)$ (the trace[10] of Σ), the so-called total variance. In statistical parlance $\frac{\lambda_i}{\operatorname{tr}(\Sigma)}$ is called the fraction of variance explained by the i-th principal component.

From a geometrical point of view, we select the first principle component (PC) by maximising the magnitude (variance) projected onto it by original random variables. We then consider orthogonal rejections of original random variables from that PC and find the second principal component following the same magnitude-maximisation objective. We repeat this process until the rejections of original random variables from the span of principle components becomes empty, whereupon the number of PCs will have reached the rank of Σ. From an algebraic point of view, principal components and their eigenvalues correspond to an eigenbasis of Σ, and thus the process is nothing more but a change of basis.

Observe, that the change of basis alters the original covariance structure. Consequently, principle component analysis (PCA) is a special case of covariance structure optimisation. Not all covariance structures are equally helpful when we deal with multivariate statistical analysis. A vector maximising the variance with respect to a factor of interest might be (and usually is) some linear combination of original coordinates. Although PCA is often used for explorative data analysis, it provides no guarantees that any principle component maximises the variance associated with a factor of interest (i.e. captures the target source of variance). There are other common methods specifically designed for the task (for example, linear discriminant analysis).

[10] The sum of diagonal entries.

3 A primer on compositional data analysis (CoDA)

3.1 Disclaimer

3.2 Compositions as equivalence classes

A D-tuple[11] $\mathbf{x} = (x_1, \dots, x_D), x_i \in \mathbb{R}_+$ is a D-part composition, if it only carries relative information about the parts, i.e. if $\mathbf{x}$ is equivalent to $\kappa \cdot \mathbf{x} = (\kappa \cdot x_1, \dots, \kappa \cdot x_1)$ for any $\kappa \in \mathbb{R}_+$. In other words, compositions are equivalent if they are equal under closure[12]

$$\mathcal{C}\left(\mathbf{x}\right) = \left(\frac{x_1}{\sum \mathbf{x}}, \dots, \frac{x_D}{\sum \mathbf{x}}\right)$$

The sample space of D-part compositions is an Aitchison simplex

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \dots, x_D) \,\middle|\, x_i \in \mathbb{R}_+; \sum_{i=1}^{D} x_i = \kappa \right\}$$

where κ is the total sum constraint. Consequently, $\mathcal{S}^D$ contains one member of each equivalence class: without loss of generality, we can assume $\kappa = 1$ in the future.

3.3 Vector space structure of $\mathcal{S}^D$

Perturbation is a binary operator $\oplus : \mathcal{S}^D \times \mathcal{S}^D \to \mathcal{S}^D$

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left(x_1 \cdot y_1, \dots, x_D \cdot y_D\right)$$

Perturbation induces an additive abelian group structure on $\mathcal{S}^D$, such that

- D-tuple $\mathbf{0} = \left(\frac{1}{D}, \dots, \frac{1}{D}\right)$ is the additive identity;
- for any $\mathbf{x} \in \mathcal{S}^D$ there is an additive inverse in $\mathcal{S}^D$ given by $\mathbf{x}^{-1} = \left(x_1^{-1}, \dots, x_D^{-1}\right)$.

Powering is a binary operator $\odot : \mathbb{R} \times \mathcal{S}^D \to \mathcal{S}^D$

$$\alpha \odot \mathbf{x} = \mathcal{C}\left(x_1^{\alpha}, \dots, x_D^{\alpha}\right)$$

It satisfies the properties of scalar multiplication:

[11] While it is tempting to call this D-tuple a vector in $\mathbb{R}_+^D$, the latter is not compatible with the algebraic structure of real vector spaces. Consequently, we cannot call the D-tuple a vector at all, unless we define a compatible vector space structure (refer to subsection 3.3)

[12] Although this definition holds for closure to any scalar $\kappa \in \mathbb{R}_+$, it is sufficient to consider $\kappa = 1$, i.e. the unit-closure.

- $\mathcal{S}^D$ is closed under $\odot$;
- $\odot$ respects the identity of field multiplication, i.e. $1 \odot \mathbf{x} = \mathbf{x}$ for any $\mathbf{x} \in \mathcal{S}^D$;
- $\odot$ respects associativity of field multiplication, i.e. $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha \cdot \beta) \odot \mathbf{x}$ for any $\alpha, \beta \in \mathbb{R}$ and $\mathbf{x} \in \mathcal{S}^D$;
- $\odot$ is distributive with respect to field addition, i.e. $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$ for any $\alpha, \beta \in \mathbb{R}$ and $\mathbf{x} \in \mathcal{S}^D$;
- $\odot$ is distributive with respect to perturbation, i.e. $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$ for any $\alpha \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$.

Consequently, perturbation and powering induce a vector space structure in $\mathcal{S}^D$ over the field $\mathbb{R}$.

3.4 Inner product space structure

The Aitchison inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \cdot \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \cdot \ln \frac{y_i}{y_j}, \text{ where } \mathbf{x}, \mathbf{y} \in \mathcal{S}^D$$

Like any other real inner product, Aitchison inner product induces a norm and a metric:

- $\|\mathbf{x}\|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a}$ – Aitchison norm;
- $d_a(x, y) = \|\mathbf{x} \ominus \mathbf{y}\|_a$ – Aitchison metric (distance).

3.5 Principles of compositional data analysis

Statistical analyses of compositional data should follow three major principles.

3.5.1 Permutation invariance

A function is permutation invariant if it yields equivalent results when the ordering of the parts in the composition is changed.

3.5.2 Scale invariance

The scale must be irrelevant. A function f defined on $\mathbb{R}_+^D$ is scale invariant if it yields the same result for all compositionally equivalent vectors, that is $f(\lambda \cdot \mathbf{x}) = f(\mathbf{x})$ is true for any $\mathbf{x} \in \mathcal{S}^D$ and $\lambda \in \mathbb{R}_+$.

3.5.3 Subcompositional coherence (dominance)

Subcompositions must behave like orthogonal projections in real analysis. For example, given a vector $\mathbf{x} \in \mathbb{R}^D$, the norm of an orthogonal projection of $\mathbf{x}$ onto any subspace $V \subset \mathbb{R}^D$ is always less than, or equal to, the norm of $\mathbf{x}$ in $\mathbb{R}^D$.

3.6 Generating systems of $\mathcal{S}^D$

Vectors in $\mathcal{S}^D$ can be (and usually are) expressed with respect to the standard basis $E = (\mathbf{e}_1, \ldots, \mathbf{e}_D)$ in the D-dimensional real vector space $\mathbb{R}^D$

$$\mathbf{x} = \sum_{i=1}^{D} x_i \cdot \mathbf{e}_i = x_1 \cdot \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + x_2 \cdot \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \cdots + x_D \cdot \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \quad \text{where } \mathbf{x} \in \mathcal{S}^D$$

At the same time, E is neither a basis, nor even a generator of $\mathcal{S}^D$ with respect to the vector space structure of $\mathcal{S}^D$:

- $E \not\subset \mathcal{S}^D$;
- there is an infinite number of coefficient vectors $(\alpha_1, \ldots, \alpha_D) : \alpha_i \in \mathbb{R}$, such that $\sum_{i=1}^{D} \alpha_i \cdot \mathbf{e}_i \notin \mathcal{S}^D$.

However, one way to acquire a generator of $\mathcal{S}^D$ is to take element-wise exponentials of every vector in E and apply closure

$$\begin{aligned} U = (\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_D) &= \left(\mathcal{C} \begin{bmatrix} e^1 \\ e^0 \\ \vdots \\ e^0 \end{bmatrix}, \mathcal{C} \begin{bmatrix} e^0 \\ e^1 \\ \vdots \\ e^0 \end{bmatrix}, \ldots, \mathcal{C} \begin{bmatrix} e^0 \\ e^0 \\ \vdots \\ e^1 \end{bmatrix} \right) \\ &= \left(\mathcal{C} \begin{bmatrix} e \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \mathcal{C} \begin{bmatrix} 1 \\ e \\ \vdots \\ 1 \end{bmatrix}, \ldots, \mathcal{C} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ e \end{bmatrix} \right) \end{aligned} \tag{1}$$

Then, any vector $\mathbf{x} \in \mathcal{S}^D$ can be written as a linear combination of U consistent with the vector space structure of $\mathcal{S}^D$

$$\mathbf{x} = \bigoplus_{i=1}^{D} \ln(x_i) \odot \mathbf{u}_i \tag{2}$$

Generator U is not linearly independent, because the following non-trivial linear combination yields the zero vector

$$\bigoplus_{i=1}^{D} 1 \odot \mathbf{u}_i = \mathcal{C} \begin{bmatrix} e \\ \vdots \\ e \end{bmatrix} = \begin{bmatrix} 1/D \\ \vdots \\ 1/D \end{bmatrix} = \mathbf{0} \in \mathcal{S}^D$$

Alternatively, without loss of generality, it is sufficient to show that $\mathbf{u}_D$ can be represented as a linear combination of $(\mathbf{u}_1, \ldots, \mathbf{u}_{D-1})$ with $D - 1$ coefficients $(-1, \ldots, -1)$, that is

$$\mathbf{u}_D = \bigoplus_{i=1}^{D-1} -1 \odot \mathbf{u}_i$$

Consequently, U is not a basis of $\mathcal{S}^D$, though any set of $D-1$ linearly independent vectors in $\mathcal{S}^D$ is. This is, of course, due to the total sum constraint of vectors in $\mathcal{S}^D$: once any $D - 1$ parts in a D-part composition are fixed the remaining part follows automatically — $\mathcal{S}^D$ is a $(D-1)$-dimensional vector space. This also implies that the covariance matrix of a random D-part composition does not have full rank, i.e. it is singular. This implication has a tremendous effect on all covariance-based statistical methods (e.g. linear models, including ANOVA, location tests and even principal component analysis).

3.7 Centred log-ratio transformation and coefficients

Coefficients with respect to a generator, as seen in equations 1 and 2, are not unique: any expression of form

$$\mathbf{x} = \bigoplus_{i=1}^{D} \ln \frac{x_i}{\lambda} \odot \mathbf{u}_i, \ \text{ where } \ \lambda \in \mathbb{R}_+$$

provides an equivalent result (which is consistent with the notion of equivalence classes in compositional data). If we choose $\lambda = \mathrm{g}_m(\mathbf{x})$ (the geometric mean of $\mathbf{x}$), then the resulting expression induces the centred log-ratio transform of $\mathbf{x}$:

$$\operatorname{clr}(\mathbf{x}) = \mathbf{x}' = \begin{bmatrix} \ln \frac{x_1}{\mathrm{g}_m(\mathbf{x})} & \dots & \ln \frac{x_D}{\mathrm{g}_m(\mathbf{x})} \end{bmatrix} \tag{3}$$

where $\mathbf{x}'$ is the vector of CLR coefficients with respect to U. The transformation is invertible. Given a vector of CLR coefficients $\mathbf{x}'$, the coefficients with respect to the standard basis of $\mathbb{R}^D$ are given by

$$\operatorname{clr}^{-1}\left(\mathbf{x}'\right) = \mathcal{C}\begin{bmatrix} \exp(x'_1) & \dots & \exp(x'_D) \end{bmatrix} \tag{4}$$

Observe, that since a vector of CLR coefficients (or any other coefficients with respect to U) has D entries, it is subject to a total sum constraint, in particular $\sum \operatorname{clr}(\mathbf{x}) = 0$ for any $\mathbf{x}' \in \mathcal{S}^D$, because

$$\ln\left(\mathrm{g}_m(\mathbf{x})\right) = \ln\left(\prod_{i=1}^{D} x_i\right)^{\frac{1}{D}} = \ln\left(\exp\left[\frac{1}{D}\sum_{i=1}^{D}\ln(x_i)\right]\right) = \frac{1}{D}\sum_{i=1}^{D}\ln(x_i)$$

Furthermore, CLR coefficients are not subcompositionally coherent, because the geometric mean is not.

Although CLR coefficients are not coordinates with respect to a basis (but coefficients with respect to a degenerate generator), they have several useful properties:

1. $\operatorname{clr}(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha \cdot \operatorname{clr}(\mathbf{x}) + \beta \cdot \operatorname{clr}(\mathbf{y})$ for any $\alpha, \beta \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$;

2. $\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \operatorname{clr}(\mathbf{x}), \operatorname{clr}(\mathbf{y}) \rangle_e$, for any $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, where $\langle \cdot, \cdot \rangle_a$ and $\langle \cdot, \cdot \rangle_e$ are the Aitchison inner product and the Euclidean inner product respectively. Naturally, this property translates to the corresponding norms and metrics.

The first property means that $\operatorname{clr} : \mathcal{S}^D \to V \subset \mathbb{R}^D$ is a vector space isomorphism[13] between a D-part Aitchison simplex and a $(D-1)$-dimensional subspace V of a real vector space $\mathbb{R}^D$, such that all coefficient vectors in V add to zero (naturally, this is also true for any results of vector space operations on objects in V). The second property means that CLR is an isometry[14] between these vector spaces. Take note, that these properties

[13] An invertible structure-preserving transformation between two algebraic structures. The structures are said to be isomorphic under such transform, i.e. indistinguishable under translated operations.

only hold for CLR vectors and not for individual components, because they are not coordinates with respect to an orthonormal basis.

3.8 Isometric (orthonormal) log-ratio transformation and coordinates

3.8.1 General construction

As shown in subsection 3.6, omitting any vector from the generator given in equation 1 results in a basis of $\mathcal{S}^D$, albeit not an orthonormal one. Yet, any basis of a finite-dimensional real inner product space can be orthogonalised and subsequently normalised to obtain an orthonormal basis[15]. Furthermore, by applying the CLR transformation to each vector in the resulting basis of $\mathcal{S}^D$, we obtain an orthonormal basis in the $(D-1)$-dimensional Euclidean space.

Let $(\mathbf{u}_1, \ldots, \mathbf{u}_{D-1})$ be an orthonormal basis in $\mathcal{S}^D$. Then, a $(D-1) \times D$ matrix Ψ, such that the i-th row of Ψ is equal to $\operatorname{clr}(\mathbf{u}_i)$, is called the contrast matrix associated with the orthonormal basis $(\mathbf{u}_1, \ldots, \mathbf{u}_{D-1})$. The rows of Ψ are called log-contrasts. Given the contrast matrix Ψ associated with an orthonormal basis $(\mathbf{u}_1, \ldots, \mathbf{u}_{D-1}) \subset \mathcal{S}^D$, the isometric (orthonormal) log-ratio transform (ILR) is a mapping $\operatorname{ilr} : \mathcal{S}^D \to \mathbb{R}^{D-1}$ defined as

$$\mathbf{x}^* = \operatorname{ilr}(\mathbf{x}) = \Psi \times \operatorname{clr}(\mathbf{x}) = \Psi \times \ln(\mathbf{x}) \tag{5}$$

The inverse transformation is

$$\operatorname{ilr}^{-1}(\mathbf{x}) = \mathcal{C}\left[\exp\left(\Psi^T \times \mathbf{x}^*\right)\right] \tag{6}$$

ILR is both an isomorphism and an isometry between $\mathcal{S}^D$ and the entire $\mathbb{R}^{D-1}$.

3.8.2 Sequential binary partition

Among all possible orthonormal bases of $\mathcal{S}^D$, there is a finite subset of bases associated with sequential binary partitions of compositional vectors [4]. A sequential binary partition (SBP) of a D-part composition is a hierarchy of parts matching the topology of a binary acyclic graph (a tree) with $(D-1)$ internal nodes and D tips (figure 1).

[14] An invertible distance-preserving transformation between two metric spaces. In case of emergent metric space structures of inner product spaces an isometry also preserves the inner product. In other words, an isometry between two inner product spaces preserves geometry.

[15] For further details, you can read about the Gram–Schmidt process.

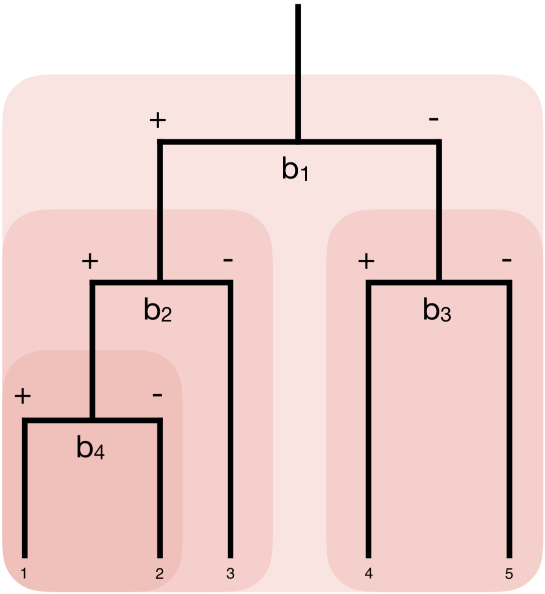

Figure 1: A sequential binary partition of a 5-part composition.

	1	2	3	4	5
b1	+1	+1	+1	-1	-1
b2	+1	+1	-1	0	0
b3	0	0	0	+1	-1
b4	+1	-1	0	0	0

Table 2: A sign matrix corresponding to the sequential binary partition of the 5-part composition from figure 1.

Coordinates with respect to an orthonormal basis induced by an SBP are called balances. A bipartition of D parts can be written as a sign matrix Φ of $(D-1)$ rows (corresponding to balances/nodes) and D columns (corresponding to parts), such that

$$\phi_{ij} = \begin{cases} +1 & \text{if component } j \text{ belongs to the left subgraph of balance } i \\ -1 & \text{if component } j \text{ belongs to the right subgraph of balance } i \\ 0 & \text{if component } j \text{ does not belong to any of these subgraphs} \end{cases}$$

Table 2 shows a sign matrix corresponding to the bipartition from figure 1.

Given a sign-matrix Φ, we can compute the corresponding contrast matrix Ψ

$$\psi_{ij} = \begin{cases} -\frac{k_i}{n_{i-}} & , \phi_{ij} < 0 \\ +\frac{k_i}{n_{i+}} & , \phi_{ij} > 0 \\ 0 & , \phi_{ij} = 0 \end{cases}$$

where $n_{i+} = \sum (\boldsymbol{\phi}_i > 0)$, $n_{i-} = \sum (\boldsymbol{\phi}_i < 0)$ and $k_i = \sqrt{\frac{n_{i-} \cdot n_{i+}}{n_{i-} + n_{i+}}}$. The utility of such orthonormal bases comes from the ability to interpret individual ILR coordinates in terms of original parts:

$$b_i = k_i \cdot \ln \frac{\mathrm{g}_m (\mathbf{x}_+)}{\mathrm{g}_m (\mathbf{x}_-)}$$

where $\mathbf{x}_+$ and $\mathbf{x}_-$ are vectors of parts in the left and right subgraphs of balance i respectively.

3.9 Additive log-ratio transformation

As shown in subsection 3.8, we can derive a basis of $\mathcal{S}^D$ by suppressing any single vector in a generator system of $\mathcal{S}^D$ as shown in equation 1. Likewise, given the same generating system, any $\mathbf{x} \in \mathcal{S}^D$ can be represented as

$$\mathbf{x} = \bigoplus_{i \in \{1,\ldots,D\} \setminus \{j\}} \ln \frac{x_i}{x_j} \odot \mathbf{u}_i, \text{ where } j \in (1, \ldots, D)$$

Here we suppress not only the j-th vector in the generator, but also the corresponding component in the composition. This component is usually called a reference part. The resulting coordinate representation is called the additive log-ratio transform of $\mathbf{x}$ with respect to the reference part j

$$\operatorname{alr}(\mathbf{x}) = \mathbf{x}' = \begin{bmatrix} \ln \frac{x_1}{x_j} & \ldots & \ln \frac{x_{j-1}}{x_j} & \ln \frac{x_{j+1}}{x_j} & \ldots & \ln \frac{x_D}{x_j} \end{bmatrix} \tag{7}$$

The inverse transformation is

$$\operatorname{alr}^{-1}(\mathbf{x}') = \mathcal{C} \begin{bmatrix} \exp(x'_1) & \ldots & \exp(x'_{j-1}) & 1 & \exp(x'_{j+1}) & \ldots & \exp(x'_D) \end{bmatrix} \tag{8}$$

These coordinates are not symmetric under the choice of reference and, unlike ILR coordinates, are given with respect to an oblique basis. Like CLR and ILR, ALR is a vector space isomorphism, though it is not an isometry[16] between the Aitchison simplex and a Euclidean vector space. ALR's utility

[16] It is, however, possible to represent CLR coefficients and ILR coordinates as a linear combination of ALR coordinates and, by extension, calculate Aitchison inner product, norm and distance in the usual Euclidean manner.

lies in its simplicity and interpretability, though care should be taken when applying statistical analyses to ALR-transformed compositions. In general, it is safe to use methods that rely solely on vector space operations and do not depend on the inner product (along with all ensuing concepts).

4 Amplicon Data Preprocessing

4.1 Problem statement

An amplicon sequencing library is a noisy collection of extracted, amplified and sequenced instances of a target DNA/RNA sequence. In most practical cases we do not know exactly what instances to expect — we have to infer them from the data, while accounting for artefacts, such as distorted and chimeric sequences, to avoid diversity inflation. Then, we need to transform the library into a count vector $\mathbf{x} \in \mathbb{N}_0^n$, where x_i is the number of observations (reads) corresponding to instance i. Bear in mind, that all steps in an experimental metagenomic workflow, from DNA extraction and purification to sequencing itself, inevitably distort the underlying community structure in one way or another, and there is little we can do about it.

4.2 Clustering-based preprocessing

Up until recently, the absolute majority of published microbiome studies have used a data preprocessing step known as OTU-picking. In general, an OTU (operational taxonomic unit) is a group of target sequences satisfying a similarity criterion. There are three main types of OTU-picking strategies: de novo, closed-reference and open-reference (the last one is a combination of the other two). De novo OTU-picking algorithms create groups from scratch and select (or compute) one representative sequence (a centroid) per group. In closed-reference OTU-picking we start with an existing set of representative sequences (references) and group reads from an amplicon library with any one of the references according to a similarity criterion without creating any new representatives. In practice, most OTU-picking algorithms involve some sort of alignment-based sequence clustering with a fixed similarity threshold (e.g. 97%). The process is usually preceded by primer excision (so as not to bias similarity scores), quality-score-based trimming and read-pair merging and is followed by chimera removal and frequency-based filtering.

A fixed similarity threshold is supposed to absorb experimental variations and group all inexact readings of a single underlying target instance. However, it also incurs a risk of merging actual distinct targets, thereby deflating diversity estimates and erasing potentially valuable statistical information. In addition to that, accurate (de novo) clustering (based on complete-linkage algorithms) is not computationally feasible, while heuristic alternatives (including advances statistical methods with adaptive similarity thresholds, e.g. [5]) are extremely unreliable [6, 7, 8]. Finally, the inability to jointly analyse datasets produced by independent OTU-picking runs (with the exception of closed-reference OTU picking) adds to the list of reasons to avoid OTU-picking altogether.

4.3 Denoising-based preprocessing

OTU-picking has largely fallen out of favour with progressive microbiome researchers (the rest are catching up rapidly). Denoising reads (that is applying error-correction) is a far better and consistent way to handle experimental variation in sequencing data, because it does not involve any similarity-based clustering. There are many different denoising algorithms, which is why students are advised to read a comparative review [9] of the three most popular ones, namely DADA2, Deblur and UNOISE3. In any case result is a set of all unique error-corrected target instances observed in an amplicon sequencing library, that is amplicon sequencing variants (ASVs) also known as sub-OTUs.

4.4 Chimeric sequences

A chimera is a synthetic combination of two or more distinct DNA templates. They appear when an aborted PCR product of one real DNA template anneals to a different template during multi-template PCR. Most amplicon sequencing libraries contain chimeras, though the exact fraction of chimeric sequences varies widely and depends on multiple factors, such as initial template concentrations and the total number of PCR cycles. It is not entirely unusual to have chimeras amount to over 50% of a sample. Consequently, all decent amplicon data preprocessing workflows include chimera detection and removal (typically as the final step). Although perfect chimera removal does not seem feasible even when we possess exact sequences of all initial DNA templates [10], the choice of upstream preprocessing steps does influence the accuracy a lot. In particular, OTU-picking hinders chimera-detection by masking non-experimental variation in closely related initial (i.e. real) template sequences. Since denoising strategies tend to reconstruct most of the real initial templates, they greatly improve and simplify chimera detection.

4.5 Copy-number variations

A single genome can encode multiple instances of a target sequence: identical (or sufficiently similar) copies distort component abundances, while distinct copies appear as separate entities (as though sequences originating from different microbial populations) and distort community diversity. Considering that, the number of ribosomal operons varies a lot (in prokaryotes as well as eukaryotes) even on a small phylogenetic scale, copy-number-variations are another significant data-obfuscating factor to be aware of, even though there is not much we can do about it now (if ever).

5 Taxonomic annotation

5.1 Problem statement

From a practical point of view, taxonomic annotation is an implicit way to augment amplicon sequencing data with functional information based on microbiological knowledge about certain taxonomic groups. However, taxonomy is a largely synthetic construct that can easily defy “natural” phylogenetic relationships between microbial populations. This is especially true about prokaryotic taxonomy, because, on the one hand, standard (i.e. eukaryotic) definitions of species (based on reproductive isolation of populations) are incompatible with prokaryotic modes of reproduction, and, on the other hand, traditional prokaryotic systematics is more concerned with phenotypic/functional traits than actual phylogeny. For example, *Shigella* and *Escherichia* are treated as two separate genera, though phylogenetically *Shigella* species fall well within the natural intra-species variation of *E. coli* [11, 12]. This phylogenetic inconsistency makes accurate taxonomy prediction (and, by extension, functional inference) almost impossible, especially given a rather minuscule amount of genetic information we have in short-read amplicon sequencing datasets.

5.2 Reference datasets

Taxonomic annotation requires a set of reference sequences with known taxonomic identities. There are three large general-purpose (that is domain-agnostic) 16S rRNA datasets: RDP, Silva and GreenGenes; only the first two of them are updated on a regular basis. All three databases are error-ridden and thus require careful filtering to improve annotation accuracy [13]. Sequences with partial, conflicting or completely inaccurate taxonomic annotations constitute up to a quarter of all records. Manually curated domain-specific databases, such as HITdb, tend to be significantly more accurate, but they are rarely available.

5.3 Alignment-based annotation

Given a database of reference sequences with known taxonomic identities, sequence alignment becomes the most straightforward approach to taxonomic annotation. This process is essentially similar to closed-reference OTU-picking with a couple of noticeable distinctions. First of all, alignment-based taxonomic annotation is usually applied to preprocessed data (i.e. chimera-checked and filtered ASVs or OTUs), though it is not entirely uncommon to keep taxonomic annotations produced as a side-product of a closed-reference OTU-picking run. Second of all, accurate alignments are computationally feasible, because we only need to align ASVs or OTU centroids. As with OTU-picking, the 97% similarity threshold was historically considered a good criterion for a perfect reference match. However, modern studies show that only 100% similarity provides any sort of annotation accuracy for partial 16S sequences (e.g. the V4 region) [14, 15]. The bigger problem is the inability to control annotation uncertainty.

5.4 Machine-learning-based annotation

Machine-learning-based taxonomic classifiers typically provide at least some way of controlling prediction uncertainty. The most popular taxonomic classifiers are the RDP classifier [16] and IdTaxa [17]. The first one is a naïve Bayesian classifier, and the second one implements a custom model inspired by decision trees. Both classifiers treat sequences as bags-of-words, that is unordered k-mer spectra. Since k-mer spectra of partial 16S sequences diverge from complete-sequence spectra, to avoid unreliable uncertainty estimates we must always use classifiers trained on the target regions we are dealing with.

5.5 Phylogenetic placement

Given a phylogenetic tree of annotated referency sequences, phylogenetic placement algorithms determine the most probable location of a query sequence under a maximum likelihood or a Bayesian phylogenetic model. Although placement is quite an old method, it has become far more accessible (i.e. user-friendly) with the advent of SEPP [18], a single-command plugin for the QIIME 2 amplicon data analysis framework that comes bundled with a precomputed GreenGenes phylogenetic tree. Phylogenetic placement is in many ways superior (at least theoretically) to both alignment-based and machine-learning-based methods. In practice, the accuracy depends not only on the quality of reference annotations (as is mostly the case with the other two methods), but also on the quality of reference phylogeny; and large-scale phylogenetic trees are never really accurate [13] and extremely computationally expensive to infer.

6 Statistical analysis: introduction

6.1 Problem statement

We can summarise the goals of most microbiome studies in two broad questions:

- how a community, as a whole, responds to spatiotemporal changes in the environment, and
- what happens to individual microbial populations in the community?

To set up some context for further discussions, we need to recognise sequencing libraries as discrete compositions (or compositional count data) [19, 20, 21]. As shown in subsection 4.1, a sequencing library is a collection of reads corresponding to a set of D distinguishable DNA or RNA sequences. During preprocessing we transform the library into a count vector $\mathbf{x} \in \mathbb{N}_0^D$, such that x_i is the number of observations (reads) corresponding to the i-th sequence[17]. The sum of all counts in a vector is bound by the sequencing depth, $n \in \mathbb{N}$, and thus pertains no information about absolute abundances in the original environment. In other words, the information contained in $\mathbf{x}$ is relative: $\mathbf{x}$ provides a finite-sample estimate of a latent canonical composition[18] (relative abundances). Consequently, microbiome data are sampled from a constrained (Aitchison) geometry and thus are incompatible with most conventional statistical methods. At the same time, compositional count data differ from canonical compositions (as per subsection 3.2) in a significant way — counts are not truly scale invariant, because the scale (sequencing depth) directly influences the variance of estimated relative abundances[19].

[17] Although we have introduced this abstract notion of preprocessing in the context of amplicon libraries, it can be easily extended to shotgun metagenomes (and high-throughput sequencing data in general) by adjusting the definition of "distinguishable sequences".

[18] In the most basic case, we can use the closure of counts $\mathbf{x} \in \mathbb{N}_0^D$ as an estimator of latent relative abundances $\mathbf{y} \in \mathbb{R}_+^D$ assuming that $\lim_{n\to\infty} \mathcal{C}(\mathbf{x}) = \mathbf{y}$.

[19] While the discrete nature of sequencing data is self-apparent, even seemingly continuous compositional measurements are counts disguised by an enormous sampling depth. For example, chemical compositions are little more than counts of molecules.

6.2 Zeros in compositional data

6.2.1 Types of zeros

There are three fundamental types of zeros in compositional data in general: rounding zeros, missing data and structural zeros. Rounding zeros and missing data are In case of compositional count data, detection sensitivity roughly translates to sampling (sequencing) depth, though this is not entirely true. Missing data and structural zeros arise in joint analyses of datasets with different sampling histories. Let N and M be two sets of components observed in compositional count vectors $\mathbf{x} \in \mathbb{N}^{|N|}$ and $\mathbf{y} \in \mathbb{N}^{|M|}$, such that $N \cap M \neq N \cup M$. Vectors $\hat{\mathbf{x}}, \hat{\mathbf{y}} \in \mathbb{N}_0^{|N \cup M|}$ in the joint dataset will have $(N \cup M) \setminus (N \cap M)$ zeros that can be either missing data or structural zeros. Missing data are omitted (unmeasured) components and as such are related to rounding zeros. Structural zeros correspond to parts that are fundamentally absent from either N or M. Regardless of their origin, zeros are not compatible with compositional data analysis, because compositions with zero parts are mathematically undefined. To deal with the problem, we can replace zeros by some positive values and/or omit them from the dataset. In practice, we fully discard heavily zero-inflated parts (e.g. parts with exactly zero, or very few, observations in more than 50% of all count vectors in the dataset; exact criteria are subject to optimisation) and replace zeros in the remaining data.

6.2.2 Zero replacement

The most basic way to remove zeros in count data is to simply replace zeros by a small number $\lambda \in \mathbb{N}$ (usually 1) to all counts. This approach is called additive zero replacement. Given a count vector $\mathbf{x} \in \mathbb{N}_0^D$, $\hat{\mathbf{x}} = \mathbf{x} + \lambda = (x_1 + \lambda, x_2 + \lambda, \ldots, x_D + \lambda)$. Additive zero replacement has several significant drawbacks:

1. it masks missing data and structural zeros;

2. it fully neglects the uncertainty associated with rare/missing parts;

3. it non-uniformly (and non-linearly) distort ratios between original positive observations[20];

4. it distorts the covariance structure: zero-inflated variables become associated by virtue of being consistently filled by the same constant value;

The so called Bayesian Monte Carlo zero replacement strategy implemented in package ALDEx2 [22] is a more advanced variation on additive zero replacement. Given an initial compositional count vector $\mathbf{x} \in \mathbb{N}_0^D$, we take $\hat{\mathbf{x}} \sim \text{Dirichlet}(\mathbf{x}+1)$. The support of a Dirichlet distribution with D parameters (called concentrations) is a set of all multinomial probability vectors of D outcomes (parts) with expectation $\mathrm{E}\left[\text{Dirichlet}(\mathbf{x}+1)\right] = \mathcal{C}\left(\mathbf{x}+1\right)$. This procedure is roughly equivalent to using a non-informative Dirichlet prior to replace zeros in estimated relative abundances. The scale of concentrations (that is counts in our case) controls variance. Consequently, by taking multiple samples $\hat{\mathbf{x}} \sim \text{Dirichlet}(\mathbf{x}+1)$ for each vector $\mathbf{x}$ in our initial dataset, we inject uncertainty into downstream analyses, thereby alleviating the second and, to some extent, the first downside of pseudocount zero replacement, though it still falls short of providing a rigorous framework for missing data and structural zeros. Since $\lim_{n\to\infty} \mathcal{C}\left(\mathbf{x}+1\right) = \mathcal{C}\left(\mathbf{x}\right)$, the method addresses the third downside of pseudocount zero replacement (at least asymptotically). Although this strategy is quite effective and simple, it requires careful handling, because Dirichlet samples originating from the same initial compositional count vector appear as faux (technical, rather than biological) replicates (thereby inflating the number of degrees of freedom). In the context of linear modelling, mixed-effects models provide a rich framework for dealing with faux replicates.

[20] Let $\mathbf{x} \in \mathbb{N}^D$ be a count vector sorted in descending order. Then, $\ln \frac{x_i+\lambda}{x_i} \leq \ln \frac{x_j+\lambda}{x_j}$ for any $i < j$.

References

[1] M. Bordewich *et al.*, "On the information content of discrete phylogenetic characters," *Journal of Mathematical Biology*, vol. 77, pp. 527–544, 2018.

[2] P. O. Lewis *et al.*, "Estimating Bayesian Phylogenetic Information Content," *Systematic Biology*, vol. 65, pp. 1009–1023, 2016.

[3] Y. Wang *et al.*, "Optimal Eukaryotic 18S and Universal 16S/18S Ribosomal RNA Primers and Their Application in a Study of Symbiosis," *PLoS ONE*, vol. 9, p. e90053, 2014.

[4] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal, "Isometric Logratio Transformations for Compositional Data Analysis," *Mathematical Geology*, vol. 35, pp. 279–300, 2003.

[5] S. W. Olesen *et al.*, "dbOTU3: A new implementation of distribution-based OTU calling," *PLOS ONE*, vol. 12, p. e0176335, 2017.

[6] W. Chen *et al.*, "A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs," *PLoS ONE*, vol. 8, p. e70837, 2013.

[7] M. A. Jackson *et al.*, "A heritability-based comparison of methods used to cluster 16S rRNA gene sequences into operational taxonomic units," *PeerJ*, vol. 4, p. e2341, 2016.

[8] R. C. Edgar, "Accuracy of microbial community diversity estimated by closed- and open-reference OTUs," *PeerJ*, vol. 5, p. e3889, 2017.

[9] J. T. Nearing *et al.*, "Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches," *PeerJ*, vol. 6, p. e5364, 2018.

[10] R. Edgar, "UCHIME2: improved chimera prediction for amplicon sequencing," *bioRxiv*, p. 074252, 2016.

[11] J. R. Johnson, "Shigella and Escherichia coli at the crossroads: machiavellian masqueraders or taxonomic treachery?," *Journal of Medical Microbiology*, vol. 49, pp. 583–585, 2000.

[12] E. A. Pettengill *et al.*, "Phylogenetic Analyses of Shigella and Enteroinvasive Escherichia coli for the Identification of Molecular Epidemiological Markers: Whole-Genome Comparative Analysis Does Not Support

Distinct Genera Designation," *Frontiers in Microbiology*, vol. 6, p. 1573, 2016.

[13] R. Edgar, "Taxonomy annotation and guide tree errors in 16S rRNA databases," *PeerJ*, vol. 6, p. e5030, 2018.

[14] R. C. Edgar, "Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences," *PeerJ*, vol. 6, p. e4652, 2018.

[15] R. C. Edgar, "Updating the 97% identity threshold for 16S ribosomal RNA OTUs," *Bioinformatics*, vol. 34, pp. 2371–2375, 2018.

[16] Q. Wang *et al.*, "Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy," *Applied and Environmental Microbiology*, vol. 73, pp. 5261–5267, 2007.

[17] A. Murali *et al.*, "IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences," *Microbiome*, vol. 6, p. 140, 2018.

[18] S. Janssen *et al.*, "Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information," *mSystems*, vol. 3, pp. e00021–18, 2018.

[19] G. B. Gloor, J. M. Macklaim, M. Vu, and A. D. Fernandes, "Compositional uncertainty should not be ignored in high-throughput sequencing data analysis," *Austrian Journal of Statistics*, vol. 45, no. 4, pp. 73–87, 2016.

[20] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome Datasets Are Compositional: And This Is Not Optional," *Frontiers in Microbiology*, vol. 8, p. 2224, 2017.

[21] T. P. Quinn, I. Erb, M. F. Richardson, and T. M. Crowley, "Understanding sequencing data as compositions: an outlook and review," *Bioinformatics*, vol. 34, no. 16, pp. 2870–2878, 2018.

[22] A. D. Fernandes, J. N. Reid, J. M. Macklaim, T. A. McMurrough, D. R. Edgell, and G. B. Gloor, "Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis," *Microbiome*, vol. 2, no. 1, p. 15, 2014.